Verantwoorde algoritmisering: zorgen waardengevoeligheid en transparantie voor meer vertrouwen in algoritmische besluitvorming?*

Stephan Grimmelikhuijsen & Albert Meijer

Algoritmen nemen een steeds prominentere rol in binnen het openbaar bestuur. Algoritmen zouden objectievere en efficiëntere beslissingen kunnen nemen dan mensen. Echter, recente schandalen laten zien dat er nog veel problemen kleven aan het gebruik van algoritmen bij de overheid. Vanuit ethische hoek wordt steeds meer nadruk gelegd op eisen die worden gesteld aan algoritmen en wij willen deze eisen koppelen aan bestuurswetenschappelijke inzichten over het gebruik van technologie in de publieke sector. Wij pleiten daarom voor verantwoorde algoritmisering – verantwoorde organisatorische praktijken rondom het gebruik van algoritmen – en betogen dat dit nodig is om het vertrouwen van burgers te behouden. Wij onderscheiden twee pilaren van verantwoorde algoritmisering – waardengevoeligheid en transparantie – en laten zien hoe elk van deze pilaren samenhangt met algoritmisering en op welke manieren dit zou kunnen bijdragen aan vertrouwen. We eindigen het artikel met een agenda voor empirisch onderzoek naar verantwoorde algoritmisering en vertrouwen.

Inleiding

Algoritmen, in het bijzonder *machine learning*-algoritmen, zijn al jaren bezig aan een gestage opmars in ons dagelijks leven. Meer recentelijk hebben ook overheidsorganisaties de mogelijkheden van *machine learning*-algoritmen en big data ontdekt.[1] Op allerlei beleidsterreinen worden dergelijke algoritmen ingezet (Zouridis, Van Eck, & Bovens, 2020). Zo kan de politie met behulp van een 'slim' algoritme beter voorspellen waar een verhoogd niveau van criminele activiteit zal plaatsvinden (Meijer & Wessels, 2019) en gebruikt het Centraal Orgaan opvang Asielzoekers (COA) slimme algoritmen om te voorspellen in welke stad een immigrant het best tot zijn recht komt (Van Wijnen, 2019). In een recente verkenning

* Dr. S.G. Grimmelikhuijsen is universitair hoofddocent Publiek Management aan de Universiteit Utrecht, Departement Bestuurs- en Organisatiewetenschap. Prof. dr. A.J. Meijer is hoogleraar Publiek Management aan de Universiteit Utrecht, Departement Bestuurs- en Organisatiewetenschap.

1 De term 'machine learning' (ML) is in 1959 gemunt door AI-pionier Arthur Samuel. Waar een algoritme kan worden gedefinieerd als 'a finite set of rules that provide a sequence to solve a specific task or certain problem', is machine learning het uitvoeren van deze taak zonder expliciet hiervoor geprogrammeerd te zijn (Samuel, 1959).

constateerde het CBS dat inmiddels meer dan de helft van de overheidsinstanties 'iets' met *machine learning*-algoritmen doet (CBS, 2018).

Deze snelle opkomst voor algoritmen brengt een grote belofte met zich: algoritmen kunnen sommige patronen sneller herkennen dan mensen van vlees en bloed en een algoritme handelt – ogenschijnlijk – zonder willekeur of vooroordeel. Tegelijkertijd noemen critici belangrijke nadelen van algoritmen in het openbaar bestuur: algoritmen zijn net zo bevooroordeeld als de onderliggende data en kunnen inbreuk maken op privacy. Los van het leed dat individuen wordt aangedaan door bijvoorbeeld verdachtmakingen op onterechte gronden en privacyschendingen, kunnen deze nadelen van het gebruik van algoritmen leiden tot een erosie van vertrouwen in de overheid.

In dit artikel betogen wij dat verantwoorde algoritmisering nodig is om het vertrouwen van burgers in dergelijke overheidsbesluiten te behouden. Het bekendste voorbeeld in Nederland dat laat zien hoe onverantwoorde algoritmen kunnen leiden tot afname van dat vertrouwen is het Systeem Risico Indicatie (SyRI). SyRI werd officieel gelanceerd in 2014 door het ministerie van Sociale Zaken en Werkgelegenheid als technologisch instrument voor de bestrijding van fraude. Het systeem kan door verschillende publieke organisaties worden gebruikt om belastingfraude, uitkeringsfraude en arbeidsfraude te monitoren. Het systeem bevat een grote database met onder andere data over de financiële situatie van burgers, hun uitkeringsgeschiedenis, opleidingsniveau en integratie in de Nederlandse samenleving. Het risicomodel dat ten grondslag ligt aan SyRI, bepaalde op basis van de data of burgers een hoog of laag risico op fraude geven.

Hoewel dergelijke systemen voor risicobepaling lange tijd weinig publieke aandacht kregen, raakte SyRI zeer omstreden. Verschillende maatschappelijke organisaties en individuele burgers kwamen in verzet tegen dit algoritme en spanden een rechtszaak aan. SyRI is inmiddels verboden door de rechter omdat dit systeem volgens de rechter potentiële uitkeringsfraudeurs op discriminerende gronden uitlichtte. Daarnaast was de werking van het algoritme en de onderliggende data niet transparant en dit past niet in een democratische rechtsstaat (Van Schendel, 2019). Met het gebruik van algoritmen zoals SyRI riskeren overheden het vertrouwen van burgers te verliezen. Om vertrouwen van burgers in het gebruik van algoritmen te creëren is er daarom een steeds sterkere roep om ethische eisen aan algoritmen te stellen (Mittelstad, Allo, Taddeo, Wachter & Floridi, 2016; Ananny, 2016): er is een noodzaak tot verantwoorde algoritmisering.

