BeleidsonderzoekOnline

DOI: 10.5553/BO/221335502019000006001

Evaluatieonderzoek, Big Data en Artificiële Intelligentie: een verkenning

Frans L. Leeuw

Aanbevolen citeerwijze bij dit artikel
Frans L. Leeuw, 'Evaluatieonderzoek, Big Data en Artificiële Intelligentie: een verkenning', *Beleidsonderzoek Online* juni 2019, DOI: 10.5553/BO/221335502019000006001

1 Achtergronden

Beleid van overheden en andere (semipublieke) organisaties vindt steeds meer 'in digitalis' plaats. Deels omdat het om opvattingen en gedragingen van mensen gaat die (volledig) 'in digitalis' plaatsvinden en daar ook 'beïnvloed' worden door middel van beleidsinterventies (denk aan digitale piraterij, cyber pesten of de handel in e-currencies). Deels omdat er digitale instrumenten worden ingezet bij het invoeren van 'oorspronkelijk' niet-digitaal beleid (denk aan e-governance, e-health, e-diplomacy, tele-leren). En deels omdat de digitale wereld uit zichzelf noopt tot digitaal beleid: denk aan de effecten van (digitale) zorgplichten van internetproviders, het (digitale) privacy & data-beleid van instellingen en nieuwe vormen van 'digitale privacy' (Koops, 2018),[1] maar ook regelgeving rondom (big) data en Artificiële Intelligentie.

Voor beleidsevaluatoren zijn er de afgelopen twintig jaar met andere woorden nieuwe, spannende en uitdagende probleemvelden bij gekomen. Tegelijkertijd is vastgesteld dat de belangstelling van de professie voor deze thematiek lange tijd vrij minimaal geweest. Leeuw en Leeuw (2012) beschreven deze 'gap' voor internationale evaluatie-tijdschriften. Tabel 1 is gebaseerd op een telling van woorden die enig inzicht gaven in de aandacht van evaluatie-onderzoekers voor 'cyber, internet en digitaal'. De tijdschriften publiceerden in de onderzochte periode een kleine 2 miljoen woorden waarvan er maar een paar honderd de digitale wereld betroffen.

Tabel 1 Resultaten van literatuuronderzoek (Leeuw & Leeuw 2012)

Journal (data were collected in January/ February 2011)	Words internet or web or digital or cyber found in full text	Words internet or cyber or digital or web found in keywords	Words digital and policy found in full text/ keywords
Evaluation (1995-)	77 times	----	5/0
Evaluation and Program Planning (1981-)	228 times	*	43/0*
Assessment and Evaluation in Higher Education (1981-)	174 times **	3 times	25/**
Evaluation and Research in Education (1981)	11 times **	----	8/0
Educational Evaluation and Policy Analysis (1981)	67 times	1 time	3/0
Evaluation Review (1981-)	76 times	3 times	9/0
Evaluation and the Health Professions (1987-)	103 times	4 times	2/0
New Directions for Evaluation (1981-)	127 times	*	6/*
American Journal of Evaluation (1981-)	219 times	4 times	18/0
The Journal of Multidisciplinary Evaluation (2004-)	8 times	*	1/0

Notes:
[1] Number of times the words ´internet´, ´digital´, ´web´ or 'cyber' and 'digital and policy' appeared in 10 peer-reviewed evaluation journals (English) (since 1981 or since the first time the Journal was published – the year in parenthesis is either the starting year of the journal or the starting point of the search.).
-----: not found
*) this journal does not make it possible through an electronic way to differentiate between keywords and full text, or this service was not available during the preparation of this paper.
**) Only three (combined) search words were possible;'cyber' was excluded.
Full text: to check terms or phrases that appear anywhere in the full-text article;
Keywords: to check terms in 'keywords', as applied to articles by authors.
Data were collected in January/February 2011 and in October 2011 by the second author and checked by the first author.

Enkele jaren later werden indicatoren gebruikt, die evenmin wezen op grote belangstelling van evaluatoren voor de digitale samenleving en in het bijzonder Big Data. Zo vonden Forss en Norén (2017) in een inhoudsanalyse van een steekproef van zo'n 25 Terms of Reference (TOR)-documenten ten behoeve van evaluaties op het gebied van ontwikkelingssamenwerking, dat Big Data nog steeds zo goed als niet genoemd werden. Ook verhinderde een aantal TOR's het gebruik ervan.

Een ander signaal kwam uit een e-enquête die verspreid werd via een online koppeling naar actieve LinkedIn-groepen van evaluatoren (Høljund et al., 2017). Op deze manier was het mogelijk om meer dan 85.000 evaluatoren (wel met dubbelgangers) enkele vragen te stellen. Het bleek dat slechts weinigen in de steekproef aangaven met Big Data te werken (minder dan 1%).

De vierde bron is een rapport van UN Global Pulse over het integreren van Big Data in de monitoring en evaluatie van ontwikkelingsprogramma's (Bamberg, 2016). Naar het oordeel van Bamberger (2016: 19-21) is het 'absoluut noodzakelijk dat evaluatoren vertrouwd gaan worden met nieuwe gegevensbronnen, technologieën en methodologieën en deze integreren in hun werk. [Ik hoop] dat dit rapport niet alleen als inleidende gids voor Big Data kan dienen, maar ook als een dringende oproep tot actie (…) om ontwikkelingsagentschappen en met name evaluatoren te inspireren

om samen te werken met datascientists en analisten bij de exploratie en de toepassing van nieuwe gegevensbronnen, methoden en technologieën.'

2 Vraagstelling en structuur artikel

Dit artikel beperkt zich tot het thema van de *Big Data en Artificiële Intelligentie* (afgekort: BD/AI). Relaties van evaluatieonderzoek met andere ontwikkelingen in de digitale samenleving blijven buiten beschouwing.[2]

Twee vragen staan centraal:

- De eerste vraag is *wat BD/AI te bieden hebben aan evaluatieonderzoek van (digitaal) beleid?* Daarover gaat paragraaf 4.

