Bestuurlijke borging van integere algoritmes

dr. J. Goudsmit & prof. dr. ir. G.J.de Ridder*

Trefwoorden: algoritme, integriteitsrisico, waarden, value sensitive design

Compliance heeft toe te zien op het integer gebruik van algoritmes. Algoritmes zijn zelden waardenvrij. In hun implementatie wordt al dan niet bewust stelling genomen in de spanningen tussen meerdere conflicterende waarden. De uitdaging is om te borgen dat deze stellingnames bewust plaatsvinden in de daarvoor gepaste gremia.
De conflicten tussen relevante waarden rondom de implementatie van algoritmes kunnen gezien worden als integriteitsrisico's. Door gestructureerd waardenconflicten rondom het ontwerp en de inzet van algoritmes te identificeren, analyseren en beoordelen kan een organisatie borgen dat deze maatschappelijk betamelijk zullen zijn. In dit proces kan gebruik gemaakt worden van de methode van *Value Sensitive Design*. Hiermee kan men niet alleen de waarden van belanghebbenden grondig verkennen maar ook zoeken naar beheersmaatregelen.
Zodoende wordt uiteindelijk bewust besloten of het nettointegriteitsrisico dat voortkomt uit de stellingname acceptabel is. Dit besluit hoort thuis in de gebruikelijke beslisstructuur van de organisatie. Hierin kan compliance op gepaste wijze tegenwicht bieden. Op deze manier wordt de betamelijkheid van algoritmes bestuurlijk geborgd.

Inleiding

Algoritmes zijn niet nieuw. Deze recepten voor berekening worden al millennia door wiskundigen gebruikt. Zelfs de ontwikkeling van kunstmatige intelligentie voert terug tot pionierend werk in de jaren zestig.[1] In de afgelopen twee decennia zijn de ontwikkelingen echter onmiskenbaar in een stroomversnelling geraakt. Als maatschappij voelen we ons oncomfortabel bij het rappe tempo waarin deze nieuwe mogelijkheden worden benut om wezenlijke beslissingen te maken.

Zo voorspellen algoritmes onze zorgbehoefte[2] of het risico op recidive.[3] Een recente casus in Nederland was het fraudebestrijdingssysteem SyRI van Sociale Zaken, dat in februari 2020 door de rechtbank verboden werd.[4] Voorspellingen van algoritmes hebben wezenlijke impact op ons leven. Daarom verwachten we van organisaties en overheden dat algoritmes met gepaste zorgvuldigheid ontwikkeld en geïmplementeerd worden. De maatschappelijke discussies laten zien dat dit nog niet altijd goed gaat.[5] We krijgen maar al te vaak reden om te twijfelen of onze belangen en waarden wel goed meegenomen worden in de afweging.

Een onderzoek naar het gebruik van algoritmes door de overheid, uitgevoerd naar aanleiding van zorgen van kamerleden Jetten en Bruins-Slot in 2017 liet zien dat nagenoeg de helft van de respondenten algoritmes bleek in te zetten, waarbij in meer dan driekwart van de gevallen sprake was van *machine learning*.[6] De Europese Commissie vormde in juni 2018 de *High Level Expert Group on AI*, die benadrukt dat kunstmatige intelligentie niet alleen robuust, wettig, maar ook ethisch verantwoord moet zijn. Een uiteenzetting van relevante ethische

* Jeroen Goudsmit is kerndocent Compliance & Integriteit Management aan de Vrije Universiteit Amsterdam en tevens compliance officer ter bestrijding van witwassen, terrorismefinanciering en sanctie- overtredingen bij de Rabobank. Jeroen de Ridder is als universitair hoofddocent filosofie verbonden aan de Vrije Universiteit Amsterdam en als bijzonder hoogleraar christelijke filosofie aan de Rijksuniversiteit Groningen. Zijn onderzoek en onderwijs richt zich onder andere op de ethiek en epistemologie van algoritmes.

1 Voor een toegankelijke introductie, zie S. Blauw, 'Wat is een algoritme?', *De Correspondent* 2 juli 2019, https://decorrespondent.nl/10306/wat-is-een-algoritme/150191584708-20794aae en H. Fry, *Hello World: How to be Human in the Age of the Machine*, Doubleday 2019.