Verantwoorde algoritmisering kan worden begrepen als een specifieke manifestatie van wat door Owen, Macnaghten, Stilgoe (2012, pp. 757-758) verantwoorde innovatie wordt genoemd. Zij omschrijven dit als een verplichting tot zorgvuldige innovatie op basis van responsiviteit ten opzichte van de omgeving, ethische oordeelsvorming en kennis van percepties en relevante feiten. Op basis hiervan kunnen we verantwoorde algoritmisering als volgt omschrijven: *Verantwoorde algoritmisering verwijst naar het afwegen van ethische dilemma's bij het gebruik van*

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004002

algoritmen in organisaties, gebaseerd op responsiviteit ten opzichte van de omgeving en kennis van de (mogelijke) impact.

In dit artikel beargumenteren we dat verantwoorde algoritmisering kan bijdragen aan het vertrouwen van burgers. Daarbij bouwen we voort op debatten over ethische eisen aan algoritmen (Mittelstad et al., 2016; Ananny, 2016) en combineren deze met kennis over het gebruik van technologie in de publieke sector (Orlikowski, 1992; Zuurmond, 1994; Gil-Garcia, 2012). Daarbij stellen we niet de technologie an sich centraal, maar het gebruik van de technologie in een organisatie. 'Algoritmisering' gaat verder dan alleen het introduceren van een algoritme: een algoritme krijgt immers pas betekenis in het gebruik in een organisatie. Algoritmisering heeft daarom betrekking op de ontwikkeling, implementatie en het gebruik van algoritmen. Op al deze facetten dient er sprake te zijn van verantwoord gebruik.

We maken deze claim preciezer door in dit artikel in eerste instantie de term 'algoritmisering' te preciseren aan de hand van klassiek werken over informatisering van organisaties. Vervolgens gaan wij expliciet in op de vraag wat nu vertrouwen van burgers is om goed te begrijpen hoe dit kan worden geschaad door het gebruik van algoritmen. Daarna werken wij op basis van bovengenoemde definitie twee aspecten van verantwoorde algoritmisering uit, te weten *waardengevoeligheid* en *transparantie*. Wij laten zien op welke wijze waardengevoelige en transparante algoritmisering kan bijdragen aan vertrouwen in overheidsorganisaties. We eindigen het artikel met een agenda voor empirisch onderzoek naar verantwoorde algoritmisering en vertrouwen.

Algoritmisering: meer dan alleen een technisch issue

Vaak wordt verantwoord gebruik van algoritmen vanuit een technologisch perspectief benaderd. In dit geval gaat het dan vooral over hoe men algoritmen kan ontwerpen op een manier dat deze zonder vooroordelen en accuraat werken. Dit is uiteraard van groot belang, maar met alleen het technologisch perspectief zijn we nog niet bij een verantwoord gebruik van deze algoritmen. De organisatorische praktijken rondom algoritmen dragen ook bij aan het (on)verantwoorde gebruik ervan. Immers, algoritmen krijgen pas betekenis wanneer zij worden geïmplementeerd in organisaties en worden gebruikt door medewerkers. De techniek en de inbedding ervan zijn niet los van elkaar te beschouwen (Orlikowski, 1992; Gil-Garcia, 2012).

Daarom hanteren we hier de term 'algoritmisering' om te duiden dat het ons gaat om een organisatorisch proces en niet alleen een losgezongen nieuwe technologie. We bouwen voort op vroeg wetenschappelijk werk over informatisering in het openbaar bestuur van Arre Zuurmond (1994, pp. 42-48) en onderscheiden de onderstaande dimensies van algoritmisering:

1 *Technologie*. Het proces van algoritmisering begint met de introductie van een nieuwe technologie in een organisatie. Het algoritme kan werken in een min

of meer losstaand beslisondersteuningssysteem of volledig geïntegreerd zijn in de organisationele infrastructuur.

2 *Technische expertise*. Om een algoritme te kunnen gebruiken is een diversiteit aan expertise noodzakelijk. Experts die weten hoe een algoritme of systeem werkt, zijn nodig om technische problemen op te lossen en om de algoritmen te onderhouden en updaten.

3 *Informatie-relaties*. Slimme algoritmen worden gevoed door data (vaak *big data*) en produceren nieuwe soorten informatie en patronen. Dit betekent dat de informatie-relaties in en om een overheidsorganisatie zullen veranderen.

4 *Organisatiestructuur*. Als gevolg van de veranderende informatie-relaties ontstaan er nieuwe samenwerkingen en nieuwe verhoudingen tussen afdelingen of departementen. Zo kunnen afdelingen die de informatie en algoritmen beheren, een sterkere en meer coördinerende rol in de organisatie krijgen.

5 *Informatiebeleid*. Als gevolg van de introductie van een algoritme en het belang ervan zullen overheidsinstanties zich meer en meer gedwongen voelen om beleid te ontwikkelen over het gebruik ervan. Algoritmen zijn van strategisch belang, net als geld, personeel en bevoegdheden en moeten 'met beleid' worden gebruikt en onderhouden.