- De tweede vraag komt aan de orde in paragraaf 5: *wat heeft evaluatieonderzoek te bieden als het gaat om het analyseren/onderzoeken van de betrouwbaarheid, validiteit en enkele andere aspecten van Big Data en AI?*

Voorafgaand daaraan omschrijf ik in paragraaf 3 enkele begrippen.

Ten slotte worden in paragraaf 7 enkele conclusies getrokken waarbij ik met een schuin oog ook naar de toekomst kijk.

3 Enkele begripsomschrijvingen

De term *Big Data* verwijst naar de enorme omvang van diverse soorten gegevens en de daarbij passende 'analytics'. Er is geen minimale hoeveelheid vereist om van BD te kunnen spreken, maar het is wel duidelijk dat het gaat om vele terabytes (TB), petabytes (PB) en exabytes (EB) met gegevens. Stephens et al. (2015) vergeleken enkele jaren geleden al vier Big Data-terreinen en stelden de vraag wie over een paar jaar de 'grootste' zal zijn: astronomie (opslag 1 EB jaarlijks), Twitter (1-17 PB), YouTube (1-2 EB) of genomics (2-40 EB).

Naast aandacht voor deze (eerste) V(olume) wordt ook over vier andere V's gesproken. V # 2 heet 'velocity', de snelheid waarmee Big Data gegenereerd worden. Het viraal gaan van sociale mediaberichten, het mondiaal real time checken van creditcardbetalingen, het meten van vervoersstromen via smart apparaten en tal van andere voorbeelden laten zien waar het bij deze V om gaat. V # 3 heet 'variety' en betreft de diverse typen data: documenten, audio, video, foto's, (f)mri-metingen, QS [quantified self]-gegevens (De Kogel & Cornet, 2016: 79 en 81)[3] en 'klassieke' (= register en via bevraging verkregen) data, m.a.w. gestructureerde en ongestructureerde data. V # 4 staat

voor 'veracity': het verschijnsel dat er van alles en nog wat aan data verzameld wordt dan wel zichzelf verzamelt, en er vragen te stellen zijn bij de betrouwbaarheid ervan. Soms wordt nog over V # 5 gesproken: 'value' (de waarde vooral in economische zin van Big Data).

In een recent KNAW-rapport (2018) schrijft de commissie dat het 'belangrijkste onderscheid tussen "gewone" data en Big Data in de hoeveelheden en grote verscheidenheid aan soorten (hoge dimensionaliteit) van gegevens zit, die op talloze terreinen kunnen worden vergaard, opgeslagen, gekoppeld en geanalyseerd. De hoeveelheden en hoge dimensionaliteit blijven explosief toenemen'. Tegelijkertijd zijn Big Data niet veel waard als er geen *analytics* beschikbaar zijn en gebruikt worden: opslag en beheer van data, methoden om ermee te werken, inclusief procesregels en 'governance' van de data(analyse/duidingen/conclusies). Tezamen wordt wel gesproken over het Big Data Eco Systeem.

Bij *Artificiële Intelligentie* spelen algoritmes een cruciale rol. Een algoritme is in essentie een lijst van stappen/instructies om met een computerprogramma een probleem op te lossen. Er bestaan diverse 'soorten' (of niveaus) van AI, waarbij wel verwezen wordt naar deep learning als de thans meest geavanceerde vorm van machine learning. AI wordt ook wel omschreven als een systeem dat (tot op zekere hoogte) autonoom kan waarnemen, handelen en zich aanpassen. In toenemende mate beschikken machines dankzij AI over 'menselijke kwaliteiten' als waarnemen, leren, plannen, realiseren, zoeken en handelen.

Er zijn tal van toegankelijke publicaties beschikbaar waarin deze begrippen uitgebreider en met voorbeelden worden toegelicht. Zie daarvoor o.a. Klous & Wielaard (2014); Domino (2015); Gandomi & Haider (2015), maar ook deze site: https://www.linkedin.com/pulse/intelligent-things-its-all-machine-learning-roger-attick/ (geraadpleegd 30 januari 2019).

4 Vraag 1: wat hebben BD/AI te bieden aan evaluatieonderzoek van (digitaal) beleid?

Ten eerste kunnen met Big Data sneller maatschappelijke trends *gescand* worden dan mogelijk is met traditionele gegevensverzameling/bronnen (zoals enquêtes). Dat geldt in ieder geval voor trends die zich vooral *in digitalis* afspelen. Denk aan (de aanpak van) cyber pestgedrag, digitale piraterij, handel in verdovende middelen, wapens en mensen op het Dark Web, maar denk ook aan *gaming*, communicatie en voorlichting via YouTube video's (in plaats van door middel van papieren handleidingen). Maar ook als het om niet primair digitale ontwikkelingen gaat, is openstaan voor het

gebruik van Big Data (doorgaans) verstandig voor een evaluator. Een voorbeeld geeft een studie van trends in faillissementen in Nederland. Willemsen en Leeuw (2017) vergeleken de officiële CBS-statistieken over faillissementen over een periode van zo'n tien jaar met gegevens die via query's en Google Correlate uit zoekgedrag van mensen in die periode waren op te diepen. Ze probeerden de in de CBS-statistiek aangetroffen ontwikkelingslijn van aantallen faillissementen na te bootsen op basis van gegevens over online zoekgedrag. Dat lukte wonderwel. En: de op grond van Google-zoekgedrag/Correlate gemaakte voorspellingen ten aanzien van de ontwikkeling in faillissementen (ná de meetperiode: in dit geval 2014/2015) zouden zelfs wat beter zijn geweest met gebruikmaking van Big Data dan zonder. Daas et al.[4] deden hetzelfde met de CBS-enquête naar consumentenvertrouwen. Daar werden de resultaten uit enquêtes vergeleken met uitkomsten uit sentimentsanalyses en ook hier met verbluffende overeenkomsten.

