Tussen data en theorie

Goudsmit, Jeroen; Teuwen, Jonas

published in
Tijdschrift voor Toezicht
2020

DOI (link to publisher)
10.5553/TvT/187987052020011001008

document version
Publisher's PDF, also known as Version of record

Essay

Tussen data en theorie

Het venijn zit in de aard

Jeroen Goudsmit en Jonas Teuwen*

Algoritmen helpen om op grote schaal beslissingen te nemen. Het is echter lastig om achteraf toe te zien op de kwaliteit van deze beslissingen. Toezicht zou zich met name moeten richten op het wordingsproces van algoritmes: de stappen die genomen worden om te komen van probleemomschrijving tot een geïmplementeerd algoritme. In dit proces worden immers de principiële keuzes gemaakt die bepalend zijn voor de manier waarop het algoritme zal handelen en de wijze waarop het uitwerking heeft op de maatschappij. Door deze keuzes expliciet en onder de juiste overwegingen te maken verkleint het risico op misdragingen. Toezichthouders en ontwikkelaars van algoritmen tezamen kunnen hier een handreiking voor opstellen.

Inleiding

De recente ontwikkelingen binnen artificiële intelligentie (AI) zullen door niemand onopgemerkt gebleven zijn. De drijvende kracht achter de recente ontwikkelingen is de enorme groeispurt die *machine learning*, en dan met name *deep learning* heeft doorgemaakt. Deze disciplines houden zich bezig met het ontwikkelen van algoritmes geijkt aan de hand van data, waarbij *deep learning* zich toespitst op algoritmes gemodelleerd naar neurale netwerken. Het zijn niet zozeer de theoretische inzichten die het tot hier gebracht hebben, maar de combinatie van een enorme toename van datacollectie én een sprong in de bredere beschikbaarheid van de benodigde *hardware* voor de complexe berekeningen.

Voor de recente (her)intrede van *machine learning* waren er een aantal rekentaken die voor een mens relatief eenvoudig leken, maar voor een computer een ondoenlijke taak bleken. In 1997 versloeg de computer Deep Blue tot ieders verbazing de toenmalige wereldkampioen schaken Garry Kasparov met 3½–2½. Dit systeem gebruikte een combinatie van *brute force* rekenwerk met een door een mens bedacht heuristisch evaluatiesysteem.[1] Tot voor kort bleek het spel Go door de astronomische hoeveelheid van mogelijke posities onaantastbaar voor een dergelijke heuristische aanpak. *Deep learning* heeft hier echter verandering in gebracht. Zo is recent in het vakblad *Nature* een AI-algoritme beschreven dat de wereldkampioen Go, Lee Sedol, met 4-1 versloeg.[2]

Het blijft niet bij spelletjes. Ook de steeds beter wordende machinevertalingen zijn gebaseerd op *deep learning*[3] Net als bij zelfrijdende auto's zitten hier vaak de grote AI-spelers van de wereld zoals Facebook, Uber en Google achter. Ook binnen de geneeskunde zijn er al algoritmes die het net zo goed doen als een ervaren

* Dr. J.P. Goudsmit is kerndocent Compliance en Integriteit Management bij de Vrije Universiteit Amsterdam. Dr. J.J.B. Teuwen is Associate Staff Scientist bij het Nederlands Kanker Instituut en tenure-track researcher bij de afdeling radiologie, nucleaire geneeskunde en anatomie aan het Radboud Universitair Medisch Centrum.

