Algoritme: de mens in de machine

Casusonderzoek naar de toepasbaarheid van richtlijnen voor algoritmen

Job Spierings

Sander van der Waal

Maart 2020

waag
technology & society

'The difference, and it is a crucial one, lies in the wholly new scale of ambition and intervention entertained by high modernism.'

James C. Scott, ‘Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed’ (1998)

Casusonderzoek toepasbaarheid van conceptrichtlijnen voor algoritmen

Job Spierings
Sander van der Waal

Samenvatting

In opdracht van het ministerie van Binnenlandse Zaken en Koninkrijksrelaties heeft Waag een onderzoek gedaan naar 'richtlijnen voor het gebruik van algoritmen' op basis van twee casussen bij de politie en het UWV.

Onderzoek

Politie (keuzehulp bij formulier online aangifte internetoplichting) en UWV (Werkverkenner) stelden uitgebreide documentatie beschikbaar over het ontwerp van beide applicaties. Naast bureauonderzoek hielden we gestructureerde interviews met acht teams die in verschillende rollen betrokken zijn bij ontwerp en gebruik van deze applicaties. Tot slot organiseerden we een werksessie waarbij ook een aantal andere overheidsorganisaties vertegenwoordigd waren.

Aanbevelingen

Aanbeveling 1-3 gaan inhoudelijk over de richtlijnen zelf. Aanbeveling 4 en 5 zijn maatregelen die het inbedden en toepassen van de richtlijnen bij overheden kunnen faciliteren.

1. Verscherp de definitie

Er bestaan aanzienlijke verschillen in wat onder een algoritme wordt verstaan. Analytisch en technisch onderlegde gesprekspartners zien ze overal: een bureaucratisch proces is voor hen ook een algoritme. Andere geïnterviewden lijken de term algoritme gelijk te stellen aan "puur de techniek" of kunstmatige intelligentie. Dit maakt de reikwijdte van de richtlijnen ondoorzichtig.

De risico's die de richtlijnen adresseren doen zich juist voor op het snijvlak van algoritme en ambtenaar. Om te voorkomen dat de richtlijnen te beperkt of juist te breed worden opgevat adviseren we het onderwerp van de richtlijnen aan te passen en daarbij gebruik te maken van het onderzoek van Maranke Wieringa. Zij definieert een algoritmisch systeem als een socio-technische verzameling bestaande uit een combinatie van technische onderdelen, sociale praktijken en (organisatie)cultuur.[1] De overheid gebruikt deze socio-technische verzamelingen om beslissingen te nemen voor of over burgers. De richtlijnen kunnen dan ook niet goed worden toegepast zonder ook (het grotere) vraagstuk rondom ketenverantwoordelijkheid en geautomatiseerde besluiten te adresseren en expliciet te maken hoe deze richtlijnen zich hiermee verhouden.[2]

[1] Maranke Wieringa, 'What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability', in: Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA.

[2] Zie: Marlies van Eck, 'Geautomatiseerde ketenbesluiten & rechtsbescherming': https://pure.uvt.nl/ws/portalfiles/portal/20399771/Van_Eck_Geautomatiseerde_ketenbesluiten.pdf

2. Benoem rollen

Definieer voor wie de richtlijnen zijn en maak ze voor deze rollen specifiek en compact. Het toepassen van de richtlijnen zal voor een belangrijk deel bestaan uit het organiseren van interne dialoog en afstemming over het treffen van de juiste maatregelen bij ontwikkeling en gebruik van algoritmen. Wij adviseren de hierbij betrokken rollen expliciet te benoemen en stellen zes rollen voor die bij toepassing altijd betrokken zouden moeten: proceseigenaar, ontwerper/onderzoeker, business/regel-analist, softwareontwikkelaar, communicatiemedewerker en beslismedewerker.

De mate waarin deze rollen betrokken en verantwoordelijk zijn verschilt: in ontwerp- en testfase zijn kleine hoeveelheden beslismedewerkers betrokken en houdt architect/onderzoeker dagdagelijks overzicht. Als het algoritme is geïmplementeerd en onderdeel is van lijnactiviteiten, is die verhouding omgekeerd. Afhankelijk van de fase (ontwerp, pilot, test, implementatie) hebben rollen een andere betrokkenheid met en verantwoordelijkheid voor de toepassing van de richtlijnen. Uitgangspunt is dat voor alle rollen in iedere fase de betrokkenheid geborgd moet zijn.