Ten tweede kunnen Big Data gebruikt worden *om contexten waarin beleidsprogramma's, waaronder wetgeving werken, te scannen en om baseline-gegevens te reconstrueren (ook ná de introductie van (nieuwe) beleidsprogramma's en wetten*). Het in kaart brengen van contexten door gebruik te maken van Big Data gebeurt op verschillende manieren. In 2014 kwam de *Global Database of Events, Language and Tone* (Gdelt) beschikbaar, waarin gegevens over maatschappelijke ontwikkelingen, waaronder politieke conflicten sinds 1979 te vinden zijn. Gdelt gebruikt het nieuws van tv-uitzendingen, bladen, en websites vanuit veel landen, in meer dan 100 talen en doet dat aan de hand van wat aan mensen, locaties, organisaties, thema's, bronnen, afbeeldingen, emoties (zoals conflicten) te vinden is. Het bevat de lengte- en breedtegraad voor elke gebeurtenis – *geotagged* op stadsniveau. Voor evaluaties is dit een rijke bron die kan worden gebruikt om diverse contexten (sociale, historische, bestuurlijke) weer te geven en in de evaluaties te betrekken; wel dient rekening te worden gehouden met de mate waarin media objectiviteit de facto nastreven en realiseren. Bail (2014) presenteert voorbeelden op het niveau van organisaties, bijvoorbeeld om hun cultuur te leren kennen. In plaats van gegevens daarover via interviews te verzamelen (inclusief de daarbij behorende non-response problemen en 'research fatigueness'), laat hij zien hoe screenscraping-technologie wordt gebruikt om informatie van websites, maar ook uit boeken, YouTube-video's, toespraken, lezingen en ondernemingsblaadjes op te halen is. Datawarehouses en repositoria zoals Lexis-Nexis of ProQuest bevatten digitale kopieën van de meeste tijdschriften, kranten en tijdschriften op de wereld. Bader et al. (2017) gebruiken Google Street View om 'een veelbelovend alternatief te bieden voor onderzoekers om wijkomgevingen in steden in kaart te brengen en om te onderzoeken hoe de plaatselijke omstandigheden binnen een groter geografisch bereik variëren'. Naast

de aandacht voor de voordelen van het gebruik van 'virtuele' systematische sociale observaties in de buurt bespreken ze ook de ermee gepaard gaande valkuilen en uitdagingen (Bader et al., 2017: 20). De meeste van deze data zijn in staat om baselines achteraf te reconstrueren. Dat is belangrijk, omdat een van de problemen bij evaluatieonderzoek is dat metingen soms pas van start kunnen gaan als het nieuwe beleid of de nieuwe wet al is uitgerold, zonder dat er sprake is geweest van een *nul-meting*. Met gebruikmaking van digitale gegevens kan dit – deels – wel. Neem het *MIT Billion Prices-project*: hoe hoog prijzen in een bepaalde regio vóór het wijzigen van of invoeren van (nieuw) beleid waren, kan met deze gegevens gemakkelijk worden 'gereconstrueerd'. Hetzelfde geldt voor Google Street View-gegevens die kunnen worden gebruikt om de ontwikkeling in de infrastructuur en gebouwen in steden te volgen.

Ten derde zijn Big Data van belang voor het *uitvoeren van impact analysis/doelbereiking en/of effectmeting van beleidsinterventies*. Onderzoekers hebben in de Big Data-wereld (soms) toegang tot gegevens die steeds meer (bijna) real-time beschikbaar zijn. Ik geef twee voorbeelden. *Het eerste betreft Google-zoekgedrag*. Dat soort data zijn minder gevoelig voor leugen en bedrog dan enquêtes (per telefoon of computer).[5] Over hoe dat soort gegevens gebruikt kunnen worden in een evaluatieonderzoek, verwijs ik naar een studie naar (justitieel) beleid om online gedrag te beïnvloeden, in casu digitale piraterij. H.B.M. Leeuw (2017) promoveerde op een onderzoek naar de werking van beleidsinterventies gericht op dat gedrag. Een van de doelen van het onderzoek was na te gaan of en hoe online zoekgedrag, gerelateerd aan digitale (muziek)piraterij, in de VS veranderde na de implementatie van een specifieke antipiraterij-interventie (het Copy Right Alert System of CAS). Dankzij Big Data was het mogelijk om zoekopdrachten te selecteren die afkomstig waren uit de VS en om een voormeting te doen om te bepalen hoe dit zoekgedrag er *voor* de implementatie van CAS uitzag. De dataset die hiervoor werd gebruikt, was gebaseerd op Google Trends, die alle zoekopdrachten van zijn gebruikers opslaat. Om de wijzigingen in het online zoekgedrag met betrekking tot digitale piraterij te meten, was het nodig om een aantal sleutelwoorden in te voeren in Google Trends en die vervolgens binnen een bepaalde periode (een jaar voor en een jaar na de implementatie van het CAS) te meten in de 'digitale empirie'. Vervolgens konden tijdtrends worden gemaakt om in kaart te brengen hoe het gedrag in de loop van de tijd veranderde. De volgende stap was om deze trends te vergelijken met de trends die werden gegenereerd op basis van gegevens die een periode na de implementatie van het CAS weerspiegelen. Grote verschillen tussen deze twee sets van trends in de tijd *zouden* er op (kunnen) wijzen dat het online zoekgedrag in verband met digitale piraterij is veranderd door de invoering van CAS. Zo'n conclusie was methodologisch echter niet te trekken, omdat veranderingen in online gedrag ook (geheel of

gedeeltelijk) het gevolg kunnen zijn van een of meer andere factoren. Dat is het probleem van de (causale) attributie of contributie, iets waar Big Data (maar ook álle andere 'typen' data) niet dé oplossing voor zijn.[6] Wat de onderzoeker vervolgens deed, was het vergelijken van twee jurisdicties (Canada/VS), waarbij CAS wel in de VS was uitgerold maar niet in Canada, en de resultaten van die vergelijking nam hij mee in zijn verdere analyse. Ook analyseerde hij met behulp van klassieke (enquête)data welke mechanismen een rol speelden bij digitale piraterij van, in zijn geval, studenten. Op theoretisch niveau was zo een link te leggen tussen de verschillende typen data en het vraagstuk van de invloed van anti-piraterij-interventies.