1. Heuristieken zijn vuistregels die helpen om complexe problemen behapbaarder te maken ten koste van precisie en volledigheid. Voor een omschrijving van de AI achter Deep Blue verwijzen we naar R.E. Korf, 'Does Deep-Blue use AI?', *AAAI Technical Report* 1997.
2. D. Silver e.a., 'Mastering the game of Go without human knowledge', *Nature* 2017, 550(7676), p. 354-359.
3. M. Ranzato, G. Lample en M. Ott, 'Unsupervised machine translation: A novel approach to provide fast, accurate translations for more languages', Bericht op Facebook Engineering, 31 augustus 2018, beschikbaar via https://engineering.fb.com/ai-research/unsupervised-machine-translation-a-novel-approach-to-provide-fast-accurate-translations-for-more-languages/.

doi: 10.5553/TvT/187987052020011001008

medisch specialist. Zo heeft een Nederlandse start-up recent haar AI-algoritme dat gebruikt kan worden om mammogrammen te beoordelen vergeleken met 101 ervaren borstradiologen. Hieruit bleek dat haar algoritme het in deze test minstens even goed doet als de gemiddelde radioloog.[4] Op dit moment zijn er meer dan 40 producten gecreëerd door AI-startups die door de Amerikaanse Food and Drug Administration goedgekeurd zijn om klinisch te gebruiken met soortgelijke claims.

Het is aan toezichthouders om te bepalen hoe ze met deze ontwikkelingen om willen gaan. Vanuit prudentieel toezicht heeft De Nederlandsche Bank afgelopen zomer de discussie op verantwoord gebruik van AI in de financiële sector aangezwengeld.[5] Ze introduceerde hier zes principes samengevat in het acroniem 'SAFEST': *soundness*, *accountability*, *fairness*, *ethics*, *skills* en *transparency*. Met deze principes wordt richting gegeven aan de belangrijke thema's bij het inzetten van AI-algoritmes. Daarnaast heeft de Autoriteit Persoonsgegevens vanuit de Algemene verordening gegevensbescherming (AVG) een wettelijke basis voor toezicht op AI, en meldt deze toezichthouder zich te richten op het vormgeven van een stelsel voor toezicht op AI en algoritmes waarin persoonsgegevens worden gebruikt.[6] De Europese Commissie heeft in juni 2018 de AI *High Level Expert Group* aangesteld. Deze groep heeft als doelstelling de implementatie van de Europese Strategie op AI te ondersteunen. Hiertoe heeft ze aanbevelingen gedaan over de inzet van AI, met nadruk op menselijk toezicht en non-discriminatie.[7]

De crux is het borgen van de uitkomst van algoritmische beslissingen. Dit begint al vroeg. In dit essay betogen we dat het wordingsproces van algoritmen – van de eerste probleemstelling tot de implementatie van het daadwerkelijke algoritme – het moment is om in te grijpen. De kloof tussen technici en bestuurders moet hiervoor wezenlijk geslecht worden, omdat beide groepen zonder wederzijds begrip en dialoog niet de nodige keuzes weloverwogen kunnen maken.

Te vaak wordt de nadruk gelegd op het controleren van deze kwaliteit wanneer het algoritme reeds geïmplementeerd is in een organisatie. Hier resten dan slechts twee opties:

1. Het één voor één controleren van de beslissingen door een mens als buffer in het proces te plaatsen met als optie in te grijpen bij elke handeling.
2. Het controleren van de beslissingen voor een (deel)populatie door het proces periodiek te evalueren met als optie hierna in te grijpen.

We betogen dat de betere optie is om de kwaliteit te borgen tijdens het wordingsproces. Dit is immers het moment waarop de wezenlijke keuzes gemaakt worden: hoe het model werkt en hoe het ingebed is in het proces dat het dient.

Eerst analyseren we het begrip 'algoritme' en het gewicht dat hieraan is komen te hangen. Vervolgens bespreken we voorgenoemde twee opties en beargumenteren we dat deze omwille van menselijke beperkingen en de te grote potentie voor misdragingen ontoereikend zijn. Daarna behandelen we enkele van de keuzes die gepaard gaan bij het ontwikkelen van algoritmes en betogen we dat deze meer principieel dan technisch van aard zijn, waarmee we beargumenteren dat bestuurlijke betrokkenheid cruciaal is. Ten slotte concluderen we dat verbinding nodig is tussen onderzoekers en ontwikkelaars van algoritmes en de beleidsmakers en toezichthouders die dezen in goede banen moeten leiden. Er dient gezamenlijk een praktische handreiking gemaakt te worden, waarin houvast gegeven wordt aan praktikanten in het toepassen van de meer algemene principes over verantwoord AI-gebruik.