6 *Monitoring en evaluatie*. In navolging van het beleid zullen organisaties methoden en systemen opzetten om te kijken of algoritmen nuttig en effectief maar ook verantwoord en legitiem zijn. Algoritmen en de systemen waar zij onderdeel van zijn, zullen van tijd tot worden geëvalueerd en waar nodig worden aangepast.

De dimensies hangen sterk met elkaar samen, maar kunnen deels ook los van elkaar bewegen. Zo zullen meer technologische expertise en staand informatiebeleid nodig zijn om een goed monitoring- en evaluatiesysteem te ontwikkelen. Tegelijkertijd betekent een goed informatiebeleid niet automatisch dat er ook geëvalueerd wordt en kan een hoge mate van technologische expertise ook ongecontroleerd blijven. De samenhang tussen de zes dimensies laat tevens zien hoe complex het verantwoordelijkheidsvraagstuk eigenlijk is en het onderstreept in elk geval ons punt dat we breder moeten nadenken over verantwoord gebruik van algoritmen en dit dus niet (alleen) aan de *techies* moeten overlaten.

We kunnen het voorbeeld van het gebruik van het Criminaliteit Anticipatie Systeem (CAS) door de politie in Amsterdam gebruiken om deze verschillende elementen toe te lichten. Met behulp van het CAS voorspelt de politie *high impact crimes*, zoals woninginbraken en overvallen. Om dit te doen wordt de kaart van Amsterdam opgedeeld in vierkantjes van 125 bij 125 meter. Voor deze vakjes verzamelt de politie data over zaken zoals criminaliteitshistorie, afstand tot bekende verdachten, afstand tot de dichtstbijzijnde snelwegoprit en socio-economische gegevens (Willems & Doeleman, 2014). Op basis van deze informatie voorspelt het algoritme van CAS de kans op een incident in een bepaalde periode. Gebieden met een hoge kans op een *high impact crime* kleuren rood en gelden als een risicogebied.

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004002

Het systeem is ontwikkeld om de besluitvorming over de inzet van de politie te verbeteren. Daarmee vereist het systeem enerzijds expertise op het gebied van *decision-support systems* binnen de politie en tegelijkertijd verandert ook de expertise van informatiespecialisten want zij moeten de uitkomsten van het CAS kunnen interpreteren. Voor het systeem maakt de politie niet alleen gebruik van allerlei politiedata maar ook openbare data over bijvoorbeeld de socio-economische kenmerken van wijken. Het gebruik van het CAS wordt geïntegreerd in overlegstructuren – *briefings* – om politieagenten te informeren en de invulling van deze overlegstructuren wordt ook aangepast op het gebruik van het CAS. Voor het gebruik van het CAS heeft de politie ook specifiek beleid ontwikkeld. En het CAS wordt gemonitord op het functioneren van de techniek (*back end*), het functioneel gebruik van het interface en de mate waarin het wordt gebruikt door de informatiespecialisten. Daarmee heeft dit systeem op allerlei manieren invloed op de wijze waarop de Amsterdamse politie omgaat met haar informatie en planning en raakt dit het werk van allerlei medewerkers op verschillende manieren.

Het voorbeeld laat zien dat de introductie van een algoritme leidt tot een verschuiving in de werkwijze ten aanzien van de wijze waarop agenten op straat worden aangestuurd. Deze verschuiving kan grote betekenis hebben voor de relatie tussen de organisatie en de omgeving. Wij zijn hierbij met name geïnteresseerd in de vraag op welke wijze dit invloed heeft op het vertrouwen van burgers. Onze veronderstelling is dat specifieke eisen aan de invulling van informatisering – verantwoorde informatisering – nodig zijn om het vertrouwen van burgers te behouden.

Vertrouwen: ziet de burger de overheid als competent, welwillend en integer?

Vertrouwen van burgers in de overheid is onderwerp van veel wetenschappelijke literatuur in de politicologie en bestuurskunde. Door de veelheid aan publicaties en definities lijkt het lastig om een rode draad of algemene definitie te ontdekken. Toch zien we over disciplines heen dat er voor vertrouwen een tamelijk breed gedeelde definitie is; deze zullen wij ook in dit artikel hanteren. Volgens Rousseau, Sitkin, Burt & Camerer, 1998, p. 395) is vertrouwen*:* 'a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another'.

Kern van deze definitie wordt gevormd door de 'positieve verwachtingen' die iemand heeft met betrekking tot de intenties of het gedrag van een ander. Die 'ander' is in ons geval de overheid. De positieve verwachtingen die men kan hebben, worden gevormd door de mate waarin iemand een overheidsinstantie als betrouwbaar ervaart. Er is veel onderzoek gedaan naar de constituerende dimensies van betrouwbaarheid en over het algemeen worden de volgende drie het meest genoemd als dimensies van betrouwbaarheid (zie bijvoorbeeld Mayer, Davis, & Schoorman, 1995; McEvily & Tortoriello, 2011; Grimmelikhuijsen & Knies, 2017):

- *Gepercipieerde competentie*: deze dimensie heeft betrekking op de vraag in hoeverre een burger de overheid ziet als capabel, professioneel en effectief in het oplossen van problemen.
- *Gepercipieerde welwillendheid*: de gepercipieerde welwillendheid beziet iemands beeld van de intenties van een overheidsinstantie. Welwillendheid gaat over de vraag of burgers de indruk hebben dat de overheid werkt in het belang van burgers en het publieke belang in meer algemene zin.
- *Gepercipieerde integriteit*: de derde dimensie heeft betrekking op de integriteit van een overheidsinstantie (soms eerlijkheid genoemd in de literatuur). Een overheid wordt gezien als integer wanneer deze als oprecht en waarheidsgetrouw wordt gezien.