3. Vereenvoudig de opzet

Op basis van ons onderzoek denken we dat de richtlijnen zijn in te delen in vier categorieën:

1. Wet en beleid: Expliciet moet worden aangetoond dat het toegepaste algoritme en het omliggende proces geschikt en proportioneel zijn voor het doel zoals vastgelegd in wet, beleid en regels. Deze afweging moet publiek beschikbaar zijn. Dit omvat ook de richtlijnen voor validatie en gegevensherkenning. In de toepassing van de richtlijnen kan het nuttig zijn te verwijzen naar richtlijnen en standaarden voor specifieke domeinen of vakgebieden (Algemene beginselen van behoorlijk bestuur, de verderop genoemde Standard for Public Code[3], BurgerServiceCode[4]).

2. Impact assessment: de context waarin het algoritme wordt toegepast is bepalend voor de manier waarop je de richtlijnen wil toepassen. Maak daarbij het volgende onderscheid:
    A. Impact van het algoritme op de beslissing/proces
    B. Impact van de beslissing of het proces op de burger
    In procesontwerpen worden (on)voorziene risico's vaak geadresseerd door menselijke monitoring/interventie. Zowel de kwaliteit als de kwantiteit hiervan moet geborgd worden.

3. Verantwoording: wie is verantwoordelijk voor de toepassing van de richtlijnen? Hoe krijgt die verantwoordelijkheid vorm, wie is daar intern bij betrokken? Hoe wordt geborgd dat verantwoordelijkheid betekenisvol gedragen kan worden met voldoende domeinkennis, technologisch inzicht en overzicht over het proces? Hoe is dit auditeerbaar en toetsbaar?

[3] https://standard.publiccode.net/

[4] https://www.noraonline.nl/wiki/BurgerServiceCode

4. Uitlegbaarheid: maak onderscheid voor wie de uitlegbaarheid beschikbaar moet zijn, op welk moment en welke maatregelen daarvoor zijn genomen.

    A. Publieke uitlegbaarheid: leg publiek en transparant uit waarom dit algoritme wordt ingezet, op welke wetgeving en beleid dat is gebaseerd en waar algoritme en systeem voor worden geoptimaliseerd.
    B. Collegiale uitlegbaarheid, borging, bewaking en verantwoording: leg uit wat er in het operationele proces nodig is om de juiste werking van het algoritme te borgen. Waar moeten bijvoorbeeld (beslis)ambtenaren rekening mee houden bij het monitoren of begrijpen van uitkomsten van algoritmen?
    C. Uitlegbaarheid van een specifieke beslissing: spontaan aangeboden uitleg in begrijpelijke taal over de totstandkoming van een beslissing, gericht op een gelijkwaardige informatiepositie voor de burger. Gekoppeld aan laagdrempelige manieren om in contact te komen.

4. Koppel toepassing richtlijnen aan dialoog, voorlichting en verantwoording

De richtlijnen zijn bedoeld als kader/instrument voor het (kunnen) toepassen van geschikte maatregelen. Daarnaast werken de richtlijnen op dit moment vooral als kader voor een gestructureerde dialoog maar zijn nog geen zelfstandig bruikbaar instrument. Daarvoor moet worden voorzien in tooling om ketenverantwoordelijkheid in kaart te brengen, interne dialoog te organiseren en resultaten daarvan vast te leggen.

Dit komt ook tegemoet aan de behoefte van mensen om te weten wanneer men het goed doet. Hierbij kan het organisaties helpen als er een *impact assessment* beschikbaar is.

De resultaten van deze processen kunnen vervolgens worden ingezet bij voorlichting en verantwoording.