Algoritmes nemen we serieus

Wanneer een exacte wetenschapper denkt aan een algoritme, dan denkt hij aan een stapsgewijze procedure; een wiskundig recept om iets te berekenen. Voorbeelden gaan terug tot ver voor de jaartelling, zoals de simpele staartdeling of Euclides' algoritme voor het berekenen van de grootste gemene deler. Het spreekt voor de mate waarin dit begrip vanzelfsprekend is dat het pas in 1936 door opeenvolgend Alonzo Church en Alan Turing zorgvuldig werd gedefinieerd.[8]

In het dagelijks taalgebruik wordt de term algoritme vaak veel breder gebruikt dan de puur wiskundige notie van Church en Turing.[9] Zo worden verwijzingen naar algoritmes gebruikt als synecdoche voor het bredere proces waar ze onderdeel van zijn. Enerzijds maakt dit het gesprek gemakkelijker, omdat je zo de details over het vormen van het algoritme en de manier waarop je het inzet onder het tapijt kan vegen. Anderzijds vertroebelt dit de discussie, omdat dit precies is waar de problematiek zit. Het is belangrijk te zien dat het algoritme *per se* niet het probleem is, maar juist alles er omheen.

4. A. Rodriguez-Ruiz e.a., 'Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists', *JNCI: Journal of the National Cancer Institute* 2019, nr. 9, p. 916-922.
5. De Nederlandsche Bank, *General principles for the use of Artificial Intelligence in the financial sector*, 2019.
6. Autoriteit Persoonsgegevens, *Focus AP 2020-2023: Dataprotectie in een digitale samenleving*, 2019.
7. Europese Commissie, *Building Trust in Human Centric Artificial Intelligence*, 2019, beschikbaar via https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence.
8. In 1928 stelden David Hilbert en Wilhelm Ackermann het *Entscheidungsproblem*, wat vraagt om een algoritme te formuleren dat logische uitspraken analyseert en bestempelt als waar dan wel onwaar. Om te kunnen bewijzen dat dit theoretisch onmogelijk is, was het nodig om dit intuïtieve begrip eerst zorgvuldig te definiëren. De twee resulterende noties vormen de basis van de moderne wiskundige studie van berekenbaarheid.
9. T. Gillespie, 'Algorithm', in: *Digital Keywords: A Vocabulary of Information Society and Culture*, Princeton Studies in Culture and Technology 2016, p. 18-30.

Evenzo kwalijk is het gebruik van de term algoritme als talisman – een magisch object dat gevrijwaard is van alle kritiek. Dit taalgebruik speelt in op ons vertrouwen in het gekwantificeerde en geeft oneigenlijke legitimiteit aan een uitspraak puur en alleen omdat een algoritme betrokken was. Ook dit vertroebelt de discussie. Een uitkomst controleer je op zijn merites, niet op wie of wat de uitkomst bepaald heeft.

Mens als onvolmaakte buffer tegen algoritmes

Wanneer een algoritme beslissingen neemt dan kunnen we terecht twijfelen aan de kwaliteit van deze beslissingen. Organisaties die algoritmes op enige wijze inzetten in een proces dat leidt tot beslissingen, doen er goed aan om hierop te reflecteren. Je zou simpelweg een mens kunnen vragen de uitkomsten van het algoritme één voor één te controleren. In veel situaties is dit niet effectief, werkt het averechts en leidt het er enkel toe dat een lagere kwaliteit van beslissingen worden verhuld.
Mensen zijn berucht slecht in het wegen van meerdere stukken informatie. Zo leidt *anchoring* ertoe dat we volkomen irrelevante informatie meewegen en zodoende onze beslissingen oneigenlijk bijstellen.[10] Afwijken van deze *anchors* is zelfs voor rechters een sterke uitdaging.[11] Dit maakt menselijke controle in het proces ineffectief.