2 Z. Obermeyer, B. Powers, C. Vogeli, & S. Mullainathan, 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science* 2019 *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.

3 J. Dressel & H. Farid, 'The accuracy, fairness, and limits of predicting recidivism', *Science Advances* 2018, 4(1), 1–5. https://doi.org/10.1126/sciadv.aao5580.

4 B. Nagtegaal, 'Geen hoger beroep tegen uitspraak over fraudebestrijdingssysteem SyRI', *NRC* 23 april 2020. Zie voor het hele dossier: https://www.nrc.nl/dossier/syri-fraudesysteem/.

5 Voor algoritmes in rechtspraak, zie bijvoorbeeld: 'Partnership on AI', *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System 2019*.

6 S. Doove & D. Otten, *Verkennend onderzoek naar het gebruik van algoritmen binnen overheidsorganisaties, bijlage bij Kamerstukken II 2018/19, 26643, 588*.

beginselen vormt een belangrijk onderdeel van hun aanbeveling.[7] De Raad van Europa benadrukt de impact die het grootschalig gebruik van algoritmes kan hebben op mensenrechten en stipt eveneens ethisch besef aan.[8] Zij stelt dat algoritmes gewild of ongewild prioritering aanbrengen in waarden – een uitermate ongemakkelijke realisatie.[9] De impact van algoritmes op menselijke waarden is een rode draad in de maatschappelijke discussie en de Europese aanpak.

De compliance professional heeft al enkele jaren van doen met deze discussie over algoritmes. Zo biedt de AVG aanknopingspunten voor het spreken over algoritmes die van doen hebben met natuurlijke personen. De Autoriteit Persoonsgegevens heeft dit onderwerp dan ook benoemd als een van haar drie focusgebieden[10] en de eerste contouren van het toezichtkader[11] hierop zijn reeds geschetst. Vanuit gedrags- en prudentieel toezicht maakten de Autoriteit Financiële Markten en De Nederlandsche Bank een analyse van de mogelijkheden voor kunstmatige intelligentie in de verzekeringssector.[12] In datzelfde jaar bracht DNB een discussiedocument uit over principes die ze belangrijk acht in het werken met kunstmatige intelligentie in de financiële sector.[13] Ook hier vormen menselijke waarden de kern van het debat, met een gemene deler in het aanstippen van ethisch besef en behoorlijkheid.

Maatschappelijk onbetamelijk gedrag kan leiden tot een verlies van vertrouwen in de organisatie. De compliance officer speelt een belangrijke rol in het borgen van integer gedrag. Dit moet niet te nauw gelezen worden: de integriteit van algoritmische processen behoort evenzeer tot het takenpakket. Dit is voor menig compliance officer nieuw terrein. In dit artikel beschrijven we daarom hoe dit op hoofdlijnen aangepakt kan worden. Ten eerste laten we zien dat (nieuwe) technologie nooit waarde-neutraal is en dat de ontwikkeling en implementatie ervan onvermijdelijk waardenconflicten met zich meebrengt. Dat geldt evenzeer voor algoritmes. Dat betekent dan ook dat algoritmes noodzakelijkerwijs stelling nemen in waardenconflicten. Ten tweede benadrukken we dat maatschappelijke betamelijkheid neerkomt op het zorgvuldig overwegen en beslechten van deze waardenconflicten in het ontwerp en de toepassing van het algoritme. Dit vereist dat de organisatie bewust positie inneemt en hiernaar handelt, wetende dat een positie hoe dan ook ingenomen zal worden door het algoritme. Ten derde pleiten we er daarom voor waardenconflicten expliciet inzichtelijk te maken, net zoals dat gebeurt met andere integriteitsrisico's. Hiervoor geven we enkele handvatten. De nodige beslissingen kunnen op deze manier op de juiste tafel genomen worden, met gepast tegenwicht vanuit compliance. Zodoende concluderen we dat organisaties waardenconflicten rondom algoritmische beslissingsprocessen moeten identificeren, analyseren en hierop een positie moeten innemen om maatschappelijke betamelijkheid te realiseren.

