Inzicht in transparantie

Een essay over trade-offs achter algoritmische besluitvorming

Strycharz, J.; Steenhuisen, B.; van der Voort, H.

DOI
10.5553/Bk/092733872020029004005

Publication date
2020

Document Version
Final published version

Published in
Bestuurskunde

Inzicht in transparantie

Een essay over trade-offs achter algoritmische besluitvorming*

Joanna Strycharz, Bauke Steenhuisen & Haiko van der Voort

Algoritmen in het openbaar bestuur worden steeds vaker bekritiseerd om een gebrek aan transparantie. Wetgever en de burger verwachten dat besluiten die genomen worden op basis van algoritmen en die van invloed zijn op het leven van individuen, transparant zijn. Dit artikel beredeneert de organisatorische context achter dit idee van transparante algoritmen. We zien transparantie als een waarde, net als waarden als veiligheid, duurzaamheid of privacy. Daarmee wordt al snel duidelijk dat transparantie op gespannen voet kan staan met andere waarden. Indien niet alle waarden maximaal geborgd kunnen worden, ontstaat er waardenconflict. Hoe gaan organisaties met waardenconflict om bij algoritmische besluitvorming? Daarover is niet veel bekend. In deze bijdrage theoretiseren we de omgang met waardenconflict in organisaties. We concluderen dat er meer aandacht nodig is voor de transparantie van de organisatie achter waardenconflict, in plaats van de transparantie van algoritmen. We sluiten af met een onderzoeksagenda.

De roep om transparante algoritmen

In de film 'Minority Report' is de technologie zo ver, dat moordenaars kunnen worden opgespoord voordat ze hun moord zullen begaan. Soms weten ze zelf niet dat ze iets kwaads zouden gaan doen. Goed idee, zo lijkt aan het begin van de film. De twijfel volgt: kloppen de voorspellingen eigenlijk wel? Op basis van welke data worden de voorspellingen gemaakt? En wie construeert en interpreteert de informatie? Worden er onschuldige mensen gearresteerd en hoe kunnen we daarachter komen?

Ook heden ten dage is er de twijfel over de toepassing van algoritmen in het openbaar bestuur. De algoritmen in 'Minority Report' bleken corrupt, net als de mensen die ze maakten en interpreteerden. Zo ontstond een duidelijk beeld van goeden en kwaden. Dat duidelijke beeld is er niet altijd in het huidige openbaar bestuur. De twijfels voeden wel een roep om transparantie (Leenes, 2016). Ook zonder verkeerde intenties blijken algoritmen vaak bevooroordeeld en dan willen we graag dat iemand daarvoor betekenisvol verantwoording over kan afleggen.

* J. Strycharz, Msc is universitair docent Persuasive Communication aan de Universiteit van Amsterdam, Faculteit Maatschappij- en Gedragswetenschappen. Dr. ir. B.S. Steenhuisen is universitair docent Organisatie & Governance aan de TU Delft, Faculteit Techniek, Bestuur en Management. Dr. H.G. van der Voort is universitair docent Organisatie & Governance aan de TU Delft, Faculteit Techniek, Bestuur en Management.

Algoritmen in het openbaar bestuur worden steeds vaker bekritiseerd om hun gebrek aan transparantie. Wetgever en de burger verwachten dat algoritmen die in het openbaar bestuur toegepast worden en die van invloed zijn op het leven van individuen, verklaarbaar zijn (Preece, Harborne, Braines, Tomsett, & Cakraborty, 2018). Deze verwachting is terug te zien in recente wetgeving zoals de Algemene verordening gegevensbescherming (AVG), waarin rechtmatigheid, behoorlijkheid en transparantie van algoritmische besluiten als kernbeginselen opgenomen zijn en nader zijn uitgewerkt in verschillende normen. In Nederland is de Autoriteit Persoonsgegevens als toezichthouder verantwoordelijk voor het toezicht op de toepassing van AI en algoritmen waarbij persoonsgegevens worden gebruikt.[1]

De roep om transparantie van algoritmen is intuïtief. We zien enerzijds dat analytische functies verder dan voorheen geautomatiseerd worden. De computer neemt meer denkfuncties van mensen over. Anderzijds zien we dat besluiten op basis van deze algoritmen kwetsbare mensen kunnen raken. Waarom patrouilleert de politie zo vaak in mijn wijk? Waarom word ik geselecteerd voor fraudeonderzoek? Waarom word ik niet uitgenodigd voor een sollicitatiegesprek?

Maar hoe eenvoudig is het om transparantie van algoritmen te garanderen? Dit artikel beredeneert de organisatorische context achter transparantie van algoritmen. We vatten 'transparantie' hierbij op als een waarde. Dit verheldert al snel waarom transparantie lastig is. Er zijn namelijk meerdere waarden in het spel als besluiten in het openbaar bestuur met hulp van algoritmen worden genomen, zoals privacy en effectiviteit. Deze kunnen bovendien met elkaar conflicteren. De belangrijkste waarden zetten we uiteen. Vervolgens beredeneren we de organisatorische context achter de manier waarop met waardenconflict omgegaan wordt. Een logische conclusie is dat transparantie van algoritmen betrekking kan hebben op de transparantie van de organisatie. Ook deze conclusie benaderen we kritisch, vanuit de theorie over waardenconflict. We sluiten dit essay af met een onderzoeksagenda over de omgang met waardenconflict bij algoritmische besluitvorming.