Sterker nog, mensen maken de beslissingen zelfs slechter. Onderzoek naar *bounded rationality* laat zien dat menselijke beslissingen niet puur rationeel zijn. Zo is ons denken onderhevig aan *biases* die onze beslissingen beïnvloeden waardoor niet-rationele keuzes gemaakt worden.[12] Zo ontstaat de voorkeur over te gaan op actie daar waar dit de norm is, zelfs wanneer dit leidt tot slechtere uitkomsten.[13] Zodoende werkt een mens als buffer soms averechts.
Daarnaast hebben mensen ook persoonlijke doelstellingen die haaks kunnen staan op die van het proces dat ze dienen.[14] Dit fenomeen is ook te beschouwen binnen het gebruik van algoritmes binnen de radiologie. Zo blijkt dat wanneer een *deep learning* systeem gebruikt wordt om ziektes te zoeken in thorax röntgenfoto's, het algoritme het op zichzelf beter doet dan wanneer gecombineerd met een ervaren en gespecialiseerd radioloog.[15] Een mens als buffer maakt de beslissingen dus zeker niet vanzelfsprekend beter.

Periodieke toetsing als zwakke stok achter de deur

Als het kalf verdronken is, dempt men de put. Droevig voor het eerste kalf, maar de uitwerking is nog te overzien. Niet zo bij algoritmes, die juist ingezet kunnen worden om in korte tijd op grote schaal ingrijpende beslissingen te nemen. Zodoende kan een algoritme al voor zijn periodieke keuring serieuze nadelige uitwerking hebben op de maatschappij en de betrokken organisatie. Daarnaast is een serieuze keuring nagenoeg onmogelijk tenzij er hiervoor randvoorwaarden zijn geschapen.
Er bestaan tal van voorbeelden waarin algoritmen in korte tijd veel invloed konden hebben. Denk hierbij aan *targeted advertising* waarbij advertenties heel specifiek gemaakt worden om een boodschap te brengen die goed resoneert bij de gebruiker, waar dit zonder *targeting* niet het geval zou zijn. Om verdere invloed te beperken heeft Twitter onlangs besloten om politieke advertenties niet op haar platform toe te laten.[16] Zelfs al wordt de invloed van dergelijke advertenties naderhand als onwenselijk vastgesteld, dan hebben ze al negatieve gevolgen gehad die lastig te herstellen zijn.
De negatieve uitwerking van een algoritmische beslissing op een persoon is lastig te bewijzen. Hoe kun je immers doorhebben dat het algoritme jou verkeerd behandeld heeft? Hierdoor is het niet vanzelfsprekend dat een algoritme dat mensen benadeelt ook veel klachten zal vangen. Bij een periodieke toetsing van het algoritme zal deze nadelige uitwerking dus ook niet snel opvallen. Andersom kan de uitwerking soms pijnlijk duidelijk worden, op zo'n manier dat het vertrouwen in de organisatie meteen geschaad is. Zo heeft de New York State Department for Financial Services slechts enkele dagen na het losbarsten van een discussie over mogelijke discriminatie op grond van geslacht in het algoritme voor creditcardaanvragen een onderzoek gestart.[17]
Tenslotte is het achteraf keuren van algoritmische beslissingen niet eenvoudig. Een dergelijke keuring heeft immers meer informatie nodig dan noodzakelijk is voor het uitvoeren van het algoritme. Om te toetsen of