Een interessante achterliggende vraag hier is in hoeverre percepties van betrouwbaarheid – het stellen van vertrouwen in een algoritme – overeenkomen met de daadwerkelijke betrouwbaarheid ervan. Met andere woorden, een algoritme kan ongerechtvaardigd worden vertrouwd door burgers (en ambtenaren!), omdat niemand precies weet hoe het algoritme en de onderliggende data in elkaar steken. Juist het SyRI-incident laat zien wat voor desastreuze gevolgen dit heeft voor vaak kwetsbare burgers. Zoals we hebben gezien in het voorbeeld van SyRI, kunnen de grote verschuivingen in het functioneren van publieke organisaties met name de gepercipieerde welwillendheid en integriteit negatief beïnvloeden, juist wanneer men niet kritisch genoeg is geweest op de betrouwbaarheid van een algoritme. Het gestelde vertrouwen en de feitelijke betrouwbaarheid van algoritmen moeten overeenkomen om een uiteindelijke breuk in vertrouwen te voorkomen.

Tegelijkertijd is het niet zo dat ieder incident met algoritmen per definitie leidt tot minder vertrouwen. In de wetenschappelijke literatuur worden diverse zaken naar voren gebracht die reputatieverlies kunnen voorkomen, zoals het succesvol ontwijken van de schuldvraag (Hood, 2007) of, wanneer we wat optimistischer van aard zijn: transparante en tijdige communicatie (Coombs, 2007). Zo laten Grimmelikhuijsen, De Vries en Zijlstra (2018) zien dat wanneer een organisatie haar eigen fouten toegeeft in een perscommuniqué, er slechts beperkt vertrouwensverlies optreedt. Met deze nuancering in het achterhoofd zullen herhaalde incidenten en crises een risico vormen voor het vertrouwen op de lange termijn. Wanneer een overheid niet op verantwoorde wijze gebruikmaakt van algoritmen, is het waarschijnlijk dat incidenten zoals SyRI zich vaker zullen voordoen en dus een risico vormen voor het vertrouwen van burgers in de overheid in het algemeen, en in algoritmen in het bijzonder.

Verantwoorde algoritmisering is daarom noodzakelijk om het vertrouwen van burgers in de overheid te behouden. Op basis van de literatuur onderscheiden wij twee pilaren van verantwoorde algoritmisering: waardengevoeligheid en transparantie.

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004002

Tabel 1 Richtinggevende vragen voor waardengevoelige algoritmisering

Dimensie van algoritmisering	Richtinggevende vraag
Technologie	Hoe worden de verschillende waarden van het specifieke toepassingsgebied geïdentificeerd en meegenomen in het ontwikkelen van het algoritme?
Technische expertise	Zijn de technische experts die het algoritme ontwikkelen en onderhouden, zich bewust van verschillende waarden en kunnen zij waardengevoelige keuzes herkennen?
Informatie-relaties	Worden waardengevoelige keuzes inherent aan de onderliggende data en patronen die naar voren komen uit het combineren van datasets, erkend?
Organisatiestructuur	Is de verantwoordelijkheid voor de afweging van waarden bij het gebruik van algoritmen in een duidelijke positie in de organisatie belegd?
Informatiebeleid	Heeft de organisatie een beleid waarin wordt omschreven hoe wordt omgegaan met waardengevoelig gebruik van algoritmen?
Monitoring en evaluatie	Heeft de organisatie een systeem voor het monitoren en evalueren van uitkomsten van algoritmen waarbij rekening wordt gehouden met diverse waarden?

Pilaar 1: Verantwoorde algoritmisering door waardengevoeligheid

De eerste pilaar van verantwoorde algoritmisering betreft het introduceren van waardengevoeligheid bij het ontwikkelen van algoritmen. Dit klinkt logisch, maar is in de kern verschillend van hoe er doorgaans naar technologische innovaties gekeken wordt. In eerste instantie is de dominante focus op innovaties de potentiële winst in efficiëntie en effectiviteit, waarbij andere waarden uit het oog worden verloren of zelfs worden gezien als barrière voor vernieuwing. De literatuur over waardengevoeligheid – *value sensitive design* (Friedman, 1996; Van den Hoven, 2007; Friedman et al., 2008) – benadrukt dat aandacht voor deze andere waarden cruciaal is.

Friedman, Kahn & Borning (2008, p. 69) definiëren waardengevoelig ontwerp als: '(...) a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process.' Voor een waardengevoelige algoritmisering in het openbaar bestuur moeten we begrijpen welke waarden een rol spelen in algoritmische besluitvorming. Hierbij is er dus een noodzaak tot responsiviteit ten opzichte van de omgeving: er moet rekening worden gehouden met verschillende belanghebbenden bij een besluit en welke waarden voor hen belangrijk zijn. Deze waarden moeten worden geïdentificeerd en vervolgens moet er worden gekeken hoe deze worden meegenomen in het proces van algoritmisering, dat wil zeggen in expertise, informatiebeleid, structuur et cetera. Tabel 1 laat zien welke richtinggevende vragen bij iedere dimensie van algoritmisering van belang zijn.