Waardenconflicten zijn onvermijdelijk

Met het gebruik van algoritmes zijn, net als met alle vormen van technologie, waarden gemoeid. Heel algemeen gesteld is een waarde iets wat een persoon of groep mensen belangrijk vindt in het leven.[14] Voor algoritmes zijn dat enerzijds technisch-functionele waarden zoals efficiëntie, effectiviteit, sensitiviteit, specificiteit, gebruiksvriendelijkheid en compatibiliteit met bestaande toepassingen, maar anderzijds ook morele en politieke waarden, zoals eerlijkheid, rechtvaardigheid, non-discriminatie, privacy, informed consent, transparantie, verantwoordelijkheid en betrouwbaarheid.

Zo'n veelheid van potentieel relevante waarden levert onvermijdelijk waardenconflicten op. En wel om twee redenen. Ten eerste omdat verschillende waarden verschillende kanten op kunnen trekken. Daarbij kunnen er zowel conflicten binnen de groepen van technisch-functionele waarden of morele en politieke waarden ontstaan, als tussen die twee groepen. Een voorbeeld van dat eerste is de afweging tussen sensitiviteit (niet missen van waar je naar op zoek bent) en specificiteit (niet aanwijzen waar je niet naar op zoek bent). Wie met een algoritme geen enkel fraudegeval wil missen, zal een heleboel gevallen onterecht als frauduleus aanmerken. Een voorbeeld van een conflict tussen de twee groepen van waarden is dat tussen voorspel-

[7] High Level Expert Group on Artificial Intelligence, *Policy and investment recommendations for trustworthy Artificial Intelligence* 2019, https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

[8] Council of Europe, *Algorithms and Human Rights* 2018, www.coe.int/freedomofexpression.

[9] Council of Europe, *Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems* 2020, https://go.coe.int/lBwbG

[10] Autoriteit Persoonsgegevens, *Focus AP 2020-2023: Dataprotectie in een digitale samenleving* 2019.

[11] Autoriteit Persoonsgegevens, *Toezicht op AI & Algoritmes* 2020.

[12] Autoriteit Financiële Markt & De Nederlandsche Bank, *Artificiële Intelligentie in de verzekeringssector - een verkenning* 2019.

[13] De Nederlandsche Bank, *General principles for the use of Artificial Intelligence in the financial sector* 2019.

[14] B. Friedman, P. H. Kahn Jr. & A. Borning, 'Value Sensitive Design and Information Systems' in: *The Handbook of Information and Computer Ethics* 2009 (pp. 69–101). John Wiley & Sons, https://doi.org/10.1002/9780470281819.ch4.

lende kracht en privacy. Het verzamelen, opslaan en gebruiken van meer persoonsgegevens kan de precisie en voorspellende kracht van een algoritme vergroten, maar staat op gespannen voet met het respecteren van privacy en informed consent.
Een tweede reden waarom waardenconflicten onvermijdelijk zijn is dat er verschillende partijen betrokken zijn bij algoritmes: ontwikkelaars, gebruikers, klanten, maar indirect ook de bredere organisatie, personen van wie de gegevens verzameld, opgeslagen en verwerkt worden, de overheid en toezichthouders. Deze partijen hebben eigen belangen en zullen dan ook uiteenlopend gewicht toekennen aan verschillende waarden. Ontwikkelaars zullen efficiëntie en nauwkeurigheid hoog in het vaandel hebben, gebruikers de gebruiksvriendelijkheid, en de klanten willen hun privacy gerespecteerd hebben. Omdat betamelijkheid vereist dat je als organisatie rekening houdt met de waarden en belangen van alle direct en indirect betrokkenen, zijn waardenconflicten onvermijdelijk.
Niet elk waardenconflict is even fundamenteel of onoplosbaar. Er zijn ten minste twee wegen die bewandeld kunnen worden om conflicten te verminderen of op te lossen. Ten eerste is dat nadere specificatie van of reflectie op de relevante waarden.[15] Een waarde als veiligheid bijvoorbeeld is zonder verdere invulling veel te breed. Technologie zal maar zelden 100% veilig zijn, dus is de vraag welke risico's aanvaardbaar zijn en voor wie ze aanvaardbaar zijn. Hoe weeg je grotere risico's op kleine ongelukken af tegen kleine risico's op hele grote ongelukken? Tellen alleen fysieke ongelukken of moet psychologische schade meegewogen worden in veiligheid? Door de betrokken waarden te definiëren en te specificeren, wordt een conflict tussen twee waarden vanzelf minder of lost het op.
Een tweede mogelijkheid om waardenconflicten te mitigeren of op te lossen is (innovatief) herontwerp. Een aansprekend voorbeeld hiervan uit een andere sector is de Oosterscheldekering.[16] Na de Watersnoodramp van 1953 was het plan in eerste instantie om de Oosterschelde geheel af te sluiten met een dam om zo bescherming te bieden tegen toekomstige overstromingen. Begin jaren zeventig zwollen de protesten hiertegen echter aan, vanuit zowel de visserij als de milieubeweging. Deze partijen pleitten voor het verhogen van de dijken als alternatief voor volledige afsluiting. De waarden die hier botsten waren enerzijds veiligheid en anderzijds werkgelegenheid en milieubehoud. In 1976 werd uiteindelijk besloten tot het aanbrengen van schuifdeuren in een gedeelte van de kering. Die deuren staan normaal open maar kunnen bij storm dicht. Deze oplossing realiseerde een compromis tussen de botsende waarden waar alle partijen tevreden mee waren.
Hoewel deze twee opties – het nader specificeren van waarden en het mitigeren van conflicten via herontwerp en innovatie – belangrijke eerste stappen zijn om waardenconflicten te verminderen, valt niet te verwachten dat alle conflicten erdoor verdwijnen. Er zullen situaties overblijven waarin er waardenknopen doorgehakt moeten worden, zoals we verderop nog nader zullen illustreren. Dat is geen zuiver technische of bedrijfseconomische kwestie, maar vereist afweging van de belangen van direct en indirect betrokken partijen. Het is daarom van belang dat de geëigende gremia binnen de beslisstructuur van de organisatie zowel de belangenafweging maken als de beslissing nemen.