Het belang van transparantie

Algoritmen die in het openbaar bestuur worden ingezet, baseren zich vaak op data over burgers. Zulke algoritmen waarbij persoonsgegevens worden gebruikt, zijn onderworpen aan de AVG. Een belangrijk doel van de AVG is dat algoritmen weliswaar effectief zijn, maar tegelijk voor mensen begrijpelijk en daarmee betrouwbaar zijn. In de literatuur worden vier belangrijke redenen voor meer transparantie beschreven, namelijk verantwoording, controle, verbeteringen en verdere ontwikkelingen. Ten eerste faciliteert transparantie verantwoording over besluiten die zijn genomen met behulp van algoritmen. Hier wordt over het

1 Zie https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/toezicht_op_ai_en_algoritmes.pdf.

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004005

algemeen de noodzaak van redenen of rechtvaardigingen voor een specifiek resultaat bedoeld, en niet een beschrijving van de innerlijke werking van algoritmen in het algemeen (Adadi & Berrada, 2018). In deze zin is transparantie belangrijk voor de ethiek van algoritmen: zo kunnen besluitvorming en de logica achter de besluiten worden uitgelegd (Etzioni & Etzioni, 2017). Ten tweede geeft transparantie een beter zicht op onbekende kwetsbaarheden en tekortkomingen van algoritmen, en helpt zij om snel fouten te identificeren en te corrigeren (zogenoemde *debugging*). Transparantie maakt het dus mogelijk om de werking van algoritmen te controleren. Ten derde streeft men ernaar om algoritmen altijd te verbeteren. Een model dat kan worden uitgelegd en begrepen, kan ook gemakkelijker worden verbeterd (Adadi & Berrada, 2018). Daarmee kan transparantie bijdragen aan de verbetering van bestaande algoritmen. Als laatste dient transparantie als middel voor verdere ontwikkelingen in het veld van kunstmatige intelligentie. Transparantie over de werking van een algoritme is een handig hulpmiddel om verder te leren, informatie over impact van algoritmen te verzamelen en daarmee kennis op te doen. Zo draagt transparantie niet alleen bij aan verbetering van bestaande algoritmen, maar ook aan vooruitgang in kunstmatige intelligentie. In het algemeen is het doel van transparantie om het menselijk begrip en vertrouwen in besluitvorming te verbeteren, om onpartijdige en rechtvaardige beslissingen te bevorderen en om verdere ontwikkelingen te bevorderen (Das & Rad, 2020).

Richtlijnen voor ethische algoritmen

Transparantie is echter niet het enige principe dat zou moeten gelden voor algoritmen. Als algoritmen worden toegepast voor besluitvorming in het openbaar bestuur, kan deze toepassing ernstige gevolgen hebben voor individuen. Daarom heeft de High-Level Expert Group on Artificial Intelligence die door de Europese Commissie in het leven is geroepen, richtlijnen voor toepassing van algoritmen en kunstmatige intelligentie gepubliceerd. Deze richtlijnen zijn gebaseerd op drie componenten. Ten eerste moeten algoritmen aan alle toepasselijke wetgeving voldoen. Ten tweede moeten ze ethisch verantwoord zijn en ervoor zorgen dat ethische principes en waarden worden nageleefd. Ten derde moeten ze robuust zijn, zowel technisch als sociaal, aangezien algoritmen zelfs goedbedoeld onbedoelde schade kunnen veroorzaken (High Level Independent Group on Artificial Intelligence, 2019). Terwijl er duidelijke regels in de AVG staan over bijvoorbeeld het gebruik van persoonlijke gegevens voor algoritmen, is het vaak moeilijker om aan de tweede en derde component te voldoen. In deze sectie beschrijven we waarden die van belang zijn voor ethische en robuuste algoritmen. Deze omvatten transparantie, maar ook andere waarden die van belang zijn. Over deze waarden moeten vaak afwegingen worden gemaakt door bestuurders die de verantwoordelijkheid dragen voor het toepassen van algoritmen.

Ethische algoritmen zijn gebaseerd op vier hoofdprincipes: respect voor menselijke autonomie, voorkomen van schade, eerlijkheid en de bovengenoemde trans-

parantie en verklaarbaarheid. De menselijke autonomie omvat het idee dat ieder mens een 'intrinsieke waarde' bezit, die nooit mag worden verminderd, aangetast of onderdrukt door anderen (McCrudden, 2008) – ook wanneer algoritmen worden ingezet. Individuen moeten dus niet worden gezien als 'objecten' die door algoritmen worden beoordeeld, maar als individuen met menselijke waardigheid. Mensen die omgaan met algoritmen (door ze in te zetten of over die algoritmische besluiten te nemen), moeten in staat zijn om volledige zelfbeschikking over zichzelf te behouden. Bij de inzet van algoritmen staat de mens centraal. Dat betekent dat het belangrijk is dat bestuurders overzien hoe algoritmen worden ingezet.