We verwachten dat ieder element van waardengevoelige algoritmisering invloed heeft op de drie dimensies van vertrouwen in de overheid. Zo is te verwachten dat wanneer er rekening wordt gehouden met de waarden van relevante belanghebbenden, zoals uitkeringsontvangers, de kans kleiner is dat algoritmen zodanig ontwikkeld en gebruikt worden dat zij kwetsbare burgers op oneerlijke of discriminerende wijze benadelen. Denk hierbij aan het algoritme dat werd gebruikt door SyRI waardoor een groep toch al kwetsbare burgers extra hard werden geraakt, doordat zij op oneerlijke wijze door het systeem waren geoormerkt als mogelijke bijstandsfraudeur. Wanneer er bij de ontwikkeling rekening was gehouden met een waarde als gelijkheid, had een algoritme wellicht deze uitwerking niet gehad. Ook was de kans op dit drama veel kleiner geweest wanneer er door goede evaluatie bewuster gekeken zou zijn naar een diversiteit aan waarden in plaats van alleen effectiviteit.

Kortom, wanneer er gebrek aan waardengevoeligheid is, kan algoritmisering desastreus zijn voor het vertrouwen van burgers, en dan in het bijzonder de ethische dimensies van vertrouwen. Wanneer door waardengevoelige algoritmisering burgers op een eerlijkere en minder bevooroordeelde wijze behandeld worden, zullen zij een overheidsinstantie naar verwachting als meer integer en welwillend (in het belang van de burger) ervaren. Voor het CAS is het zeer belangrijk dat helder kan worden aangetoond dat het systeem niet discriminerend werkt richting bepaalde groepen of bepaalde wijken.

Ten slotte zouden twee aspecten van algoritmisering kunnen bijdragen aan vertrouwen in de competentie (percepties van efficiëntie, professionaliteit) van een overheidsinstantie. We verwachten dat het ontwikkelen van informatiebeleid en evaluatiestructuren bijdraagt aan meer weloverwogen keuzes met betrekking de werking van het algoritme. Dit draagt bij aan het idee dat overheden op een rationele en professionele wijze omgaan met algoritmen in hun organisatie. Tegelijkertijd bestaat het risico dat het opstellen van dergelijk beleid de verwachtingen ten aanzien van het gebruik van algoritmen verder kan opschroeven, hetgeen kan leiden tot meer teleurstelling wanneer hieraan niet kan worden voldaan (vergelijk Grimmelikhuijsen & Porumbescu, 2017).

Samenvattend vraagt de eerste pilaar van verantwoorde algoritmisering om een grotere mate van aandacht voor de waardengevoeligheid in alle facetten van algoritmisering: niet alleen bij de ontwikkeling van het algoritme, maar ook in alle andere organisationele aspecten moet een diversiteit aan waarden worden erkend. Op deze manier wordt het risico op bevooroordeelde of oneerlijke algoritmische besluitvorming verkleind en blijft het vertrouwen behouden.

Pilaar 2: Verantwoorde algoritmisering door transparantie

Een tweede pilaar van verantwoorde algoritmisering is transparantie. Transparante algoritmisering komt in de literatuur niet aan de orde, maar algoritmische transparantie is in de academische literatuur inmiddels vaak besproken en wordt

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004002

gezien als een belangrijke manier om ervoor te zorgen dat overheden en ontwikkelaars verantwoording afleggen over algoritmen (Diakopoulos, 2016). We nemen dit begrip daarom als uitgangspunt en vertalen het naar transparante algoritmisering.

Algemeen wordt erkend dat de transparantie van veel algoritmen zwaar tekortschiet (Lepri, Oliver, Letouzé, Pentland & Vink, 2018. Maar wat is algoritmische transparantie eigenlijk? En op welke wijze zou dit kunnen bijdragen aan vertrouwen? We definiëren algoritmische transparantie als volgt: *Een algoritme is transparant als externe actoren de mogelijkheid hebben om toegang te krijgen tot een algoritme om deze te beoordelen en wanneer de uitkomsten van een algoritme op begrijpelijke wijze interpreteerbaar zijn voor mensen.* Deze definitie bestaat uit twee kerncomponenten: toegankelijkheid en interpreteerbaarheid. Met name bepaalde soorten *machine learning*-algoritmen schieten tekort op deze dimensies. Ten eerste zijn beslisuitkomsten vaak moeilijk uit te leggen en dus moeilijk interpreteerbaar. Voor burgers die te horen krijgen dat ze, bijvoorbeeld, verdacht worden bijstandsfraude, is het cruciaal om te weten *waarom* een algoritme deze aanbeveling heeft gedaan. Omdat *machine learning*-algoritmen vaak zelf uit patronen en combinaties van data uitkomsten destilleren en deze patroonherkenning vaak niet voor mensen te interpreteren is (Burrell, 2016), schiet deze dimensie er vaak bij in. Daarom wordt in AI-onderzoek nu veel aandacht besteed aan de manier waarop algoritmen goede interpreteerbare uitleg kunnen genereren (Guidotti et al., 2018; Mittelstadt et al., 2019).