Algoritmes nemen stelling in waardenconflicten

We kunnen deze algemene inzichten over technologie en waarden verder invullen voor algoritmes. Uiteraard zijn voor algoritmes de eerder genoemde technisch-functionele waarden van belang. Maar de maatschappelijke discussie benadrukt daarnaast waarden als gelijkheid, solidariteit en menselijke waardigheid.[17] Toezichthouders leggen nadruk op behoorlijkheid (*fairness*), rechtmatigheid en transparantie.[18] Veel organisaties hebben daarnaast eigen kernwaarden en leggen bijvoorbeeld nadruk op waarden als openheid, persoonlijkheid en verantwoordelijkheid. Deze waarden worden door verschillende personen op uiteenlopende manieren geïnterpreteerd en gewaardeerd. De uitdaging is dus om zorgvuldig stelling te nemen in het spanningsveld tussen deze technisch-functionele en menselijke en moreel-politieke waarden.
Een reflex kan zijn om dit gesprek te laten voor wat het is. Zijn waarden niet te vaag en te subjectief om systematisch te verdisconteren? En kun je niet prima een algoritme implementeren zonder uitvoerig over waarden na te denken? Zoals we reeds zagen is het echter onvermijdelijk dat algoritmes positie kiezen in het spanningsveld tussen verschillende waarden, simpelweg door het maken van ontwerpkeuzes. Een mooi voorbeeld hiervan zien we in modellen voor natuurlijke taalverwerking, zoals het model word2vec. Dit model kent betekenis toe aan woorden op

[15] I. van de Poel, 'Values in Engineering Design. In *Philosophy of Technology and Engineering Sciences* 2009 (Vol. 9)'; Elsevier, https://doi.org/10.1016/B978-0-444-51667-1.50040-9
[16] Zie noot 15.
[17] Zie noot 7.
[18] Zie noot 13.

basis van hun gebruik in een enorme verzameling bestaande teksten. Bij gebruik blijkt dat het model precies de stereotypen uit de verzameling teksten reproduceert.[19] Volgens het model is '*man to computer programmer as woman is to homemaker*'.[20] Ongetwijfeld zal dit een correcte weergave zijn van hoe de woorden gebruikt worden in een verzameling bestaande teksten, maar de vraag natuurlijk is of we onze modellen klakkeloos traditionele man/vrouw rolverdelingen willen laten bevestigen. Dat is een onbedoelde stellingname met verregaande consequenties. Een bevredigende manier om de zogeheten *genderbias* uit het model te halen is nog niet bekend.[21] Correctheid en gelijkheid blijken dus vooralsnog lastig technologisch in balans te brengen. Volledige mitigatie van het waardenconflict is dus (nog) onmogelijk. Wat wel kan, is het conflict tussen deze waarden expliciet maken. Transparantie bieden over de tekortkoming kan een werkbare tussenoplossing zijn, zoals reeds toegepast door bedrijven als Google.[22]