Het tweede voorbeeld maakt gebruik van *satellietdata* (Watmough et al., 2019). De onderzoekers richten zich op het monitoren van de voortgang van een van de *Sustainable Development Goals* (SDG's) van de VN (armoedevermindering). Het volgen van deze ontwikkeling en proberen vast te stellen wat de impact van interventies is, vereist frequente, actuele gegevens over sociale, economische en ecosysteemomstandigheden. In de wereld van de SDG's wordt nog vaak gebruik gemaakt van enquêtes onder huishoudens, die frequent afgenomen moeten worden. De onderzoekers wilden dit anders doen en vroegen zich af of satellietgegevens zouden kunnen helpen bij het monitoren van de voortgang van armoedevermindering op het platteland van Kenia. Zij stelden twee vragen: (i) Kan de rijkdom van huishoudens worden voorspeld op basis van satellietgegevens? (ii) Is een socio-ecologisch geïnformeerde multilevel-analyse van de satellietgegevens in staat om de variantie in vermogens van huishoudens te verklaren?

> 'We found that satellite data explained up to 62% of the variation in household level wealth in a rural area of western Kenya when using a multilevel approach. This was a 10% increase compared with previously used single-level methods, which do not consider details of spatial landscape use. The size of buildings within a family compound (homestead), amount of bare agricultural land surrounding a homestead, amount of bare ground inside the homestead, and the length of growing season were important predictor variables. Our results show that a multilevel approach linking satellite and household data allows improved mapping of homestead characteristics, local land uses, and agricultural productivity, illustrating that satellite data can support the data revolution required for monitoring SDGs, especially those related to poverty and leaving no one behind.'

Ten vierde kan via evaluatieonderzoek zicht worden gekregen op gewenste en ongewenste (*neven)effecten van technologieën* (in de digitale wereld). Het is een gegeven dat de invloed van technologie gepaard gaat met uitkomsten en gevolgen die (door sommigen)

verwacht (en gekend) zijn, maar vermoedelijk veel vaker met gevolgen waar dat niet voor geldt. Mediafilosoof Marshall McLuhan wees ooit op een achterliggend mechanisme: 'Wij vormen onze gereedschappen en daarna vormen zij ons.'[7] Voorbeelden zijn er talloos, van positieve (de internet-utopie van de vroege jaren negentig) tot dystopische; denk bij dat laatste aan omvangrijke online 'misinformatie' die met internet mogelijk blijkt (maar door weinigen voorzien werd bij het begin van deze revolutie).

Ten vijfde: we zien inmiddels de eerste voorbeelden van de manier waarop Big Data kunnen bijdragen aan het vinden van die beleidsdoelen die, als ze gerealiseerd zouden worden, maatschappelijk het 'meest' effectief zijn. Dit werk gebeurt tegen de achtergrond van de studie van Kleinberg et al. (2015) naar het voorspellen van beleidsproblemen. Zij hebben deze benadering aan de hand van een eenvoudig droogte-/regenprobleem toegelicht.

> 'One policy maker facing a drought must decide whether to invest in a rain dance to increase the chance of rain. Another seeing clouds must decide whether to take an umbrella to work to avoid getting wet on the way home. Both decisions could benefit from an empirical study of rain. But each has different requirements of the estimator. One requires causality: do rain dances cause rain? The other does not, needing only prediction: is the chance of rain high enough to merit an umbrella?'

Andini et al. (2018) hebben tegen deze achtergrond *machine learning* toegepast om te zien of de keuze van de groep mensen die in Italië in 2014 in aanmerking kwam voor een 'tax credit' van 80 euro per maand, wel de meest aangewezen groep was om het beleidsdoel 'to boost household consumption', te realiseren. De groep die door de politiek was uitgekozen, betrof werknemers en anderen met een bruto jaarinkomen tussen € 8,145 en € 26,000 maar dat bleek niet de 'beste' groep te zijn (in termen van de te bereiken en bereikte resultaten). Hun conclusie was dat 'the effectiveness of the program would have significantly increased if the beneficiaries had been selected *according to a transparent and easily interpretable ML algorithm*'.[8] Dat algoritme hebben de onderzoekers ontwikkeld. Daarmee konden ze berekenen hoeveel groter het effect was geweest van deze interventie als de selectie van de doelgroepen/het beleidsprobleem *op díe manier* was vastgesteld in plaats van volgens de (klassieke) werkwijze (dus zonder machine learning).

5 Vraag 2: Wat heeft evaluatieonderzoek te bieden als het gaat om het analyseren/onderzoeken van de betrouwbaarheid, validiteit en enkele andere aspecten van Big Data en AI?

Ging het in het voorafgaande over hoe met behulp van BD/AI de kwaliteit, bruikbaarheid en relevantie van evaluatieonderzoek bevorderd kan worden, in déze paragraaf draai ik het om: *wat heeft evaluatieonderzoek te bieden als het gaat om het analyseren/onderzoeken van de betrouwbaarheid, validiteit en enkele andere aspecten van Big Data en AI?*

Enkele voorbeelden uit de praktijk die het belang van aandacht voor *evaluaties van BD/AI* duidelijk maken, beschrijf ik in box 1.

Box 1: Belang van evaluatieonderzoek naar Big Data en AI

Watson's IBM-rol bij het diagnosticeren van kanker werd door sommigen als behoorlijk belangrijk en overtuigend neergezet, maar werd door anderen gekritiseerd.[9] Een van de kritiekpunten was dat er, zonder dat daar helder over gecommuniceerd was, niet altijd 'echte' patiëntgegevens gebruikt werden, maar hypothetische. Vragen die een evaluator in zo'n geval stelt: Wat gebeurde waar, wanneer en waarom? Op grond van welke assumptie(s) werd gemeend dat deze aanpak kon? Wat zijn de gevolgen voor de kwaliteit (waaronder de validiteit) van de diagnoses?