Ten tweede mogen algoritmen geen schade veroorzaken of verergeren of anderszins een nadelige invloed hebben op individuen. Dit betreft individuele of collectieve schade die ook immaterieel kan zijn, bijvoorbeeld schending van de privacy. Er moet dus voor worden gezorgd dat algoritmen niet kwaadaardig gebruikt kunnen worden (High Level Independent Group on Artificial Intelligence, 2019). Dit is bijzonder belangrijk voor kwetsbare groepen in de maatschappij die meer aandacht verdienen bij het toepassen van algoritmen in het openbaar bestuur. Bestuurders moeten zich beseffen dat asymmetrie van macht vaak tot onbedoelde schade kan leiden bij het toepassen van algoritmen in besluitvorming over individuen.

Ten derde moeten algoritmen eerlijk zijn. Eerlijkheid is een moeilijk te definiëren begrip wat het ook lastig maakt voor bestuurders om te bepalen wanneer een algoritme eerlijk werkt. De High Level Independent Group on Artificial Intelligence (2019) stelt voor dat eerlijkheid uit twee componenten bestaat. Aan de ene kant zorgen eerlijke algoritmen voor een gelijke en rechtvaardige verdeling van zowel baten als kosten, en dat beoordelingen door deze algoritmen vrij zijn van oneerlijke vooroordelen, discriminatie en stigmatisering. Aan de andere kant moeten individuen wier leven beïnvloed wordt door algoritmen, in staat zijn om de door algoritmen genomen besluiten aan te vechten.

Verklaarbaarheid is het laatste component van ethische algoritmen. Algoritmische processen moeten transparant zijn, de doelen van deze algoritmen moeten openlijk worden gecommuniceerd en beslissingen genomen op basis van algoritmen – voor zover mogelijk – verklaarbaar zijn voor alle stakeholders. Algoritmen die in het openbaar bestuur worden toegepast, mogen dus geen 'back box' algoritmen zijn die alleen door hun ontwikkelaar begrepen worden. Dit is bijzonder belangrijk omdat zonder deze informatie individuen de algoritmische beslissingen niet kunnen aanvechten.

Vanuit technisch oogpunt is het ook belangrijk dat algoritmen robuust zijn. Dat betekent dat ze beschermd zijn tegen kwetsbaarheden waardoor ze kunnen worden uitgebuit door bijvoorbeeld hackers. Verder is het belangrijk dat ze correcte voorspellingen doen of besluiten nemen. Een hoge mate van nauwkeurigheid is vooral van belang in situaties waarin de algoritmen mensenlevens recht-

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004005

streeks beïnvloeden (High Level Independent Group on Artificial Intelligence, 2019).

Waardenconflicten

De richtlijnen leiden tot de vraag of alle principes tegelijkertijd kunnen worden gerespecteerd. Zijn verklaarbare algoritmen ook de eerlijkste? Voorkomen ze de meeste schade? Iets concreter: verklaarbaarheid wordt van groot belang geacht voor de toepassing van algoritmen in het openbaar bestuur. Maar het heeft ook een belangrijk nadeel. Transparante algoritmen moeten vaak minder ingewikkeld zijn, terwijl 'more complex models enjoy much more flexibility than their simpler counterparts, allowing for more complex functions to be approximated' (Barredo Arrieta et al., 2020, p. 30). De meest nauwkeurige algoritmen zijn meestal niet erg verklaarbaar (bijvoorbeeld diepe neurale netwerken, *random forest* algoritmen en *support vector machines*), en de best interpreteerbare modellen zijn meestal minder nauwkeurig (bijvoorbeeld lineaire of logistische regressie) (Adadi & Berrada, 2018). Dit betekent dat het steeds moeilijker wordt om uit te leggen hoe een algoritme besluiten neemt naarmate de complexiteit van dat algoritme toeneemt. Tegelijkertijd zijn meer complexe algoritmen vaak meer nauwkeurig. Aldus is er een waardenconflict tussen verklaarbaarheid enerzijds en nauwkeurigheid anderzijds.