10. Zie de paragraaf *Anchoring and Adjustments* in: A. Tversky en D. Kahneman, 'Judgment under Uncertainty: Heuristics and Biases', *Science* 1974, 185(4157), 1124 LP-1131.
11. T. Mussweiler en F. Strack, 'Numeric Judgments under Uncertainty: The Role of Knowledge in Anchoring', *Journal of Experimental Social Psychology* 2000, nr. 5, p. 495-518.
12. Dit geldt ook in het toezichtdomein, zie bijvoorbeeld R. Jansen en M. Aelen, 'Biases in toezicht: wat zijn het en hoe kunnen we ermee omgaan?', *TvT* 2015, afl. 1, p. 521.
13. Keepers in profvoetbal springen bijvoorbeeld disproportioneel vaak naar links of rechts, conform de norm, in plaats van dat ze in het midden blijven staan, zie M. Bar-Eli, O.H. Azar, I. Ritov, Y. Keidar-Levin en G. Schein, 'Action bias among elite soccer goalkeepers: The case of penalty kicks', *Journal of Economic Psychology* 2007, nr. 5, p. 606-621.
14. Zo zijn Amerikaanse rechters onderhevig aan prikkels die maken dat het gunnen van borgtocht voor hen persoonlijk meer risico's heeft dan voordelen, zie S.R. Wiseman, 'Fixing bail', *George Washington Law Review* 2016, nr. 2, p. 417-479.
15. E.J. Hwang e.a., 'Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs', *JAMA Network Open* 2019, nr. 3, e191095.
16. R. Wassens, 'Twitter legt beperkingen op aan advertenties rond gevoelige thema's', *NRC* 15 november 2019.
17. S. Mansoor, 'A Viral Tweet Accused Apple's New Credit Card of Being "Sexist." Now New York State Regulators Are Investigating', *Time* 11 november 2019, beschikbaar via https://time.com/5724098/new-york-investigating-goldman-sachs-apple-card/.

doi: 10.5553/TvT/187987052020011001008

het algoritme discrimineert op geslacht zul je moeten weten wat het geslacht is van de mensen over wie het algoritme beslist. Dit zal echter vaak niet opgeslagen worden op gronden van dataminimalisatie. Het niet opslaan of meenemen van dit soort informatie in het algoritme doet echter weinig om het risico op discriminatie langs deze as uit te sluiten, omdat geavanceerde algoritmes vaak in staat zijn het geslacht te schatten uit andere kenmerken in de data.[18]

Principiële keuzes onderliggend aan algoritmes

Een organisatie gebruikt niet zomaar een algoritme; dit wordt gedaan om een bepaald probleem op te lossen. Zo begint het proces om een algoritme te maken met een heldere doelstelling met bijbehorende succescriteria.[19] Het belang van het maken van volledige en weloverwogen keuzes in dit stadium is haast niet te onderschatten.
Het is verleidelijk te denken dat algoritmen waardevrij zijn. Dit is naïef. Bij het ontwerpen van een algoritme worden wezenlijke keuzes gemaakt over hoe het functioneert. Dit is niet enkel zo omdat de ontwerper daar zin in heeft; keuzes maken is een essentieel en theoretisch onvermijdelijk onderdeel van het maken van algoritmes.[20] In feite wordt met deze keuzes de basis gelegd voor elke beslissing waar het algoritme bij betrokken zal zijn.
De technische keuzes moeten in lijn zijn met de context die voortkomt uit de doelstelling en de bijbehorende succescriteria. Doorgaans zit er een groot gat tussen de abstractie waarin de bestuurlijke doelstellingen zijn geformuleerd en de wiskundige precisie waarmee de daaropvolgende technische keuzes gemaakt moeten worden. Dit is precies waar het risico op misdragende algoritmes zit.
We pleiten voor het dichten van dit gat door een sterkere interactie tussen (senior) management, domeinexperts en technici. Toen in de jaren negentig *data mining* steeds vaker gehanteerd werd, is onder andere met steun vanuit het *European Strategic Programme on Research in Information Technology* het *Cross-Industry Standard Process for Data Mining* (CRISP-DM) gevormd.[21] Dit procesmodel geeft een industrie-onafhankelijke en *tooling*-onafhankelijke methodologie voor *data mining*, en wordt ruim twintig jaar na dato nog altijd veel gebruikt. Dit model haalt al enkele praktische overwegingen aan, maar benadrukt maar in beperkte mate welke overwegingen bewust gemaakt moeten worden met de betrokkenen en welke overwegingen puur technisch van aard zijn.