Een volgende valkuil is om direct aan de slag te gaan met een oppervlakkig besef van de relevante waarden. Het volstaat niet om enkel te realiseren dat bepaalde waarden van belang zijn. Algemene waarden moeten specifiek gemaakt worden en vertaald naar de concrete situatie. Neem bijvoorbeeld de waarden gelijkheid en non-discriminatie. Die klinken misschien eenduidig genoeg, maar kunnen vertaald worden naar wezenlijk verschillende technische randvoorwaarden.[23] De al aangestipte modellering van recidivisme raakte in opspraak na beschuldigingen van racisme.[24] Hierbij werden andere technische vertalingen gekozen van de waarde non-discriminatie dan door de oorspronkelijke producent gehanteerd werden. Een grondigere analyse van de waarde had de kans geboden beter te begrijpen waar de betrokkenen echt om zouden geven. Door deze analyse vroegtijdig uit te voeren en mee te nemen als randvoorwaarden in het ontwikkelproces had veel leed voorkomen kunnen worden, zowel voor de betrokkenen als voor het bedrijf. Een ander voorbeeld is de discussie over het belang van *explainable AI* voor transparantie en verantwoording.[25] Wat houdt het precies in dat algoritmisch ondersteunde beslissingen 'explainable' moeten zijn? Moeten gebruikers en klanten of burgers daarvoor in detail begrijpen hoe het algoritme in elkaar steekt en tot resultaten komt? Waarschijnlijk niet. Het draait eerder om helderheid over welke verantwoordelijkheden waar belegd zijn, welke factoren bepalend zijn geweest voor een beslissing, en de mogelijkheid tot beroep. Door dit vroeg in het ontwikkelproces te expliciteren, kan voorkomen worden dat er nodeloze of onhaalbare eisen aan het ontwerp worden gesteld.

Ten slotte kan het misverstand bestaan dat discussies over het specificeren van waarden en omgaan met waardenconflicten slechts operationele kwesties zijn die zonder bestuurlijke inmenging afgehandeld kunnen worden. Niets is minder waar. Het maatschappelijk draagvlak voor de inzet van een algoritme staat of valt met een zorgvuldige en weloverwogen stellingname in waardenconflicten. Deze stellingname is een fundamentele keuze die raakt aan de *license to operate*. Ze moet dan ook genomen worden op de juiste tafel, in het besef dat dit een beslissing is over risico-acceptatie. Daarbij hoort vertegenwoordiging vanuit compliance, zodat gezond tegenwicht geboden kan worden.

Een positief voorbeeld zien we in de onderzoekers die werkten aan een stedenbouwkundig voorspellingsmodel.[26] Dit model zou gebruikt gaan worden om beslissingen over de inrichting van steden te informeren. Het maatschappelijk draagvlak hiervan staat of valt met de acceptatie van dit algoritme als een legitiem instrument om beslissingen te ondersteunen. Begrijpbaarheid, nauwkeurigheid, transparantie, relevantie en gebrek aan vooringenomenheid werden erkend als cruciale waarden die allen verwezenlijkt moeten worden. En dit werd

[19] N. Garg, L. Schiebinger, D. Jurafsky & J. Zou, 'Word embeddings quantify 100 years of gender and ethnic stereotypes', *Proceedings of the National Academy of Sciences 2018*, *115*(16), 3635–3644. https://doi.org/10.1073/pnas.1720347115

[20] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama & A. Kalai, 'Debiasing Word Embedding', *30th Conference on Neural Information Processing Systems*, *NIPS 2016*, 1–9.

[21] H. Gonen & Y. Goldberg, 'Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019, Volume 1 (Long and Short Papers)*, 609–614. https://doi.org/10.18653/v1/N19-1061.

[22] D. Lee, 'Google Translate now offers gender-specific translations for some languages', *The Verge* 2018. https://www.theverge.com/2018/12/6/18129203/google-translate-gender-specific-translations-languages.

[23] S.G. Mayson, 'Bias in, bias out', *Yale Law Journal* 2019,128(8), 2218–2300.