5. Maak de richtlijnen zelf onderdeel van toetsing en doorontwikkeling

Over de impact en ontwikkeling van het grootschalig en ingrijpend inzetten van geavanceerde technologie bestaat veel onzekerheid, gevoed door angsten en hypes. Juist systemische en onvoorziene risico's gaan (pas) spelen als algoritmen op schaal worden ingezet en worden pas met de tijd zichtbaar. In de brief van minister Dekker wordt dan ook aangegeven dat het gaat om 'materie die relatief nieuw is'.[5]

Uit oogpunt van kwaliteit en borging in maatschappelijke democratie is het belangrijk om de doorontwikkeling van de richtlijnen zélf en de toepassing daarvan onderdeel te maken van een publieke dialoog. Bij dit proces horen niet alleen relevante overheidsorganisaties betrokken te worden maar ook wetenschap, professionals, maatschappelijke organisaties en burgers.

[5] https://www.rijksoverheid.nl/documenten/kamerstukken/2019/10/08/tk-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid

Inhoud

Inleiding

In november 2015 stelde Google het programma TensorFlow[6] open source beschikbaar. Plotseling was wereldwijd *cutting edge* technologie beschikbaar voor *machine learning* en het opzetten van neurale netwerken. Waar kunstmatige intelligentie tot dan toe het exclusieve domein was van universitaire en industriële onderzoeksgroepen, was deze opeens voor iedereen toegankelijk.

De zich spectaculair ontwikkelende beschikbaarheid van technologieën zorgt ervoor dat klassieke grondrechten onder druk staan (Vetzo, 2018). Experts geven aan dat impact daarvan onbekend, maar mogelijk zeer groot is. Een beroemd voorbeeld hiervan is een uitspraak van Stephen Hawking in 2016, die aangaf dat volgens hem AI "het beste of het slechtste wordt dat de mensheid ooit zal overkomen"[7]. Het maatschappelijk, op industriële schaal toepassen van nieuwe technologieën stelt onze democratie daarmee op meerdere niveaus op de proef. Voorbeelden hiervan staan bijna dagelijks in de krant, variërend van de invloed van platforms als AirBnB en Uber tot aan de vermeende invloed van algoritmes van grote platformen op meningsvorming bij democratische verkiezingen. De technische complexiteit, onvoldoende dan wel onvoldragen praktijkvoorbeelden en een veelal ronkend publiek debat plaatsen de overheid hiermee in een lastig parket.

In deze context heeft Waag tussen september 2019 en januari 2020 in opdracht van het ministerie van Binnenlandse Zaken en Koninkrijksrelaties (BZK) onderzoek gedaan naar richtlijnen voor het gebruik van algoritmen, welke door het ministerie van Justitie en Veiligheid en BZK zijn ontwikkeld. Hierbij keek Waag naar de toepassing van nieuwe technologie in overheidscontext en hoe daarbij wordt omgegaan met risico's en richtlijnen, in het bijzonder bij een casus van de politie en UWV. Dit rapport is het resultaat van dit onderzoek.

Dit rapport bestaat uit drie onderdelen: een beschrijving van de context en aanpak van het onderzoek, een overzicht van de bevindingen die daaruit naar voren kwamen en een vijftal aanbevelingen.

Waag wil graag de geïnterviewde teamleden van de politie en UWV bedanken voor de ruime tijd die zij beschikbaar stelden en de openheid en transparantie die zij gaven in hun werk.

[6] https://en.wikipedia.org/wiki/TensorFlow

[7] https://www.theguardian.com/science/2016/oct/19/stephen-hawking-ai-best-or-worst-thing-for-humanity-cambridge

Onderzoek

Onderzoeksvragen

In reactie op zorgen die er op dit gebied leven hebben twee interdepartementale werkgroepen een document opgesteld: "Conceptrichtlijnen voor het toepassen van algoritmes door overheden". In de begeleidende brief 'wettelijke waarborgen'[8] waarmee deze richtlijnen naar de Tweede Kamer zijn gestuurd is een toetsing van de richtlijnen door het Transparantielab van BZK aangekondigd. Met deze toetsing wordt nader invulling gegeven aan de motie Verhoeven / Van der Molen[9] en de motie Middendorp (onderdeel praktisch toepasbare richtlijnen).[10]

In opdracht van het ministerie van Binnenlandse Zaken en Koninkrijksrelaties heeft Waag als onderdeel van de toetsing een onderzoek gedaan naar deze richtlijnen, waarin we op basis van twee casussen antwoord zochten op de volgende vragen:

- Wat is de reden dat een richtlijn wel of niet toegepast of toepasbaar is?
- Wat zijn concrete suggesties om specifieke richtlijnen waar nodig te verbeteren ten behoeve van de praktische bruikbaarheid?