Hieronder benoemen we vier situaties waarin de noodzaak tot deze interactie duidelijk aan het licht komt rondom begrippen als interpreteerbaarheid, eerlijkheid (*fairness*), correctheid (*soundness*) en verantwoordelijkheid (*accountability*) & transparantie (*transparancy*). Deze begrippen komen uit de SAFEST-principes van DNB en de aanbevelingen van de AI HLEG.[22]

Het doel van het algoritme bepaalt wat interpreteerbaar moet zijn

Toen Gerardus Mercator in 1569 zijn wereldkaart ontwierp, had hij een duidelijk gebruiksdoel voor ogen: navigatie. De wereld past vanzelfsprekend niet op een rechthoek, en zodoende moest hij keuzes maken over hoe hij de wereldbol projecteerde op een plat vlak. Hierbij koos hij voor een algoritme dat getrouw hoeken bewaart. Zodoende reflecteren hoeken op de kaart ook daadwerkelijk hoeken op de wereldbol, waardoor zijn eindgebruikers degelijk konden navigeren. Hoeken zijn hier dus goed interpreteerbaar.
Deze keuze had het directe gevolg dat de kaart onbruikbaar is om oppervlaktes van landen te vergelijken. Oppervlakte is voor een leek oninterpreteerbaar na het toepassen van Mercator's projectie; niet omdat Gerardus zijn werk niet goed gedaan heeft, maar omdat het theoretisch onmogelijk is zowel hoeken als oppervlaktes te respecteren in kaartprojecties.[23] Zolang de kaart niet gebruikt wordt voor vergelijkingen in oppervlaktes heeft dit geen nadelige effecten. In het algemeen is een dergelijke wezenlijke afweging nodig om te bepalen wat goed door de eindgebruiker geïnterpreteerd zal moeten kunnen worden. Het is met name belangrijk om te begrijpen waar en hoe de gangbare interpretatie niet langer standhoudt: *failure analysis* in vakjargon. Niet alles hoeft goed te interpreteren te zijn zolang dit een weloverwogen keuze is en dit helder gecommuniceerd wordt.

De context van het algoritme bepaalt wat eerlijk handelen is

Om gelijke behandeling van gelijke gevallen te borgen wordt in de ontwikkeling van algoritmes vaak gekeken naar de accuraatheid van de uitkomst uitgesplitst over verschillende demografieën heen (*predictive parity*). Dit is slechts een van de vele specifieke technische noties van gelijke behandeling die gekozen zou kunnen worden.[24] Welke van deze noties een gepaste vertaling is van het lekenbegrip van gelijke behandeling, hangt sterk af van de context.
Het feit wil dat deze noties vaak mutueel exclusief zijn, zodat een technische keuze hoe dan ook gemaakt zal

18. S. Yeom, A. Datta en M. Fredrikson, 'Hunting for discriminatory proxies in linear regression models', *32nd Conference on Neural Information Processing Systems*, 2018, p. 4568-4578.
19. P. Chapman e.a., 'CRISP-DM 1.0: Step-by-step data mining guide', in: CRISP-DM Consortium 2000.
20. S. Geman, E. Bienenstock en R. Doursat, 'Neural Networks and the Bias/Variance Dilemma', *Neural Computation* 1992, nr. 1, p. 1-58.
21. Zie Project ID 25959 in CORDIS, https://cordis.europa.eu/project/id/25959. Voor details van dit procesmodel zie *supra* 20.
22. Zie *supra* 6 en *supra* 8.
23. L.P. Lee, 'The Nomenclature and Classification of Map Projections', *Empire Survey Review* 1944, nr. 51, p. 190-200.
24. S.G. Mayson, 'Bias in, bias out', *Yale Law Journal* 2019, nr. 8, p. 2218-2300.

doi: 10.5553/TvT/187987052020011001008

worden.[25] Het is zeker niet ondenkbaar dat deze keuze, in afwezigheid van degelijke kadering, niet passend is voor de situatie waarin het algoritme toegepast wordt. Het is daarom belangrijk dat deze afweging bewust gemaakt wordt op het juiste niveau in lijn met relevante wet- en regelgeving en organisatorische waarden.