Bias problemen bij predicties van bijvoorbeeld criminaliteit door politieorganisaties. Schuilenberg (2016): 'In de algoritmen zitten vooronderstellingen ingesloten omdat de variabelen op basis waarvan deze algoritmen zijn opgesteld, altijd keuzes zijn van ontwikkelaars, analisten en beleidsmakers. Technisch uitgedrukt houdt dit in dat er altijd een bias zit in de software die wordt gebruikt om criminaliteit te voorspellen. Het meest duidelijk kwam deze bias naar voren in het onderzoek van het journalistenplatform ProPublica naar risicoprofielen van de Amerikaanse justitie.' (bron: https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say)

Legacy problemen zijn de problemen die ontstaan door de erfenis van (oude en soms inadequate) ICT-systemen. Zij moeten ook worden gedetecteerd.

Assumpties waarop AI en het gebruik ervan gebaseerd zijn, zijn ook een object van evaluatieonderzoek. Welke zijn dat eigenlijk? Wat is hun status: gaat het om geloof, hoop of liefde of is er evidentie die de assumpties ondersteunt (of juist niet)? En: waar komen de assumpties vandaan?

Mission creep: gedaan met de beste bedoelingen kan het werken met Big Data/AI ook het verschijnsel oproepen, dat bepaalde activiteiten die bedoeld zijn om specifieke doelen te bereiken, onbewust en soms 'onzichtbaar' overgaan/ingezet worden om andere, bredere en verder weg gelegen doelen te realiseren. De wijkgemeente Gladsaxe in Kopenhagen experimenteerde bijvoorbeeld met een systeem dat algoritmes zou gebruiken om kinderen met een risico op misbruik te identificeren (via een vlagje in het bestand). Maar het bleek dat de 'bevlagde' families door de

gemeentelijke autoriteiten tevens getarget konden worden voor de uitrol van vroege interventies en uiteindelijk zelfs voor gedwongen verwijdering uit de buurt.[10]

In een artikel over hoe AI geëvalueerd wordt en zou moeten worden, onderscheidt Hernández-Orallo (2017)[11] *verschillende 'soorten' evaluaties*. De eerste is het vaststellen of de taak die met AI bereikt moet worden, bereikt wordt ('task-oriented'). De tweede is 'ability-oriented evaluation', waarbij nagegaan wordt wat de cognitieve capaciteiten van een AI(-benadering) zijn. En de derde is de mate waarin problemen opgelost worden met behulp van AI. De auteur geeft diverse voorbeelden van evaluatie-aanpakken om te meten hoe AI 'scoort', variërend van Turing-tests (en nieuwe varianten), wedstrijden waarmee de 'believability' van bijvoorbeeld games gemeten wordt en problems benchmarks (bijvoorbeeld gericht op het ontdekken van plagiaat, het veilig laten rijden van auto's met behulp van machine learning). De auteur behandelt ook het probleem dat er verschil is tussen AI-systemen (zoals robots) en AI-componenten (zoals specifieke algoritmes, methoden of 'tools'). Conclusies zijn dat het veld van de AI-evaluatie gefragmenteerd is (p. 431), er behoefte is aan krachtigere theorie(en) en dat 'AI requires an accurate, effective, non-anthropocentric, meaningful and computational way of evaluating its progress, by evaluating its artefacts' (p. 439). Opvallend is het ontbreken van verwijzingen naar (gedrags)wetenschappelijk evaluatieonderzoek en -methodologie.

Evaluatieonderzoek zou zich naar mijn oordeel met deze en vergelijkbare vraagstukken moeten bezighouden. Daar is nog nauwelijks sprake van. Wel zijn er interessante ontwikkelingen, die echter (groten)deels buiten de 'traditionele evaluatieprofessie' ontwikkeld worden. Daarvan noem ik er een paar.

Tijdens het jaarcongres van het ECP|Platform voor de Informatie Samenleving (15 november 2018) is een *AI Impact Assessment* (tool) gelanceerd. 'Deze AI-IA helpt bedrijven AI verantwoord in te zetten. Aan de hand van een (acht)stappenplan maken bedrijven inzichtelijk welke juridische en ethische normen een rol spelen bij de ontwikkeling en inzet van AI-toepassingen, en welke afwegingen ten grondslag liggen aan keuzes en besluiten.' Verschillende mogelijke toepassingsmogelijkheden worden genoemd, zoals 'het automatisch beoordelen of wetgeving op iemand van toepassing is. Dit kan misschien wel sneller en beter met een slim algoritme, maar een ambtenaar van vlees en bloed heeft de ruimte en de plicht om vanuit een fundamentele waarde als "menswaardigheid" te handelen en individuele afwijkende situaties te beoordelen. In de AIIA worden dergelijke ethische en juridische vraagstukken getackeld, zodat het algoritme en de ambtenaar volgens menselijke waarden kunnen samenwerken.'

Het is de vraag hoe zo'n assessment in de praktijk uitpakt: wat zijn de onderliggende assumpties, wat is de waarde van de informatie die verzameld wordt, zal het vooral een 'accountability-dingetje' worden om de behoefte aan transparantie op papier te bevredigen, zonder dat inhoudelijk geanalyseerd wordt hoe de AI 'werkt', of zal het gaan om inhoudelijke analyses?

Ook zijn er in de wetenschappelijke literatuur voorbeelden te vinden van casus waarin evaluaties van algoritmes al plaatsvinden. Box 2 presenteert een aantal casus-beschrijvingen.

Box 2: Evaluaties van algoritmes: vijf praktijkvoorbeelden

Voorbeeld 1 is een onderzoek naar de rol van algoritmes die gebruikt worden om gecompromitteerde sociale-media-accounts te detecteren, zoals gegijzelde accounts. De onderzoekers bouwden voort op COMPA ('Detecting Compromised Accounts on Social Networks') en evalueerden het (aangepaste) detectie-programma (Trang et al., 2015).

Voorbeeld 2 betreft de evaluatie van sensors die gebruikt worden bij het meten van het alcoholpercentage in het bloed (bij verkeerscontroles). Er bestaan verschillende sensoren maar daarop is wat betreft precisie, validiteit en beïnvloedbaarheid door gedrag van de 'gemetenen' het een en ander aan te merken. De onderzoekers beoogden 'less obtrusive sensor systems for breath alcohol screening' te ontwikkelen. Dit is van belang voor bijvoorbeeld de politie in het kader van verkeersongelukken of voor andere veiligheidsofficials, als het gaat om het 'alcoholvrij' betreden van werkplekken zoals fabrieken, vliegtuigen of medische faciliteiten (Ljungblad et al., 2016).