[24] J. Angwin, J. Larson, S. Mattu & L. Kirchner, 'Machine Bias', *ProPublica* 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[25] Zie hierover W. Knight, 'The dark secret at the heart of AI', *MIT Technology Review* 2017, https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/; R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. & Pedreschi, 'A survey of methods for explaining black box models', *ACM Computing Surveys* (CSUR) 2018, 51(5), 93; en noot 14.

[26] A. Borning, B. Friedman, J. Davis & P. Lin, 'Informing public deliberation: Value Sensitive Design of indicators for a large-scale urban simulation', *ECSCW 2005 - Proceedings of the 9th European Conference on Computer-Supported Cooperative Work*, *September*, 449–468. https://doi.org/10.1007/1-4020-4023-7_23.

serieus genomen. Er werd gemeten of belanghebbenden herkenden dat ontwerpkeuzes deze waarden ook daadwerkelijk reflecteren. Hiermee garandeert men het algoritme van een *license to operate*.
We zien dus dat maatschappelijk betamelijke algoritmes rekening houden met conflicten tussen de waarden van de belanghebbenden. Ten eerste is het noodzakelijk om deze waardenconflicten goed te identificeren, omdat je anders blind bent voor de risico's die hieruit voort kunnen komen. Ten tweede is het grondig analyseren van waardenconflicten nodig om stelling te kunnen nemen, omdat je anders de risico's niet goed begrijpt. Ten derde dient het juiste gremium deze stellingname te bepalen, omdat de risico-acceptatie die eruit voortvloeit in meer of mindere mate raakt aan de *license to operate*. Deze drie stappen moeten goed verankerd zijn in de gebruikelijke besluitvorming.

Behandel waardenconflicten als integriteitsrisico's

Omgaan met integriteit is de compliance officer niet vreemd. Bij wet hebben Financiële Instellingen de verplichting om toe te zien op een integere bedrijfsvoering.[27] Hiervoor is de integriteitsrisicoanalyse een bekend instrument.[28] Door systematisch een proces van risico-identificatie, -analyse en -acceptatie te doorlopen worden integriteitsrisico's bewust en op de juiste tafel aanvaard. Wij pleiten ervoor om dezelfde gedachtegang te volgen in de omgang met waardenconflicten. Waardenconflicten kunnen immers ten grondslag liggen aan reputatieschade en maatschappelijk onbetamelijk handelen – integriteitsrisico's *pur sang*. Wanneer we waardenconflicten in het ontwerp en gebruik van algoritmes inderdaad beschouwen als integriteitsrisico's is het zaak om ze systematisch in kaart te brengen en te analyseren. In het denken over de ethische aspecten van technologie en technisch ontwerp, zijn hiervoor de afgelopen decennia verschillende methodieken ontwikkeld.[29]
Eén voorbeeld hiervan dat specifiek betrekking heeft op de waarde privacy, is de methodiek van *privacy by design*.[30] Deze benadering heeft als uitgangspunt dat privacy standaard proactief en preventief gedurende het hele ontwerpproces van nieuwe applicaties en systemen meegenomen moet worden om zo te voorkomen dat het een *afterthought* wordt. Adequate bescherming vereist dat privacy het programma van wensen en eisen mede bepaalt en van beslissende invloed is op het basisontwerp, de prototypes en de uiteindelijke beoordeling van het eindproduct.
Zo'n integrale en holistische benadering van nadenken over waarden en technologie kenmerkt ook de methode van *Value Sensitive Design* (VSD).[31] Deze is in de jaren 90 ontwikkeld en in eerste instantie toegepast bij de ontwikkeling van IT-systemen. Ze biedt een integraal en omvattend raamwerk om het denken vanuit waarden te integreren in het hele ontwerp- en ontwikkelproces van nieuwe technologie. De methode is toegepast op de ontwikkeling van uiteenlopende technologieën: *computerinterfaces*, kantoorontwerp, tools voor mensen met een visuele handicap, *e-health* toepassingen enzovoorts.[32]

VSD kent drie onderdelen of lagen, die iteratief doorlopen moeten worden in een ontwerpproces en elkaar wederzijds beïnvloeden.