Ook tussen privacy en eerlijkheid moet vaak een afweging worden gemaakt: als men bepaalde gevoelige of zelfs bijzondere persoonsgegevens voor het trainen van een algoritme niet kan gebruiken, kan er geen rekening worden gehouden met de mogelijke discriminerende werking van algoritmen (Haas, 2019). Clavell, Castillo en Smith (2020) stellen dat een algoritme minder discriminerend wordt als het op data over geslacht, leeftijd, religie en ras getraind wordt (Clavell, Castillo, & Smith, 2020). Gebruik van deze data voor training is potentieel een inbreuk op privacy van individuen als iemands ras of religie als bijzondere persoonsgegevens worden gezien die door de wetgever extra beschermd zijn. De afweging tussen dataminimalisatie en eerlijkheid is hier afhankelijk van het beleid van een organisatie en van de wetgeving. Ook om te testen of een algoritme discriminerend is, moet het worden getest met gegevens die vaak als bron van discriminatie worden gezien, maar de mogelijkheden voor verwerking van deze gegevens zijn vaak beperkt op grond van de privacywet (d.w.z. de regels voor de verwerking van persoonsgegevens van speciale categorieën).

Privacy kan ook op gespannen voet staan met de nauwkeurigheid van algoritmen. Nauwkeurigheid wordt gezien als onderdeel van robuustheid van algoritmen (High Level Independent Group on Artificial Intelligence, 2019). Om de nauwkeurigheid van algoritmen te verbeteren zijn grote hoeveelheden gegevens nodig. In het openbaar bestuur gaat het vaak om persoonsgegevens. Tegelijkertijd kan het buitensporig verzamelen en verwerken van persoonsgegevens als privacyschending worden gezien en gaat dit tegen het principe van dataminimalisatie in, zoals die in de AVG ingevoerd werd. De afweging wanneer de hoeveelheid data die zijn

verwerkt voor de nauwkeurigheid van algoritmen buitensporig wordt, moet door de verantwoordelijke bestuurders worden gemaakt.

De wereld van datagedreven besluitvorming in het openbaar bestuur is vergeven van dergelijke *trade-offs*. Er zijn immers veel publieke waarden in het spel: transparantie, effectiviteit (in al haar gedaanten), privacy, zorgvuldigheid, vertrouwen, et cetera. Zoals de voorbeelden laten zien, is het vaak ingewikkeld om al deze waarden simultaan te waarborgen. Besluitvorming over algoritmen in het openbaar bestuur zal gaan over afwegen: de ene waarde gaat ten koste van de andere (Thacher & Rein, 2004; Stewart, 2006; De Graaf & Van der Wal, 2010). Daarmee rijst de vraag wie die afwegingen maakt. Idealiter worden afwegingen zoals die tussen verklaarbaarheid en nauwkeurigheid door bestuurders genomen die er publiekelijk verantwoording over kunnen afleggen, zeker als de effecten van besluiten ver reiken. In de praktijk is dat uiteraard ingewikkeld, omdat het vermogen van bestuurders en managers om algoritmen te doorgronden beperkt is. Er is veel kennis voor nodig die exclusief ligt bij de mensen die operationeel bij de algoritmen betrokken zijn.

De organisatie van waardenconflicten

Het lijkt eenvoudig om een lijst met waarden te maken die op een bepaald besluit van toepassing zijn, zoals transparantie, veiligheid, duurzaamheid, privacy, et cetera. Een lijst met waarden verheldert de complexiteit van de wens om algoritmen transparant te maken. Er zijn immers meerdere waarden in het spel. Tegelijkertijd verdoezelt een dergelijke lijst complexiteit. Een lijst kan de suggestie wekken dat besluiten over de transparantie van algoritmen een simpele afweging is tussen meerdere waarden. De suggestie wordt al snel gewekt dat een dergelijke afweging op één plek is te maken, bijvoorbeeld door een eindverantwoordelijke die de waardenconflicten erkent, oplost en verantwoordt. Het is echter de vraag of een waardenconflict zo expliciet op een bureau van een eindverantwoordelijke terechtkomt. Een aantal klassieke concepten uit de organisatiekunde doet er sterk aan twijfelen of dit gebeurt.

Ketens. Het idee achter ketenbesluitvorming is dat er een sequentie bestaat tussen activiteit, met een seriële afhankelijkheid. Met behulp van dit concept kunnen we algoritmische besluitvorming zien als een werkproces van data tot een besluit. Een simpele voorstelling: er worden data gegenereerd vanuit verschillende bronnen, deze data worden vergelijkbaar gemaakt, vervolgens worden deze geanalyseerd, geïnterpreteerd en bewerkt tot advies voor besluitvorming (Janssen & Van der Voort, 2016; Van der Voort, Klievink, Arnaboldi, & Meijer, 2019). Algoritmen kunnen in al deze werkprocessen een rol spelen. Ze kunnen bijvoorbeeld sensoren aansturen (genereren van data). De meest besproken algoritmen gaan over data-analyse. Er zijn zoveel data, dat mensen niet meer in staat worden geacht om deze te verwerken. De computer kan dat wel, maar met behulp van algoritmen die van de data interpretabele gegevens maken.