Voorbeeld 3 richt zich op predictive policing, reconstrueert en analyseert tien assumpties die daaraan ten grondslag liggen, vergelijkt wat in populaire media over de effectiviteit van (verschillende) modaliteiten van deze vorm van politiewerk gemeld wordt met wat (zeer kleine aantallen) wetenschappelijke effect-evaluaties rapporteren, en beschrijft welke uitdagingen er liggen op het vlak van de 'accountability' (Bennett Moses & Chan, 2018). Voorbeelden van assumpties die tegen het licht worden gehouden, zijn:

Assumptie 1: Data used accurately reflect reality.

Assumptie 2: The future is like the past.

Assumptie 3: Irrelevance of omitted variables.

Assumptie 4: Algorithms are neutral.

Assumptie 5: Data analytics does not unjustly discriminate.

Assumptie 6: Primacy of place.

Dit artikel is te beschouwen als een toepassing van *theory-driven evaluatieonderzoek* in de digitale wereld.

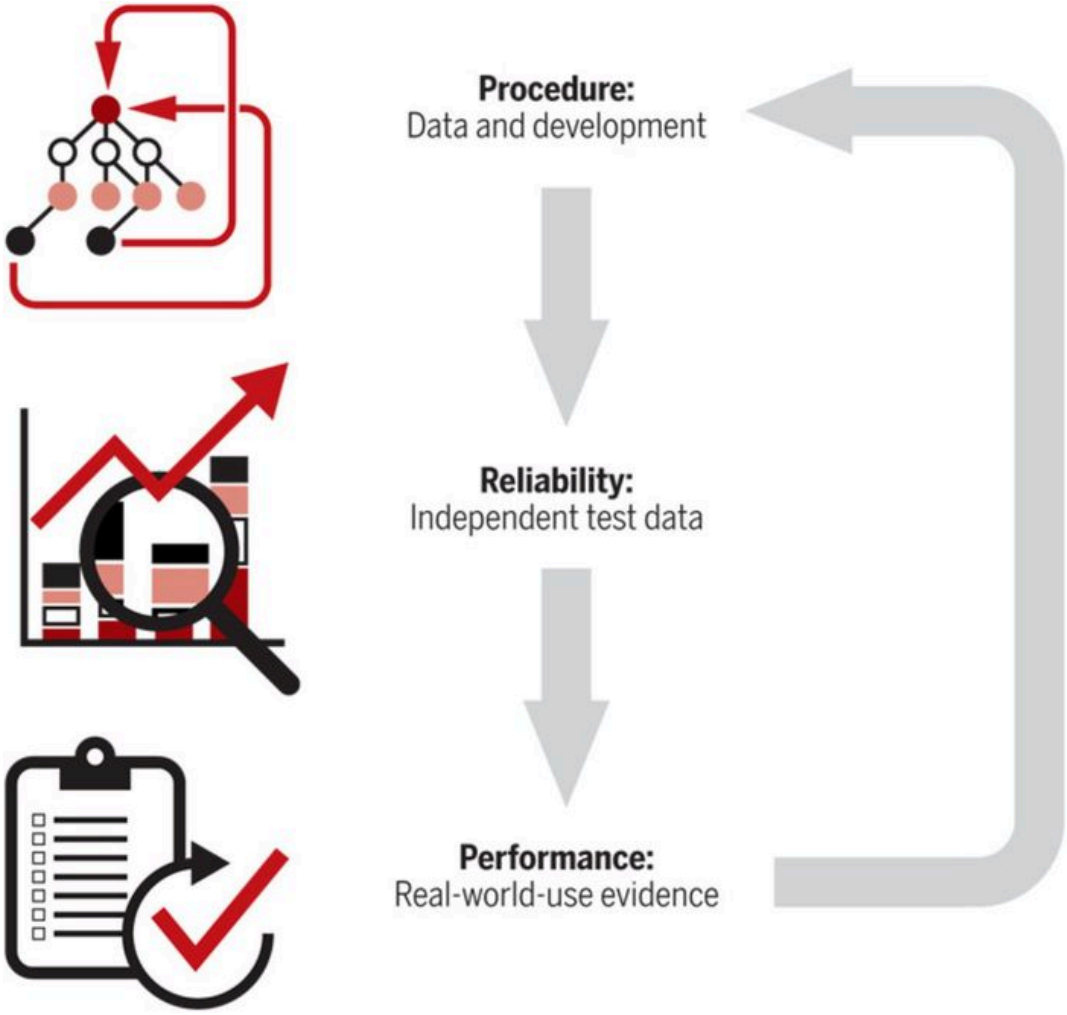
Het *vierde voorbeeld* is van Kanagasingam et al.[12] Oogproblemen werden gediagnosticeerd door AI en door een oogarts. Het ging om patiënten met diabetes die werden gezien in een praktijk voor eerstelijnsgezondheidszorg met vier artsen in West-Australië tussen december 2016 en mei 2017. In totaal stemden 193 patiënten in met het onderzoek en om foto's van het netvlies van hun ogen te laten nemen. 386 beelden werden geëvalueerd door zowel het op AI gebaseerde systeem als door een oogarts.