Ook de andere dimensie, toegankelijkheid, is vaak problematisch. Vaak wordt de broncode van algoritmen niet prijsgegeven omdat hier patent op rust, of uit angst dat als men de parameters van een algoritme kent, deze ook bespeeld kunnen worden (Mittelstadt et al., 2016). Om die reden zijn algoritmen vaak niet toegankelijk. Daarnaast is een toegankelijk algoritme voor de gemiddelde burger niet zo waardevol omdat zij niet de kennis in huis hebben om een complex algoritme te begrijpen (Kroll et al., 2016). Ook zegt een toegankelijke broncode weinig over hoe een algoritme in de praktijk uitpakt. Daarom is onze focus op algoritmisering zo relevant: het gaat niet alleen om transparantie van het algoritme zelf, maar ook over hoe deze in de organisatie wordt ingebed en gebruikt.

Wanneer we de dimensies van algoritmisering hanteren, dan krijgen de twee dimensies van transparantie als volgt vorm (tabel 2).

We verwachten dat transparante algoritmisering kan bijdragen aan vertrouwen. Theorie over procedurele rechtvaardigheid biedt inzichten in hoe algoritmische transparantie vertrouwen kan beïnvloeden. Volgens een van de grondleggers van deze theorie, Tom Tyler, kunnen individuen vrede hebben met een beslissing die negatief voor hen uitpakt als de beslisprocedure als rechtvaardig wordt ervaren (Tyler, 2006). Een van de aspecten van een rechtvaardige procedure is het krijgen van een gegronde uitleg over het besluit en de neutraliteit van de beslisser. In lijn met deze verwachting zorgt toegankelijke en interpreteerbare algoritmisering voor een eerlijker proces: dit draagt vervolgens bij aan vertrouwen in de beslissing

Tabel 2 Richtinggevende vragen voor transparante algoritmisering

Dimensie van algoritmisering	Toegankelijkheid	Interpreteerbaarheid
Technologie	Is de broncode toegankelijk om ontwerpkeuzen te beoordelen?	Geeft het algoritme inhoudelijke redenen voor een gegenereerde beslissing of aanbeveling?
Technische expertise	Zijn de functiekenmerken van de experts die betrokken zijn bij het algoritme toegankelijk?	Worden er redenen gegeven voor de betrokkenheid van verschillende soorten experts?
Informatie-relaties	Zijn de gebruikte datasets en metadata-kenmerken toegankelijk?	Geeft de organisatie uitleg over welke datasets worden gebruikt, waarom en op welke wijze?
Organisatiestructuur	Is het overzicht van de organisationele verantwoordelijkheden ten aanzien van het algoritme toegankelijk?	Worden de keuzes achter de organisationele verantwoordelijkheden uitgelegd?
Informatiebeleid	Is het informatiebeleid rond het gebruik van algoritmen toegankelijk?	Wordt het informatiebeleid rond het gebruik van algoritmen uitgelegd?
Monitoring en evaluatie	Zijn zowel de resultaten van M&E als de aanpak hiervan toegankelijk?	Wordt helder uitgelegd op welke wijze en aan de hand van welke criteria het algoritme wordt gemonitord en geëvalueerd?

en in degene die de beslissing neemt (Grootelaar & Van den Bos, 2018; Porumbescu & Grimmelikhuijsen, 2018).

Wanneer we kijken naar de dimensies van algoritmisering in tabel 2, dan kan de toegankelijkheid van datasets, algoritmen en informatiebeleid laten zien dat algoritmen niet bevooroordeeld zijn en dus bijdragen aan een rechtvaardige beslisprocedure. Tegelijkertijd sluit ook de uitlegbaarheid van algoritmisering aan bij een van de componenten van een rechtvaardige procedure. Al met al verwachten we dat het versterken van algoritmische transparantie, dus zowel toegankelijkheid als interpreteerbaarheid, bijdraagt aan de gepercipieerde welwillendheid en integriteit van overheidsinstanties. Door middel van transparante monitoring en evaluatie kunnen overheidsorganisaties laten zien dat algoritmen de gewenste effecten bereiken en dat ongewenste neveneffecten worden voorkomen. Dit kan bijdragen aan gepercipieerde competentie.

Samenvattend hebben we algoritmische transparantie gedefinieerd en toegepast op het idee van algoritmisering. We hebben vervolgens twaalf richtinggevende vragen opgesteld op basis waarvan overheidsorganisaties kunnen gaan nadenken over transparante algoritmisering (zie tabel 1 en 2). Tot slot hebben we beargumenteerd dat transparantie in termen van toegankelijkheid en interpreteerbaarheid een tweede noodzakelijk fundament is onder verantwoorde algoritmisering.