De correctheid van het algoritme hangt af van gekozen proxy

In Amerikaanse ziekenhuizen worden algoritmes gebruikt om in te schatten welke patiënten baat zullen hebben bij *high-risk care management* gericht op complexe zorgbehoeftes.[26] Een dergelijk algoritme wordt gemaakt op basis van gegevens van eerdere patiënten. Het is echter tijdrovend om te bepalen wat de zorgbehoefte van iedere eerdere patiënt precies was. Om die reden wordt er vaak van een proxy gebruikgemaakt die eenvoudiger te bepalen is en die waarschijnlijk nauw samenhangt met de gewenste uitkomst. In dit geval leek de meest voor de hand liggende proxy de gerealiseerde zorgkosten te zijn, die dan ook gebruikt werd om dit algoritme op te stellen. Een onverwacht en ongewenste gevolg hiervan is dat het algoritme gezondheidsrisico's bij Afrikaans-Amerikaanse patiënten onderschat. De zorgkosten voor deze groep zijn namelijk structureel lager door factoren als gebrek aan toegang tot goede zorg. Zo zorgt een gebrekkige vertaling van het algemene doel naar de precieze proxy in de data dat het resulterende algoritme incorrect is.

Exact ditzelfde probleem doet zich voor in het modelleren van criminaliteit aan de hand van arrestaties. Het is eenvoudig voor te stellen dat niet elke misdaad resulteert in een arrestatie en niet elke arrestatie is het gevolg van een werkelijke misdaad.[27] Wanneer er structurele verschillen bestaan in deze verhoudingen door demografische groepen heen, dan is gelijke behandeling in beginsel al nagenoeg hopeloos wanneer deze proxy gekozen wordt. Er zijn tal van manieren waarop *biases* onverhoopt kunnen worden geïntroduceerd in het verzamelen, verwerken, analyseren of evalueren van data.[28] Daarom is het juist van belang dat de ogenschijnlijk technische keuzes die in elk van deze stappen gemaakt moeten worden transparant en degelijk onderbouwd worden in interactie met de beslisser.

Het kiezen van een juiste proxy is een van de grootste uitdagingen in het maken van een nieuw model. De validiteit van de keuze kan intrinsiek niet enkel vastgesteld worden aan de hand van de data zelf. Dit maakt dat deze niet enkel technisch kan en mag zijn. Een inclusief perspectief helpt om je te wapenen tegen de blindheid voor systematische onvolkomenheden in de gekozen proxy's.

Verantwoordelijkheid nemen voor imperfectie

Bij het ontwikkelen van een algoritme kan ook reeds nagedacht worden over het onderhoudsplan van het algoritme. De geldigheid van het ontwerp is immers onderhevig aan randvoorwaarden. Zo is het algoritme ontwikkeld voor een specifieke context, en moet de context waarin het toegepast wordt hier voldoende op lijken. Daarnaast zouden voorzienbare incidenten, zoals het ontvangen van negatieve feedback op de kwaliteit van het algoritme, ertoe moeten leiden dat het model wordt herzien. Het breken van randvoorwaarden of wijzigingen in kwaliteitscriteria kan gekoppeld worden aan kalibratie en eventueel uitfasering. Door dit plan op te stellen en zich hieraan te committeren maakt dit ook het toezicht op het proces eenvoudiger.[29]

Een algoritme is zelden perfect. Dit hoeft geen probleem te zijn wanneer hiermee rekening wordt gehouden. Het is belangrijk om te beseffen dat elke algoritmische beslissing mensen onterecht binnen of onterecht buiten een categorie kan plaatsen. Dit is niet meer of minder dan een wiskundige realiteit. Het is belangrijk om hier transparant over te zijn en de lasten die hieruit voortvloeien eerlijk te verdelen.