- Zijn de richtlijnen toepasbaar in de context van de doelstelling waar binnen het algoritme is ingezet? Waarom wel of niet?
- In hoeverre helpen de richtlijnen om risico's, zoals aansprakelijkheid en bias, te beheersen?
- In hoeverre zijn de richtlijnen bij deze casus toegepast?

De eerste twee vragen worden voor iedere richtlijn apart beantwoord. De drie laatste onderzoeksvragen zijn generiek en worden samen in één paragraaf beantwoord.

Opzet 'Richtlijnen voor algoritmes'

Het document 'Richtlijnen voor het toepassen van algoritmes door overheden' bestaat uit:

- **Inleiding** die de politiek bestuurlijke context beschrijft en definieert dat de richtlijnen gaan over het ontwikkelen en gebruiken van algoritmes door de overheid en voorlichting over data-analyses door overheden aan het publiek.
- **Typologie** van algoritmen, waarbij een omgekeerd verband wordt beschreven tussen oplopende complexiteit en afnemende toepasbaarheid van technische transparantie:

[8] https://www.rijksoverheid.nl/documenten/kamerstukken/2019/10/08/tk-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid

[9] https://zoek.officielebekendmakingen.nl/kst-26643-610.html

[10] https://zoek.officielebekendmakingen.nl/kst-35200-VII-14.html

- eenvoudige beslisboom
- eenvoudig rule based
- lineaire regressie
- logistische regressie
- deep learning

- Inzet van algoritmen in 4 categorieën, die een oplopend niveau van impact hebben:
  - 1. Beschrijvend – Analyse van "Wat gebeurt er?"
  - 2. Diagnostisch – Analyse van "Waarom gebeurt het?"
  - 3. Voorspellend – Analyse van "Wat zal er gebeuren?"
  - 4. Voorschrijvend – Analyse van "Wat moet er gebeuren?"
- Definities / onderscheid tussen **technische transparantie** en **uitlegbaarheid**
- **Richtlijnen** (7)
  - Bewustzijn risico's
  - Uitlegbaarheid
  - Gegevensherkenning
  - Auditeerbaarheid
  - Verantwoording
  - Validatie
  - Toetsbaarheid
- **Categorie-indeling**
  - Categorie 1: keuzes en aannames uit wetgeving, beleidsregels af te leiden.
  - Categorie 2: keuzes, gegevens en aannames kenbaar maken.
  - Categorie 3: transparant maken van algoritmes en maatwerkgegevens
  - Categorie 4: focus op uitlegbaarheid i.p.v. technische transparantie.
  - Categorie 5: uitlegbaarheid/transparantie niet wenselijk
- **Bijlage 1: Nadere toelichting** waarin de 7 richtlijnen worden uitgelegd en sommige ervan voorzien van praktische aanbeveling voor toepassing ervan.
- **Bijlage 2: Richtlijnen inzake publieksvoorlichting over data-analyses** die zich vooral richt op een specifieke categorie algoritmen: data-analyses op basis van persoonlijke gegevens.

Onderzoeksopzet

Het onderzoek richt zich op de praktische bruikbaarheid en toepasselijkheid van een set richtlijnen voor het gebruik van algoritmen in twee specifieke casussen bij de politie en het UWV. Centraal staat of, en in hoeverre, de richtlijnen op de werkvloer bruikbaar zijn en tot zinvolle inzichten en uitkomsten leiden. Daarnaast komen onderdelen naar voren waarvan betrokkenen het van belang achten dat deze ook of explicieter door de richtlijnen worden benoemd. Het functioneren van de applicaties, het daadwerkelijke gebruik en het omringende proces en beleid is niet onderzocht.