Ten slotte, vanuit de *wetenschapsfilosofie* wordt eveneens aan dit vraagstuk bijgedragen, zoals Chapman's metablog laat zien.[13]

Mogelijkerwijze is de aandacht voor het evalueren van algoritmes in de medische wereld thans het verst ontwikkeld. Ik grijp terug op een artikel van Nicholson Price (2018), waarin de uitdagingen van de invoering van nieuwe machine-learning technieken in termen van validiteit, regulering en integratie (met bestaande praktijken) uiteengezet werden en een model gepresenteerd is hoe (evaluatie)onderzoek naar de AI-black-box (en de gevolgen van invoering van mede daarop gebaseerde interventies) wetenschappelijk verantwoord én praktisch te doen is. Zijn centrale vraag is hoe zeker providers, ontwikkelaars, regelgevers, verzekeraars, artsen en patiënten ervan kunnen zijn dat de gebruikte algoritmes accuraat en bruikbaar zijn. Voor accuraatheid is de mate te lezen waarin het benodigde databeheer aan de maat is en de analyses die ermee gedaan worden, correct zijn. Bruikbaar betreft niet alleen de mate waarin de doelgroepen die Nicholson Price onderscheidt, zoals providers en 'regulators', het nut ervan zien, maar ook de mate waarin de algoritmes bijdragen aan effectievere diagnoses en interventies dan gerealiseerd (zouden) zijn in situaties waarin daar niet mee gewerkt wordt. Nicolson Price zet vervolgens de uitdagingen waar de praktijk voor staat op een rij. Eén uitdaging is dat er gebruik wordt gemaakt van black-box algoritmes, die niet zelf verklaringen aandragen voor wat ze vinden. Evenmin benoemen bepaalde algoritmes expliciete (inhoudelijke) relaties tussen variabelen. 'For example, a neural network trained to identify tumors can identify them but uses opaque hidden layers to do so.' Ten derde is er sprake van algoritmes die zo complex zijn dat ze 'het' menselijk begrip te boven gaan. Hij attendeert ook op het hoge gehalte aan plasticiteit van (bepaalde) algoritmes: ze veranderen juist dóór het gebruiken van (nieuwe) data, die afkomstig zijn uit (veel nieuwe) meetmomenten. Samengevat ziet de auteur 'opacity' en 'plasticity' als de belangrijkste uitdagingen voor de evaluatie van algoritmes in de medische wereld. Ten slotte schetst

hij een driestappenmodel waarin hij weergeeft hoe (evaluatie)onderzoek op dit gebied is uit te voeren (figuur 1).

Figuur 1 Evaluatie & validatie van black-box algoritmes (Price, 2018)



Stap 1 is het nagaan of er sprake is van basiskwaliteit in het gebruik van (training én test) data en procedures bij het ontwerpen van de algoritmes. *Stap 2* is het meten van de performance van het algoritme met gebruikmaking van data die niet door het algoritme zijn verzameld, respectievelijk geanalyseerd (Rajkomar et al. 2019).[14] En *stap 3* is de evaluatie van het gebruik van het algoritme in de 'echte' wereld ('real-world-use evidence'). 'This third and most important step of validation applies to all sorts of black-box algorithms: they should be continuously validated by tracking successes and failures as they are actually implemented in health care settings.'

Wat is de relevantie voor beleidsmakers en onderzoekers van een dergelijke benadering? Het antwoord is eenvoudig: vergelijkbare vragen zijn al op diverse beleidsterreinen zoals justitie en veiligheid, sociale zekerheid, en verkeer en vervoer gesteld. Ik verwacht dat met de groei van het gebruik van AI bij tal van *andere* organisaties die interveniëren in de samenleving dergelijke vragen óok op de agenda komen. Denk aan het ontwikkelen van gedragsinterventies op het terrein van gezondheid, voedsel en bewegen, aan (justitiële) sancties,

aan voorlichtingsactiviteiten en 'moral persuasion' bij milieu- en klimaatbeleid en aan processen in het bedrijfsleven en de marketing. Het is dienstig daarop goed voorbereid te zijn, opdat de samenleving geïnformeerd kan worden over de manier waarop *validiteit, veiligheid en 'believability'* van het werken met AI/machine learning gewaarborgd is en wat de resultaten van het ermee werken zijn.

6 Enkele conclusies

Evaluatieonderzoek van wet- en regelgeving, interventies, programma's en andere beleidsmaatregelen is enerzijds erbij gebaat als er *meer gebruik gemaakt wordt* van BD/AI. In dit artikel heb ik verschillende toepassingsmogelijkheden gepresenteerd. Langere tijd was de aandacht in de wereld van evaluatoren hiervoor overigens (zeer) gering.

Tegelijkertijd dient er juist ook (meer) aandacht uit te gaan naar *de ándere kant* van de medaille:

- de assumpties die aan BD/AI ten grondslag liggen (inclusief het 'black box'-probleem);
- de validiteit, veiligheid en geloofwaardigheid van algoritmes;
- de bedoelde en onbedoelde consequenties van het gebruik ervan; én
- de vraag of de claims dat digitale interventies die mede gebaseerd zijn op BD/AI, effectief (of *effectiever* zijn dan andere) onderbouwd en correct zijn.

Het is te verwachten dat de rol van BD/AI bij het ontwikkelen én invoeren van (overheids)beleid eerder groter dan kleiner gaat worden. Denk aan robotica, domotica, empathische AI en 'Quantified Self-ontwikkelingen', maar ook predictiemodellen op diverse gebieden waaronder de gezondheidszorg, verkeer en vervoer maar ook het terrein van het recht en de wetgeving (Custers & Leeuw, 2017). Dat vereist een actieve en hoogwaardige aandacht van beleidsevaluatoren voor deze 'hybride werkelijkheid'.

Literatuur

Andini, M. et al. (2018). Targeting with machine learning: An application to a tax rebate program in Italy. *Journal of Economic Behavior and Organization, 56*, 86-102.

Bader, M. et al. (2017). The promise, practicalities, and perils of virtually auditing neighborhoods using Google Street View. *Annals,*

669, 18-40.

Bail, C. (2014). The cultural environment: Measuring culture with big data. *Theory and Society, 43*(3-4), 465-482.

Bamberger, M. (2016). *Integrating big data into the M and E of development programmes*. New York: Global Pulse.

Bennett Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *Policing and Society, 28*(7), 806-822.

Custers, B., & Leeuw, F.L. (2017). Legal big data: Toepassingen voor de rechtspraktijk en juridisch onderzoek. *Nederlands Juristenblad, 34*, 2449.

Domino, P. (2015). *The master algorithm*. London: Penguin.