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004005

Maar wat als ieder werkproces door verschillende personen of organisatieonderdelen wordt ondernomen? Dan is het aannemelijk dat zij zelf geconfronteerd worden met verschillende waarden en ieder van hen hierbij hun eigen, bescheiden, afwegingen maakt. Die kunnen technisch lijken, maar significant uitpakken. Zo was er het project om middels data-analyses uit te maken in hoeverre Milaan een geïnternationaliseerde stad was. Daarbij was de gehele keten van data-analist tot en met politieke besluitvormer betrokken. Na presentatie van de resultaten bleek dat de data-analisten de schijnbaar technische keuze hebben moeten maken of het ging om de gemeente of de agglomeratie Milaan. Dit bleek zeer belangrijk voor de uitkomsten (Van der Voort et al., 2019).

Professionaliteit. Algoritmen ontwikkelen is echter een vak dat wordt gedaan door zeer gespecialiseerde medewerkers. Dit bemoeilijkt al te strakke managementmethoden, als men die al zou willen hanteren. In plaats van strakke methoden wordt aan experts bij voorkeur discretionaire ruimte gelaten. Sterker nog, op basis van hun professionaliteit kunnen zij die ruimte claimen (Freidson, 2001; Noordegraaf, 2020). Daar is iets voor te zeggen, omdat de mentale modellen van professionals te complex kunnen zijn om te openbaren. De transactiekosten daarvan zijn eenvoudigweg te groot. Maar het is nu juist deze ruimte die vragen doet rijzen of transparantie haalbaar is. Het eerdergenoemde waardenconflict tussen effectiviteit en transparantie is hiervan een voorbeeld. Een professionele data scientist kan een voorkeur hebben voor neurale netwerken als basis voor een algoritme, maar vanwege de complexiteit van neurale netwerken zijn de algoritmen maar matig transparant. De voorkeur van de data scientist kan voortkomen uit professionele inschattingen van complexiteit en zijn streven naar robuustheid van het algoritme.

Prestatiemeting. Managers en bestuurders worden verantwoordelijk gehouden voor de prestaties en andere effecten van de organisatie, en wensen mede daarom grip te houden op de manier waarop die tot stand komen. Een beproefd middel is de *key performance indicator* (kpi). Die meet de prestaties van processen, deelprocessen of medewerkers. Kpi's kunnen echter een beperkte complexiteit aan. Hoe complexer een taak, hoe moeilijker het is om deze met een kpi te vangen (Okwir, Nudurupati, Ginieis, & Angelis, 2018). Gevolg is dat sommige waarden door kpi's beter worden beloond dan anderen. Daarmee krijgen de medewerkers over de gehele keten prikkels om van tevoren relevant geachte waarden, als geformuleerd in de kpi's, voorrang te geven. Dit kan ten koste gaan van professionele of ethische overwegingen (Kerpershoek, Groenleer, & De Bruijn, 2016). Zo zijn waarden als respect voor autonomie of eerlijkheid moeilijker door middel van kpi's te kwantificeren dan bijvoorbeeld robuustheid en nauwkeurigheid. Tegelijkertijd wordt het opnemen van de waarden voor betrouwbare algoritmen in kpi's door de High Level Independent Group on Artificial Intelligence (2019) gezien als een manier om hun betrouwbaarheid te garanderen.

De genoemde concepten geven weer dat waardenconflicten niet los kunnen worden gezien van de dagelijkse organisatie. Waarden worden door verschillende personen in de organisatie op verschillende wijzen geselecteerd, geïnterpreteerd

en weer gecommuniceerd naar anderen. Een logische conclusie is om de focus niet zozeer te leggen op transparantie van algoritmen, maar op transparantie van de organisatie. Hoe gaat een organisatie met waardenconflicten om? Wie doet dat? En zijn er voldoende checks & balances om goede afwegingen te maken?

De principiële onmogelijkheid van waardenafwegingen

Dus niet alleen de algoritmen, maar vooral de afwegingen tussen waarden, waarvan transparantie er een is, zouden transparant moeten zijn. De waardenafweging is een populair fenomeen in de bestuurskunde. De tragiek ervan is echter dat de metafoor een wensdroom is die ons op een utilistisch dwaalspoor zet. Leg alle ter zake doende waarden op het juiste moment aan de juiste kant van de balans en lees af. Zo ontstaat er een *trade-off*. Maar hoeveel weegt een waarde dan? Op basis van welke schaal? Strepen we dan het gewicht van de ene tegen de andere waarde weg? Waarden en hun gewichten kunnen echter impliciet zijn, hevig ter discussie staan en bovendien om allerlei redenen veranderen. Wanneer is er voldoende kennis over alle consequenties? Niet alle waarden hebben trouwens consequenties. Hoe weeg je waarden af zonder de consequenties te kennen? Wie leest op welk moment de balans eigenlijk af?

Een *trade-off* is een modieuze, maar een al even tragische metafoor. Is er een marktplaats voor waarden? Is het een kwestie van vraag en aanbod? In welke valuta wordt de marktwaarde van een waarde uitgedrukt? Of wordt een ruilhandel van ongelijke grootheden bedoeld? Wie koopt wat bij wie met wat?