Beperking van het onderzoek is dat in beide onderzochte casussen een academische context aanwezig is die elementen uit de richtlijnen al heeft toegepast omdat dit in de wetenschappelijke wereld gebruikelijk is. Toetsing vanuit casuïstiek die deze context en routine niet heeft, zal andere resultaten en inzichten opleveren.

Tijdens het bureauonderzoek hebben we een aantal gesprekken gevoerd met experts en hebben we de opzet van de interviews voorbereid. Hierbij gingen we als volgt te werk:

- Vooraf ontvingen we van Politie en UWV documentatie over de systemen en het onderzoek dat daaraan ten grondslag ligt.
- We hebben gekeken naar structuur, aanpak en inhoud van andere Richtlijnen en Toetsingskaders voor het gebruik van AI en het borgen van publieke waarden in technische systemen, waaronder applicaties/tools die inzichtelijk proberen te maken hoe eerlijkheid en statistiek samen kunnen komen in AI. Een samenvatting hiervan hebben we online gepubliceerd.[11]
- Daarnaast betrokken we meer algemeen onderzoek naar de impact van computer- en ketenbesluiten, en het verband met doenvermogen, begrip en empathie.

Omdat de richtlijnen bedoeld zijn voor een brede groep gebruikers, variërend van projectleiders tot businessanalisten, ontwikkelaars, architecten, en communicatiemedewerkers, hebben we de organisaties benaderd met het verzoek gesprekken te mogen voeren met (vertegenwoordigers van) verschillende teams en rollen. Uitgangspunt waren de volgende vier rollen die we vooraf hebben gedefinieerd:

Rol	Uitleg	Accent bij interview
Projectleider of applicatiebeheer-der	Persoon die verantwoordelijk is voor de ontwikkeling / onderhoud van het algoritme.	Focus op de mate waarin de richtlijnen nuttig zijn bij de ontwikkeling en onderhoud van het algoritme.
Afdelingshoofd	Persoon die verantwoording draagt voor de beslissingen die met het algoritme worden genomen.	Focus op de mate waarin menselijke verantwoordelijkheid voor de beslissingen die met het algoritme worden genomen kan worden gedragen.
Analist of software ontwikkelaar	Iemand die de innerlijke werking van het algoritme goed begrijpt, inclusief de domeinkennis die hierbij van belang is.	Specifieke focus op de mate van technische transparantie en hoe deze vertaald wordt naar een niet-technisch publiek.
Communicatie-specialist	Iemand die verantwoordelijk is voor de communicatie met de burger aangaande de afdeling of processen waarbij het algoritme wordt gebruikt.	Focus op de richtlijnen t.a.v. de publieksvoorlichting over data-analyses.

In beide casussen bleek deze indeling te eenvoudig. Belangrijk is de centrale rol voor een onderzoeker/ architect, die een overkoepelende, regisserende rol en verantwoordelijkheid draagt. In beide casussen hebben we met de persoon in deze rol kunnen spreken. In gestructureerde interviews spraken we met vertegenwoordigers van in totaal 8 teams die in verschillende rollen en verantwoordelijkheden betrokken zijn bij de applicaties in de casussen.

In dit verslag worden citaten uit gesprekken en interviews vermeld ter illustratie van de onderzoeksinzichten. Om focus op richtlijnen te houden is precieze context en achtergrond van het citaat minder van belang, daarom is er voor gekozen dat nergens te vermelden. Citaten zijn uit oogpunt van leesbaarheid geredigeerd.

[11] Zie: Petra Bíró, 'Algorithm says no: ethische richtlijnen voor AI systemen', (2019), https://waag.org/nl/article/algorithm-says-no-ethische-richtlijnen-voor-ai-systemen

Werksessie

Tot slot organiseerden we een werksessie over uitlegbaarheid en verantwoordelijkheid met zowel casusvertegenwoordigers als deelnemers van uitvoeringsorganisaties en andere onderzoekers. Met behulp van een prototype canvas werden ‘burgerreizen’ (een variant op klantreizen waarbij de burger niet als klant maar als mens centraal staat) visueel in kaart gebracht en in verband gebracht met de actieve en/of verantwoordelijke teams, de door deze teams beheerde algoritmen en de impact daarvan op genomen besluiten en de burger.