Forss, K. & Norén, J. (2017). Using big data for equity-focused evaluation: Understanding and utilizing the dynamics of data ecosystems. In G.J. Petersson & J.D. Breul (Eds.), *Cyber society, big data and evaluation* (pp. 171-191). Rutgers: Transaction Press.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*, 137-144.

Hernández-Orallo, J. (2017). Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement. *Artif Intell Rev, 48*, 397-447.

Høljund, S. et al. (2017). The current use of big data in evaluation. In G.J. Petersson & J.D. Breul (Eds.), *Cyber society, big data and evaluation* (pp. 35-61). Rutgers: Transaction Press.

Kleinberg, J. et al. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491-495.

Klous, S., & Wielaard, N. (2014). *Wij zijn Big Data: De toekomst van onze informatiesamenleving*. Amsterdam: Business contact.

KNAW. (2018). *Advies: Big data in wetenschappelijk onderzoek met gegevens over personen*. Amsterdam.

Kogel, C.H. de, & Cornet, L.J.M. (2016). Toepassingsmogelijkheden van Quantified Self-data: Enkele voorbeelden uit de forensisch psychiatrische praktijk. *Justitiële verkenningen, 42*(1), 79-94.

Koops, B.J. (2018). Privacy spaces. *West Virginia Law Review, 121*(2), 611-665.

Leeuw, F.L., & Leeuw, H.B.M. (2012). Cyber society and digital policies: Challenges to evaluation? *Evaluation, 18*(1), 111-127. https://doi.org/10.1177/1356389011431777

Leeuw, H.B.M. (2017). *Punish, seduce or persuade: An empirical assessment of anti-piracy interventions*. PhD Maastricht University. Den Haag: Boom Juridisch.

Ljungblad, J. et al. (2016). Development and evaluation of algorithms for breath alcohol screening. *Sensors, 16*, 469. doi:10.3390/s16040469

Nicholson Price, W. (2018). Big data and black-box medical algorithms. *Science Translational Medicine, 10*, eaao5333.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*, 1347-1358. doi:10.1056/NEJMra1814259

Schuilenburg, M. (2016). Predictive policing: de opkomst van een gedachtenpolitie? *Ars Aequi, 65*(12), 931-936.

Stephens, Z. et al. (2015). Big data: astronomical or genomical? *PLoS Biol, 13*(7), e1002195. doi:10.1371/journal.pbio.1002195

Trang, D. et al. (2015). Evaluating algorithms for detection of compromised social media user accounts. *Second European Network Intelligence Conference*, 978-1-4673-7592-4/15. doi:10.1109/ENIC.2015.19

Watmough, G. et al. (2019). Socioecologically informed use of remote sensing data to predict rural household poverty. *PNAS, 116*(4), 1213-1218.

Willemsen, F., & Leeuw, F. (2017). Big data, real-world events, and evaluations. In G.J. Petersson & J.D. Breul (Eds.), *Cyber society, big data and evaluation* (chapter 5). Rutgers: Transaction Press.

Noten

1 Koops operationaliseert het overkoepelende begrip ‘informational privacy’ in acht typen, waaronder spatial privacy, bodily privacy en behavioral privacy.

2 Denk aan ethische aspecten van digitale beïnvloeding, het

verschijnsel van de eigendomsrechten van organisaties die met (Big) Data werken, enzovoort.

3 'De term Quantified Self is in 2007 in de Verenigde Staten bedacht door Gary Wolf en Kevin Kelly, die ook een gelijknamige website introduceerden. Het gaat om "zelfmeting", ook wel "self-tracking" genoemd, en de grote hoeveelheid gegevens die in continumetingen afgetapt kan worden met behulp van "wearables" (zoals fitbit-achtige horloges), "carrieables" (in mobiele apparaten zoals een telefoon), "insideables" (bijvoorbeeld een chip die in het lichaam ingebracht wordt) en "domotica" (apparaten voor in huis of kantoor zoals lampen die reageren op beweging).'

4 Zie www.pietdaas.nl/beta/pubs/pubs/Big_data_zomerrelatiemagazine.pdf.

5 Stephens-Davidowitz gooide een knuppel in het hoenderhok van een bijna 100 jaar oud debat in de gedragswetenschappen met zijn boek: *Everybody lies* (2017). Het debat over verschillen tussen wat mensen vinden ('attitudes') en wat ze doen ('acties') gaat terug tot een artikel van Richard T. LaPiere uit 1934 in *Social Forces* (13, 230-237). Zie ook S. Donaldson en E. Grant Valone in *Journal of Business and Psychology* (2002, 17), 'Understanding self-report bias in organizational behavior research', waarin zij oorzaken analyseren van deze bias, die erop neerkomt dat '[persons] want to respond in a way that makes them look as good as possible'.

6 Op microniveau komen de zgn. A/B-experimenten die internetbedrijven frequent doen, iets meer in de buurt.

7 In zijn 'Algoritmen kunnen toveren' verwees Haroon Sheikh, *NRC Handelsblad*, 23 november 2018, naar deze uitspraak.

8 Zie voor een populairdere samenvatting: https://voxeu.org/article/effective-policy-targeting-machine-learning.

9 Zie www.bloomberg.com/opinion/articles/2018-08-24/ibm-s-watson-failed-against-cancer-but-a-i-still-has-promise (geraadpleegd 21 juli 2019).

10 Zie https://foreignpolicy.com/2018/12/25/the-welfare-state-is-committing-suicide-by-artificial-intelligence/ (geraadpleegd 26 december 2018).

11 Zie ook zijn *The Measure of All Minds. Evaluating Natural and Artificial Intelligence*, Cambridge University Press, 2017.

12 Zie

https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2703944.

13 Meaningness is a hypertext book (in progress), plus a 'metablog' that comments on it. De redactie is in handen van David Chapman. Zie deze blog voor informatie over hoe progressie in de (kwaliteit van) AI 'gemeten' kan worden: https://meaningness.com/metablog/artificial-intelligence-progress.

14 Rajkomar et al. beschrijven op toegankelijke wijze (in figuur 1 van hun artikel) stap voor stap en concreet hoe machine learning werkt in het medisch bedrijf.