Vanuit het theoretisch perspectief van March en Olsen is een inmiddels klassiek onderscheid aan te brengen tussen waardenafweging en belangenafweging. Beide hebben dan een fundamenteel andere logica (March & Olsen, 1996). De belangenafweging baseert zich op een rationeel-utilistische keuze op basis van zoveel mogelijk kennis over de consequenties. Waardenafwegingen zijn meer gevoelsmatige zoektochten die vertrekken vanuit ‘wie ben ik?’ en dat proberen te rijmen met ‘waar ben ik?’. Wellicht ingegeven door de rationele aard van de metafoor wordt met de term waardenafweging in zowel de literatuur als de praktijk gemakshalve meestal een belangenafweging bedoeld.

Van belang bij een waardenconflict, de oorzaak van *trade-offs*, is de aard van een waarde. De ene waarde is de andere niet. Transparantie kan bijvoorbeeld worden getypeerd als een zachte, procedurele, negatieve en continue tussenwaarde. Met andere woorden, typisch een waarde die het structureel aflegt in een afweging met harde, substantiële, positieve en discrete eindwaarden.

Vanuit een empirischer perspectief is gebleken dat er veel meer manieren zijn om tot *trade-offs* te komen dan alleen de waardenafwegingen. Conceptueel gezien maakt een waardenafweging altijd een bewuste koppeling tussen de waarden die conflicteren. Waardenconflicten kunnen zich echter ook onbewust of ontkoppeld oplossen. Een ontkoppelde reactie zien we bijvoorbeeld bij ‘firewalls’, waarbij een

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004005

Tabel 1 ***Manieren om met waardenconflicten om te gaan (Steenhuisen, 2009, p. 31)***

	Gekoppelde waarden	**Ontkoppelde waarden**
Expliciet	Klassiek beeld van wat een waarde-nafweging zou moeten zijn	
Impliciet		Dominante manier waarop *trade-offs* op dagelijkse basis plaatsvinden

enkele waarde in regels en procedures wordt afgeschermd van andere waarden. Denk bijvoorbeeld aan een organisatie die aparte afdelingen heeft per waarde, zoals *safety management* en *security management*. Een onbewuste benadering is bijvoorbeeld 'cycling': vandaag wordt de ene waarde en een maand later de andere geprioriteerd, enzovoort. De effectiviteit en de populariteit van deze ogenschijnlijk minder rationele vormen van conflicten managen blijkt in de praktijk verrassend hoog (Steenhuisen & Van Eeten, 2008). Organisaties zijn verrassend veel met ontkoppelen bezig. Op de noodzaak van *trade-offs* wordt weinig bewust geanticipeerd, enkel in retrospect gereageerd met interventies voor waarden die het blijkbaar nodig hebben. Er is geen minister of NS-directeur die vooraf zegt hoeveel het risico op een treinontsporing mag kosten, bijvoorbeeld. Meer toegepast op algoritmen: vaak wordt niet vooraf bepaald wanneer een fraudedetectie-algoritme nauwkeurig genoeg is en dus geen data meer nodig heeft. Een volledige waardenafweging is een zeldzaamheid. Het meeste dat op een 'bewuste koppeling' lijkt, vindt plaats in de operationele sfeer, waar onder het mom van 'professionaliteit' het beste ervan wordt gemaakt. *Trade-offs* verdienen van nature niet de transparantieprijs.

Het is zoals Oscar Wilde schreef in een toneelstuk: 'Een cynicus kent van alles de prijs en van niks de waarde. Een romanticus kent van alles de waarde maar van niks de prijs.' Met andere woorden, wie overziet aan het eind van de dag de prijs van onze waarden? Niemand misschien.

Maar er is een dieperliggend probleem. Wat is de waarde van een waarde eigenlijk? Waarden zijn theoretisch gezien imponderabilia. Ze zijn van nature incommensurabel (Weiner, 1998; Thacher, 2001). Met andere woorden, echte waarden hebben onderling geen hogere, gedeelde schaal om hun relatieve gewicht vast te stellen. De impact van algoritmische besluiten op mensenlevens laat zich daarom niet uitdrukken in een budget of een beter betrouwbaarheidspercentage, althans niet zonder we ons daar moreel gezien tegen verzetten. Met andere woorden, een waarde is iets dat je in de regel niet wil, niet kan en niet mag afwegen. Maar ook al kun je de waardenafweging filosofisch gezien als onmogelijk afdoen, in de dagelijkse praktijk zijn *trade-offs* schering en inslag.

De roep om transparantie en de dynamiek van waardenafwegingen

In discussies over het gebruik van algoritmen in het openbaar bestuur is de transparantie ervan een heikel agendapunt. Een sterke roep om meer transparantie is, echter, om drie redenen licht tragisch te noemen, wegens de dynamiek van *trade-offs*.