1. *Conceptuele analyse:* in dit onderdeel draait het om in kaart brengen en verhelderen van belanghebbenden, hun waarden en hun relatieve gewicht. Vragen die beantwoord moeten worden zijn: Wie zijn de partijen die direct en indirect belang hebben bij de nieuwe technologie? Hoe beïnvloedt de technologie hen? Wat zijn hun waarden en hoe definiëren ze die waarden? Wat is het onderlinge gewicht van de verschillende waarden en hoe kunnen ze tegen elkaar afgewogen worden?

[27] Zie *Besluit prudentiële regels Wft* (https://wetten.overheid.nl/BWBR0020420/2020-01-01).
[28] De Nederlandsche Bank (2015), *De integriteitrisicoanalyse - meer waar dat moet, minder waar dat kan*.
[29] Zie bijv. J. Van den Hoven, S. Miller, & T. Pogge, *Designing in ethics*, Cambridge University Press 2017; J. Van den Hoven, P.E. Vermaas & I. Van de Poel (red.), *Handbook of ethics, values, and technological design*, Springer 2015.
[30] Cavoukian, A. (2010). 'Privacy by Design: The 7 Foundational Principles', *The International Association of Privacy Professionals*. https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf.
[31] B. Friedman, 'Value-sensitive design', *Interactions* 1996, 3(6), 16–23; B. Friedman, P.H. Kahn Jr. & A. Borning, 'Value Sensitive Design and Information Systems', In *The Handbook of Information and Computer Ethics* 2009 (pp. 69–101). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470281819.ch4; B. Friedman & D.G. Hendry, *Value sensitive design: Shaping technology with moral imagination*, MIT Press 2019.
[32] T. Winkler & S. Spiekermann, 'Twenty years of value sensitive design: a review of methodological practices in VSD projects', *Ethics and Information Technology* 2018 https://doi.org/10.1007/s10676-018-9476-2; noot 31.

2. *Empirische analyse:* het tweede onderdeel van VSD draait om het beantwoorden van allerhande vragen over de menselijke context waarin de te ontwikkelen technologie functioneert. De vragen hebben betrekking op zowel de waarden van de betrokkenen als op het functioneren en het gebruik van de technologie. Welke waarden zeggen belanghebbenden te hanteren? Hoe prioriteren zij deze waarden? Wat zegt het gedrag van belanghebbenden over hun waarden? Daarnaast: wat zijn de psychologische en sociale voorwaarden en effecten van verschillende ontwerp-opties? Hoe interacteren mensen ermee, hoe zijn de opties van invloed op de directe organisationele en bredere context? Wat zijn de effecten op diverse groepen van belanghebbenden? De hele waaier van kwalitatieve en kwantitatieve sociaalwetenschappelijke methoden kan ingezet worden om zulke vragen te beantwoorden.
3. *Technologische analyse:* in het derde onderdeel komen de relaties tussen technologie en waarden direct in beeld. Twee soorten vragen staan centraal. Enerzijds hoe bestaande ontwerp-opties de realisatie van relevante waarden stimuleren of bemoeilijken en anderzijds hoe technologie zo ontworpen kan worden dat de eerder geïdentificeerde waarden zo goed mogelijk in acht genomen kunnen worden. Waar in het vorige onderdeel de menselijke context centraal stond, staat in dit onderdeel de technologie zelf centraal.