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004002

Tabel 3 Mogelijke relaties tussen verantwoorde algoritmisering en vertrouwen in de overheid

Dimensie van betrouwbaarheid	Waardengevoelige algoritmisering	Transparante algoritmisering
Gepercipieerde competentie	Waardengevoelige algoritmisering kan competentie versterken wanneer de organisatie laat zien dat algoritme zo wordt ingezet dat het bijdraagt aan afgewogen set aan waarden.	Transparante monitoring en evaluatie van algoritmen maken de effectiviteit van algoritmen zichtbaar. Dit kan gepercipieerde competentie vergroten.
Gepercipieerde welwillendheid	Waardengevoelige algoritmisering zorgt ervoor dat belangrijke waarden voor burgers (bijv. rechtvaardigheid) niet over het hoofd worden gezien. Dit kan bijdragen aan gepercipieerde welwillendheid.	Transparante algoritmisering zorgt ervoor dat de waarden die voor burgers belangrijk zijn, zichtbaar zijn. Dit kan bijdragen aan gepercipieerde welwillendheid.
Gepercipieerde integriteit	Waardengevoelige algoritmisering zorgt ervoor dat besluitvormers en ontwikkelaars meer oog hebben voor data-ethiek. Dit kan bijdragen aan gepercipieerde integriteit.	Transparante algoritmisering zorgt ervoor dat externe partijen toegang hebben tot algoritmen, hetgeen zorgt voor meer openheid. Dit kan bijdragen aan gepercipieerde integriteit.

Een onderzoeksagenda voor verantwoorde algoritmisering en vertrouwen

In dit artikel hebben we laten zien dat de opkomst van *machine learning*-algoritmen in het openbaar bestuur niet alleen vragen van technische aard oproept, maar juist ook vragen van organisationele aard. Dit hebben we algoritmisering genoemd. Tot dusverre wordt in de wetenschappelijke literatuur vooral de nadruk gelegd op de potentiële technische voordelen van algoritmen (efficiëntie, accuratesse) en de ethische risico's ervan (bevooroordeeld, risico's voor privacy). Wij hebben betoogd dat we verder moeten kijken dan deze eenvoudige tweedeling en dat de voor- en nadelen van algoritmen uiteindelijk worden bepaald door de manier waarop deze in een overheidsorganisatie worden ingebed en hoe de algoritmen in de praktijk worden gebruikt. We moeten in de bestuurswetenschap dus kijken naar algoritmisering als een sociaal proces in plaats van naar een algoritme als een technologisch artefact.

Ten tweede hebben wij beargumenteerd dat transparantie en waardengevoeligheid van algoritmisering kunnen bijdragen aan verschillende dimensies van vertrouwen in de overheid. Deze argumenten zijn samengevat in tabel 3.

Tabel 3 laat zien dat waardengevoeligheid en transparantie belangrijke pilaren vormen van verantwoorde algoritmisering. De mogelijke relaties tussen verantwoorde algoritmisering en vertrouwen in tabel 3 zijn echter empirisch nog niet of

nauwelijks getoetst. Om de waardengevoeligheid van algoritmisering te onderzoeken kan worden gedacht aan diepgaand kwalitatief onderzoek. Door middel van interviews en etnografisch onderzoek kunnen bijvoorbeeld impliciete waarden van betrokkenen bij het algoritmiseringsproces worden blootgelegd. Ook kan worden gedacht aan kwantitatief onderzoek om effecten van algoritmiseringsprocessen op het vertrouwen van burgers te toetsen. Zo kunnen diverse scenario's met al dan niet transparante algoritmen worden voorgelegd in een reeks surveyexperimenten om op die manier te kijken welke aspecten van transparantie bijdragen aan het vertrouwen van burgers.

Naast het toetsen van de voorgestelde relaties in tabel 3 zijn er twee specifieke gecompliceerde thema's waar we in toekomstig onderzoek aandacht aan zouden moeten besteden. Ten eerste zien we dat er bij algoritmen al snel wordt gedacht aan het implementeren van één systeem in één organisatie en dat deze implementatie in goede banen geleid moet worden. Echter, in de praktijk zien we dat data en algoritmen steeds meer onderdeel worden van netwerken van data en netwerken van organisaties (Janssen & Van den Hoven, 2015). Dit betekent dat de verantwoordelijkheid over transparante en waardengevoelige algoritmisering bij verschillende partijen ligt. Wie verantwoordelijk is voor welk deel van het algoritmiseringsproces, is dan niet eenduidig vast te stellen (Cicirelli, Guerrieri, Mastroianni, Spezzano & Vinci (2019; Zouridis et al., 2020). Meer onderzoek is nodig om inzicht te geven in hoe algoritmisering op verantwoorde wijze vorm kan krijgen in dergelijke netwerken.

Een tweede thema dat meer aandacht nodig heeft, is het zelflerende karakter van *machine learning*-algoritmen. Mogelijk is het proces van algoritmisering op een waardengevoelige en transparante manier ontwikkeld en gestart, maar is dit veranderd na verschillende iteraties van het algoritme. Een algoritme kan zodanig ontwikkeld zijn dat het geen onderscheid maakt op bijvoorbeeld etniciteit, maar na verloop van tijd kan het algoritme via een omweg toch 'leren' dat 'etniciteit' als variabele een rol speelt in het beslismodel. De dynamiek van algoritmen vereist dus vooral een goede monitoring en evaluatie door de tijd heen, en deze dimensie van algoritmisering moet daarom goed worden geborgd.

Kortom, het is belangrijk om zich te realiseren dat algoritmen niet (alleen) gaan om de techniek en dat een puur technische benadering van rechtvaardige algoritmen niet vanzelf tot verantwoord gebruik van algoritmen leidt. Hiervoor is een organisatiebrede aanpak nodig waarbij ontwikkelaars, politici en beleidsmakers zich bewust zijn van de waarden die ten grondslag liggen aan de techniek en hier ook transparant over durven zijn.