Ten eerste, de roep is een ontkoppeling op zich. Het prioriteren van een enkele waarde, zonder gevolgen voor andere waarden te overzien, laat staan mee te wegen, geeft gemakkelijk aanleiding tot andere ontkoppelde en onbewuste *trade-offs*. En dat geldt zeker voor de waarde transparantie, die in principe altijd en overal relevant is. Professionals in de operationele processen of hun managers zullen op hun beurt bewust of onbewust tegenwicht denken te moeten bieden als 'hun' waarden door de prioritering van transparantie in de knel komen.

Ten tweede meet de roep om transparantie met twee maten. Algoritmen hebben de schijn soms tegen als het om transparantie gaat. Maar is de werkwijze met algoritmen minder transparant dan de 'ouderwetse' werkwijze? Wellicht zijn de niet-transparante processen in de keten verplaatst of meer zichtbaar geworden. Een algoritme dat leidt tot surveillance in telkens dezelfde wijk, kan discriminatoir zijn. Er zijn goede redenen om daarnaar te kijken. Maar datzelfde kan gelden voor de mentale modellen van de professionals die surveilleren. Die zijn wellicht veel minder makkelijk expliciet te maken. Transparantie is wellicht altijd al laag of zelfs lager geweest.

Ten derde is de roep om meer transparantie, als de roep tot het borgen van een enkele waarde, vaak het resultaat van escalatie. Zodra een waarde in de knel komt, kan het probleem worden gesignaleerd, benoemd en gerepareerd door professionals en eventueel hun directe managers. Blijft de wanprestatie lang duren, dan kan de escalatie tot en met de politiek doorklimmen. Voor de prioriteit van een enkele kwetsbare waarde kan dit een reddingsboei zijn. Voor de transparantie van *trade-offs* echter is deze 'molen' vaker niet dan wel behulpzaam. Het omhoog tillen van *trade-offs*, uit de operationele haarvaten, leidt in de regel tot minder begrip en een meer monomane bescherming van kwetsbare waarden als reactie. Meer transparantie maar onverminderd hoge prestatie-eisen voor vele andere waarden, is een recept voor meer onbewuste, ontkoppelde *trade-offs*. De ironie is dus dat escalatie in het geval van een wanprestatie juist de zichtbaarheid en de realiteit van *trade-offs* onder druk zet.

Conclusie: van trade-offs in een organisatie naar de organisatie van trade-offs

Transparantie en betrouwbaarheid van algoritmen betekent 'alle hens aan dek' voor veel betrokken actoren in een organisatie: data-analisten, uitvoerders, managers en bestuurders. Het gaat namelijk niet enkel om het transparant maken van een formule door het bijvoorbeeld te publiceren. Het gaat ook over transparantie van de manier waarop transparantie wordt afgewogen tegen andere waarden,

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004005

zoals effectiviteit en privacy. Dat gebeurt in de context van een organisatie. Daarbij zijn vragen relevant als 'Wie weegt af?', 'Hoe doen zij dat?' en 'Waarom?'. Daarmee verschuift de discussie over transparantie van algoritmen naar de transparantie van de organisatie.

En dan nog... Met de antwoorden op deze 'recht-toe-recht-aan-vragen' weten we nog te weinig over transparantie van algoritmen. De gedachte dat waardenconflict, zoals tussen transparantie en effectiviteit, kan worden opgelost met waardenafwegingen, is wellicht te simpel. Afwegingen zijn niet zelden impliciet, zelfs voor degene die de afwegingen maakt. Bovendien voltrekken afwegingen zich zelden op een enkel moment op een enkele plaats. Transparantie als waarde kan worden ontkoppeld van andere waarden. Dat gaat al heel snel. Maak iemand exclusief verantwoordelijk voor transparantie, en het 'leed' is al geschied. Andere waarden doen voor deze persoon al snel minder ter zake. Wellicht is een andere persoon verantwoordelijk voor privacy of effectiviteit. Ieder z'n waarde. De echte afweging tussen waarden wordt elders gemaakt. Of niet. Op deze manier kunnen waardenafwegingen worden 'weggeorganiseerd'. De 'recht-toe-recht-aan-vragen' adresseren dergelijke processen maar beperkt. Relevante extra vragen zijn 'Hoe zijn verschillende waarden geïnstitutionaliseerd?', 'Zijn waardenafwegingen kenbaar?', 'Zijn professionals en bestuurders zich bewust van deze afwegingen' en 'Waarom wel/niet?'.

Zo komen we op een mogelijke empirische onderzoeksagenda. Onderzoek naar waardenafwegingen in ketens kan veel verhelderen. Waar in de keten van data-analist naar besluitvormer vinden wat voor afwegingen plaats? Hoe determineren professionele, ogenschijnlijk technische, afwegingen de besluitvorming aan het einde van de keten? Een *process tracing*-benadering (Collier, 2011) kan hier interessant zijn, bijvoorbeeld door na te gaan op basis van welke informatie, en welke data, een van tevoren geselecteerd besluit is genomen. Andersom kan, met de nodige technische kennis, het verhaal van een enkele, technische afweging door data-analisten worden verteld. Welke invloed heeft de afweging gemaakt verderop in de keten tot en met besluitvorming?