Als maatschappij is dit voor ons geen nieuw concept. Zo gebruiken we immers medicijnen waarvan we weten dat ze soms nadelige bijwerkingen hebben. Als mitigatie hierop zijn we enerzijds transparant door een bijsluiter mee te geven en anderzijds monitoren we continu op nadelige gevolgen van medicijnen en wordt ernaar gestreefd deze te minimaliseren. Exact dezelfde aanpak zou gebruikt kunnen worden wanneer een organisatie kiest om algoritmische besluitvorming in te zetten.

Conclusie

Artificiële intelligentie biedt veel kansen. Voorzichtigheid is geboden om ervoor te zorgen dat we als maatschappij niet in een van de vele valkuilen stappen. Hiervoor is sterk toezicht op de manier waarop algoritmes ingezet worden van groot belang. Het is echter niet voldoende om te focussen op toezicht achteraf of te vertrouwen op menselijke interventies. We pleiten voor meer nadruk op het wordingsproces van algoritmes. Hier zien wij twee grote kansen.

Allereerst is nadere verbinding nodig tussen onderzoekers en practicanten op het gebied van toezicht, beleid en het ontwikkelen van artificiële intelligentie. Ontwikkelaars van algoritmes willen doorgaans de juiste keuzes maken, maar hebben soms onvoldoende zicht op de context om de keuze weloverwogen te maken. Anderzijds is het voor toezichthouders en beleidsmakers lastig te

25. J. Kleinberg, S. Mullainathan en M. Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores', in: C.H. Papadimitriou (red.), *8th Innovations in Theoretical Computer Science Conference*, ITCS 2017, p. 43:1-43:23.
26. Z. Obermeyer, B. Powers, C. Vogeli en S. Mullainathan, 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science* 2019, 366(6464), p. 447-453.
27. *Supra* 24.
28. A. Olteanu, C. Castillo, F. Diaz en E. Kıcıman, 'Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries', *Frontiers in Big Data* 2019, nr. 2, p. 1-33.
29. Dit plan gaat verder dan een puur periodieke keuring. Het is immers gebaseerd op een grondige analyse van de mogelijkheden tot falen, en het zou zodoende gezien kunnen worden als een catalogisering van gebeurtenissen waarop een *review* dient te volgen.

doi: 10.5553/TvT/187987052020011001008

overzien welke regelgeving het beste risicomitigerend en kwaliteitsverhogend zal zijn. Door deze groepen nader tot elkaar te brengen is het mogelijk om concretere stappen te zetten naar een zorgvuldigere en verantwoordere inzet van algoritmes.

Ten tweede hebben *data scientists* baat bij een heldere handreiking die de relevante overwegingen aanstipt. Aan academische teksten op het gebied van algoritmische ethiek en *fairness* is niet zozeer een gebrek. Ook zijn er principiële aanbevelingen te over. Een concrete, praktische handreiking in hoe deze principes en technieken realistisch en effectief ingezet kunnen worden zou veel positieve invloed hebben.

Met het CRISP-DM-model werd houvast geboden aan het doen van *data mining*. De lessen hieruit zijn nog steeds relevant en het procesmodel wordt heden ten dage nog altijd gebruikt. De SAFEST-principes van DNB materialiseren deels als aanleiding voor dialoog tussen technici en bestuurders in het doorlopen van het CRISP-DM-proces. We hebben hierboven vier voorbeelden van dergelijke te voeren dialogen gegeven. We pleiten voor een uitbreiding van dit procesmodel in samenwerking tussen *data scientists* en toezichthouders en beleidsmaker om deze principes concreet te vertalen en houvast te geven bij de nodige gesprekken. Toetsing op de mate waarin dit proces degelijk gevolgd is, kan zo op termijn bijdragen aan extern toezicht.

doi: 10.5553/TvT/187987052020011001008