De drie soorten analyses beïnvloeden elkaar wederzijds: identificatie en specificatie van relevante waarden leidt bijvoorbeeld tot verschillende empirische en technologische vragen en andersom kunnen empirische en technologische analyses weer leiden tot nieuwe conceptuele vragen over hoe waarden precies begrepen moeten worden of over welke belanghebbenden meegenomen moeten worden. Er zou nog veel meer te zeggen zijn over de details en praktische toepassingen van VSD.[33] Hier moeten we ons beperken tot twee korte voorbeelden van situaties waarin toepassing van VSD leidde tot beter zicht op de impact van nieuwe technologie op diverse groepen belanghebbenden, om zo duidelijk te maken hoe VSD kan helpen bij het identificeren van en omgaan met waardenconflicten.[34]
De eerste casus betreft een witboek over *Augmented Reality*; een overzicht van mogelijke toepassingen, technische kwesties en belangrijke juridische en beleidsvragen. Bij het opstellen van dit witboek nodigden de schrijvers drie klankbordgroepen van belanghebbenden uit die doorgaans niet of slecht gerepresenteerd zijn: minder validen, ex-gedetineerden en vrouwen. Alle drie deze groepen leverden waardevolle inzichten die aanleiding gaven tot herzieningen in de omschrijving van AR, de mogelijke impact ervan op minderheden en kwetsbare groepen en de risico's op misbruik. Iets vergelijkbaars gold in een tweede casus: het opstellen van een beleidsdocument over het stimuleren van het gebruik van zelfrijdende auto's. Door groepen jongeren, mensen die geen auto gebruiken en mensen met een laag inkomen te betrekken bij het schrijven van dit document, ontstond een veel omvattender beeld van de relevante waarden, risico's en uitdagingen bij zelfrijdende auto's. De algemene les die hieruit getrokken kan worden, is dat het betrekken van een brede groep direct en indirect belanghebbenden cruciaal is om goed zicht te krijgen op de waarden die gemoeid zijn met de introductie van nieuwe technologie. Het zal niet altijd praktisch haalbaar zijn om dat op zo'n grondige en uitvoerige wijze te doen als in deze casussen. Maar op z'n minst is het zinvol om belanghebbenden in de eigen organisatie te consulteren en bijvoorbeeld een ondernemingsraad al vroeg in het ontwerpproces in te schakelen om mee te denken.
De methodiek van VSD biedt zodoende een beproefd recept om om te gaan met waardenconflicten als integriteitsrisico's. Consequente toepassing ervan geeft inzicht in wie de directe en indirecte belanghebbenden voor een nieuw algoritme zijn, wat hun waarden zijn, hoe die begrepen moeten worden, en de conflicten die ertussen kunnen ontstaan. Dat geeft een beeld van wat je het bruto-risico kunt noemen. Maar dat niet alleen; de methode levert eveneens inzicht op in de beheersmaatregelen die er zijn om conflicten te mitigeren of op te lossen, hetzij via interventie in de menselijke context, hetzij via technologische innovatie. De methode maakt zo niet slechts duidelijk wat het bruto-risico is, maar ook wat hiervan ondervangen kan worden en wat er uiteindelijk overblijft als netto-risico. Dat geeft helderheid over de te nemen beslissingen ten aanzien van risico-acceptatie. Het doorhakken van deze spreekwoordelijke knoop is een keuze die de beslissers samen, in goed moreel besef, moeten nemen.

[33] Het meest omvattende en actuele overzicht biedt de in noot 31 geciteerde literatuur.

[34] Zie voor meer details over de voorbeelden M. Young, L. Magassa & B. Friedman, 'Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents', *Ethics and Information Technology* 2019 *21*(2), 89–103. https://doi.org/10.1007/s10676-019-09497-z.

Conclusie

Heel algemeen gesteld is een waarde iets wat een persoon of groep mensen belangrijk vindt in het leven. Daar waar meerdere belanghebbenden zijn, is het waarschijnlijk dat hun waarden met elkaar botsen. Algoritmes nemen onvermijdelijk stelling in deze waardenconflicten. Het is zaak om deze stellingname goed overwogen en bewust te nemen op de juiste tafel in de gangbare beslisstructuur.

Door volgens de methode van VSD gestructureerd waardenconflicten rondom het ontwerp en gebruik van algoritmes te identificeren, analyseren en beoordelen kan een organisatie borgen dat algoritmes maatschappelijk betamelijk zullen zijn. De identificatie steunt op een conceptuele verkenning, waarin indirecte en directe belanghebbenden, hun waarden en de conflicten hiertussen in kaart gebracht worden. Daarna moet de interpretatie en prioritering van de waarden door de belanghebbenden empirisch in kaart gebracht worden en moet de sociale context van de diverse ontwerp-opties onderzocht worden. Het verkennen van technologische mogelijkheden kan tot slot laten zien in hoeverre de ontwerp-opties rekening houden met de geïdentificeerde waarden van belanghebbenden en of er wellicht innovatieve oplossingen zijn die de betrokken waarden nog beter verdisconteren. Pas dan kan in het daartoe geëigende gremium beoordeeld worden welke stellingname wenselijk is, in vol begrip van het te accepteren netto-risico. De compliance officer zit hier reeds aan tafel. Op deze manier kan ze haar nieuwe rol in het borgen van algoritmische integriteit serieus vervullen. ■