De vraag naar meer transparantie over het gehele ketenproces roept ook vragen op rondom professionalisering van data-analisten die algoritmen ontwerpen. Tegenwoordig spelen zij een centrale rol bij transparantie van algoritmen – zij moeten de algoritmen kunnen (laten) uitleggen. Maar zij zijn ook professionals, met een soms legitieme hang naar autonomie. Hun rol in de transparantie over *trade-offs* binnen organisaties kan worden onderzocht met een etnografie. Daarmee komen hun soms impliciete waardenafwegingen beter in beeld. Daarmee worden ook de mogelijkheden en onmogelijkheden van transparante algoritmen duidelijker. Als extraatje is de vraag interessant of hun professionaliteit vergelijkbaar is met de professionaliteit van uitvoerders. Immers, het begrip 'professionaliteit' wordt veelal op uitvoerenden toegepast, en veel minder op ondersteunende functies zoals 'data-analyse'.

Ten slotte kunnen de ideeën over waardenafwegingen in organisaties als inspiratie voor de wetgever dienen. Terwijl huidige richtlijnen zoals de richtlijnen voor betrouwbare kunstmatige intelligentie door de Europese Commissie de voorwaarden voor betrouwbare algoritmen vastleggen, zal er meer aandacht worden geschonken aan de toepassing van deze waarden in organisaties. Zo kunnen bij het ontwikkelen van algoritmen bepaalde basiswaarden in het systeem worden ingebakken.

Literatuur

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Barredo Arrieta, A., Díaz-Rodríguez, N., Ser, J. Del, Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012

Clavell, G. G., Castillo, C., & Smith, O. (2020). *Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization*. 265–271.

Collier, D. (2011). Understanding process tracing. *Political Science & Politics, 44*(4), 823–830.

Das, A., & Rad, P. (2020). *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey*.

De Graaf, G., & Van der Wal, Z. (2010). Managing conflicting values in public policy. *The American Review of Public Administration, 40*(6), 623–630.

Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics, 21*, 402–418. https://doi.org/10.1007/s10892-017-9252-2

Freidson, E. (2001). *Professionalism: The third logic*. Chicago: University of Chicago Press.

Haas, C. (2019). *The Price of Fairness - A Framework to Explore Trade-Offs in Algorithmic Fairness*.

High Level Independent Group on Artificial Intelligence. (2019). Ethics Guidelines for Trustworthy AI. In *European Commission*.

Janssen, M., & Van der Voort, H. (2016). Big data klaar voor gebruik? *Bestuurskunde, 1*, 16–20.

Kerpershoek, E., Groenleer, M., & De Bruijn, H. (2016). Unintended responses to performance management in Dutch hospital care: Bringing together the managerial and professional perspectives. *Public Management Review, 18*(3), 417–436.

Leenes, R. (2016). De voorspellende overheid. *Bestuurskunde*, (1), 38–43.

March, J. G., & Olsen, J. P. (1996). Institutional perspectives on political institutions. *Governance, 9*(3), 247–264.

McCrudden, C. (2008). Human Dignity and Judicial Interpretation of Human Rights. *The European Journal of International Law, 19*(4), 655–724. https://doi.org/10.1093/ejil/chn043

Noordegraaf, M. (2020). Protective or connective professionalism? How connected professionals can (still) act as autonomous and authoritative experts. *Journal of Professions and Organization, 7*(2), 205–223.

Okwir, S., Nudurupati, S., Ginieis, M., & Angelis, J. (2018). Performance measurement and management systems: a perspective from complexity theory. *International Journal of Management Reviews, 20*(3), 731–754.

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004005

Preece, A., Harborne, D., Braines, D., Tomsett, R., & Cakraborty, S. (2018). Stakeholders in Explainable AI. *AAAI FSS-18: Artificial Intelligence in Government and Public Sector*.

Steenhuisen, B. (2009). *Competing Public Values: Coping strategies in heavily regulated utility industries*.

Steenhuisen, Bauke, & van Eeten, M. (2008). Invisible Trade-Offs of Public Values: Inside Dutch Railways. *Public Money & Management, 28*(3), 147–152. https://doi.org/10.1111/j.1467-9302.2008.00636.x

Stewart, J. (2006). Value Conflict and Policy Change. *Review of Policy Research, 23*(1), 183–195. https://doi.org/10.1111/j.1541-1338.2006.00192.x

Thacher, D. (2001). Conflicting values in community policing. *Law and Society Review, 35*(4), 765–798.

Thacher, David, & Rein, M. (2004). Managing Value Conflict in Public Policy. *Governance, 17*(4), 457–486. https://doi.org/10.1111/j.0952-1895.2004.00254.x

van der Voort, H. G., Klievink, A. J., Arnaboldi, M., & Meijer, A. J. (2019). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly, 36*(1). https://doi.org/10.1016/j.giq.2018.10.011

Weiner, M. (1998). The clash of norms: dilemmas in refugee policies. *Journal of Refugee Studies, 11*(4), 432–453.