Literatuur

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93-117.

Bestuurskunde 2020 (29) 4
doi: 10.5553/Bk/092733872020029004002

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1-12. doi:10.1177/2053951715622512

Centraal Bureau voor Statistiek (CBS). (2018). *Verkennend onderzoek naar het gebruik van algoritmen binnen overheidsorganisaties* [Exploratory investigation of the use of algorithms in government organizations]. Verkregen van https://www.cbs.nl/nl-nl/maatwerk/2018/48/gebruik-van-algoritmen-door-overheidsorganisaties

Cicirelli, F., Guerrieri, A., Mastroianni, C., Spezzano, G., & Vinci, A. (Eds.). (2019). *The Internet of Things for smart urban ecosystems*. Cham: Springer.

Coombs, W.T. (2007). Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corporate Reputation Review, 10*(3), 163-176.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56-62. doi:10.1145/2844110

Friedman, B. (1996). Value-sensitive design. *Interactions, 3*(6), 16-23.

Friedman, B., Kahn, P.H., & Borning, A. (2008). Value sensitive design and information systems. In Himma, K.R. & H.T. Tavani (eds.) *The handbook of information and computer ethics*. (pp. 69-101). Hoboken (NJ), United States: John Wiley & Sons.

Gil-Garcia, J.R. (2012). *Enacting electronic government success: An integrative study of government-wide websites, organizational capabilities, and institutions* (Vol. 31). Springer Science & Business Media.

Grimmelikhuijsen, S., & Knies, E. (2017). Validating a scale for citizen trust in government organizations. *International Review of Administrative Sciences, 83*(3), 583-601. doi: 10.1177/0020852315585950

Grimmelikhuijsen, S., & Porumbescu, G.A. (2017). Reconsidering the expectancy disconfirmation model: Three experimental replications. *Public Management Review, 19*(9), 1272-1292.

Grimmelikhuijsen, S., Vries, F. de, & Zijlstra, W. (2018). Breaking bad news without breaking trust: The effects of a press release and newspaper coverage on perceived trustworthiness. *Journal of Behavioral Public Administration, 1*(1).

Grootelaar, H.A., & Bos, K. van den. (2018). How litigants in Dutch courtrooms come to trust judges: The role of perceived procedural justice, outcome favorability, and other sociolegal moderators. *Law & Society Review, 52*(1), 234-268. doi:10.1111/lasr.12315

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR), 51*(5), 93.

Hood, C. (2007). What happens when transparency meets blame-avoidance? *Public Management Review, 9*(2), 191-210.

Hoven, J. van den. (2007). ICT and value sensitive design. In Goujon,P., Lavelle, S., Duquenoy, P., Kimppa, K. & V. Laurent (eds.) *The information society: Innovation, legitimacy, ethics and democracy in honor of Professor Jacques Berleur SJ* (pp. 67-72). Boston, MA: Springer.

Janssen, M., & Hoven, J. van den. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? *Government Information Quarterly, 32*(4), 363-368.

Kroll, J.A., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., & Yu., H. (2016). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633-706.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611-627. doi:10.1007/s13347-017-0279-x

Mayer, R., James, H.D., & Schoorman, D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734. doi:10.5465/amr.1995.9508080335

McEvily, B., & Tortoriello. M. (2011). Measuring trust in organisational research: Review and recommendations. *Journal of Trust Research, 1*(1), 23-63. doi: 10.1080/21515581.2011.552424

Meijer, A., & Wessels. M. (2019). Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration, 42*(12), 1-9. doi: 10.1080/01900692.2019.1575664

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288).

Orlikowski, W.J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science*, *3*(3), 398-427.

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy*, *39*(6), 751-760. doi:10.1093/scipol/scs093

Porumbescu, G.A., & Grimmelikhuijsen, S. (2018). Linking decision-making procedures to decision acceptance and citizen voice: Evidence from two studies. *The American Review of Public Administration, 48*(8), 902-914. doi:10.1177/0275074017734642

Rousseau, D.M., Sitkin, S.B., Burt, R.S. & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393- 404. doi: 10.5465/amr.1998.926617

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210-229.

Schendel, S. van. (2019). The challenges of risk profiling used by law enforcement: Examining the cases of COMPAS and SyRI. In Reins, L. (ed.) *Regulating new technologies in uncertain times* (pp. 225-240). The Hague: TMC Asser Press.

Tyler, T.R. (2006). *Why people obey the law*. Princeton: Princeton University Press.

Wijnen, J.-F. van. (2019, 27 mei). COA test algoritme voor gerichte plaatsing vluchtelingen. FD.nl

Willems, D., & Doeleman, R. (2014). Predictive Policing – wens of werkelijkheid?, *Tijdschrift voor de Politie, 76*(4-5), 39-42.

Zouridis, S., Eck, M. van, & Bovens. M. (2020). Automated discretion. In Evans, T. & P. Huper (eds.) *Discretion and the quest for controlled freedom.*(pp. 313-329) Cham: Palgrave Macmillan.

Zuurmond, A. (1994). *De infocratie: Een theoretische en empirische heroriëntatie op Weber's ideaaltype in het informatietijdperk*. Den Haag: Phaedrus.