Onderzochte casussen

Politie: Keuzehulp aangifte online oplichting

Online oplichting wordt wel gezien als 'het nieuwe fietsendiefstal': relatief makkelijk om te doen maar moeilijk om te bewijzen. Het standaardformulier dat de politie hiervoor op haar website heeft werkte niet naar tevredenheid. Een deel van de aangiftes werd te summier ingevuld en er zijn aangiftes waar er vermoedelijk sprake is van een conflict met een verkoper over kwaliteit of levertijd. In dat geval is de aangever meer geholpen met verwijzing naar civielrechtelijke stappen omdat er juridisch gezien geen sprake is van oplichting.

De webapplicatie[12] biedt een adviserende functie aan burgers bij het doen van online aangifte van internetoplichting (Artikel 326 Wetboek van Strafrecht). Deze ondersteuning wordt aan de burger als optie aangeboden, het is ook mogelijk rechtstreeks naar het reguliere aangifteformulier te gaan. De burger vertelt eerst, in een vrij tekstveld, wat er is gebeurd. Met behulp van tekstanalyse stelt de applicatie vervolgens aanvullende vragen. Op basis van de tekst en aanvullende antwoorden classificeert het model de aangifte en geeft de aangever advies en handelingsperspectief over wat de meest zinvolle vervolgactie is. In hoofdlijnen zijn er twee opties:

- De burger *voelt* zich opgelicht maar er is juridisch waarschijnlijk sprake van bijvoorbeeld wanprestatie of een ander civielrechtelijk conflict. Het advies aan de burger kan bijvoorbeeld zijn om de levering alsnog af te wachten of contact te zoeken met een brancheorganisatie als Thuiswinkel.org voor bemiddeling.
- De applicatie schat in dat online aangifte zinvol is en helpt burger in stappen het aangifteformulier in te vullen.

De applicatie is ontworpen en ontwikkeld door het AI-lab van de Nationale Politie in Driebergen. De inhoudelijke expertise en strategie op dit domein ligt, in afstemming met de portefeuillehouder digitalisering en cybercrime, bij het Landelijk Meldpunt Internetoplichting in Heemskerk, een initiatief van de eenheid Noord-Holland. Het operationele proces waarin de aangiften worden beoordeeld en verwerkt valt onder verantwoordelijkheid van de afdeling Operationele Informatieverwerking (OIV) in Driebergen.

In het kader van dit onderzoek hebben we met vertegenwoordigers van deze drie teams gesproken. Betrokkenen die we niet gesproken hebben zijn onder meer de (deels externe) ontwikkelteams, portefeuillehouder dienstverlening (verantwoordelijk voor de website, belegd bij eenheid Oost-Nederland), Officier van Justitie/OM en lokale korpsen die uiteindelijk verantwoordelijk zijn voor opsporing en vervolging.

[12] https://www.politie.nl/themas/internetoplichting.html

UWV: Werkverkenner

Het UWV werkbedrijf is verantwoordelijk voor het ondersteunen van WW'ers bij het vinden van betaald werk. De ondersteuning bestaat uit dienstverlening zoals een workshop, webinar, online trainingen en begint vaak met een persoonlijk gesprek op een UWV-vestiging. Omdat zowel het aantal adviseurs als de beschikbare dienstverlening beperkt is moet het werkbedrijf bepalen hoe snel WW'ers in aanmerking komen voor een gesprek en welke dienstverlening dan het best ingezet kan worden.

Hiervoor heeft het UWV de Werkverkenner ontwikkeld.[13] Nadat een WW-uitkering is toegekend krijgt de WW'er in de online 'Werkmap' een vragenlijst te zien. De WW'er geeft antwoord op vragen over arbeidsverleden, de persoonlijke situatie en een inschatting van de eigen kansen op de arbeidsmarkt. Gecombineerd met gegevens over arbeidsverleden en opleiding voorspelt de Werkverkenner vervolgens de kans op het vinden van betaald werk binnen een jaar (uitgedrukt in een percentage), aangevuld met een diagnose van persoonlijke belemmeringen en mogelijkheden van WW'ers om betaald werk te vinden (een percentage per factor). De aanvrager krijgt de percentages niet te zien, tenzij de WW'er er naar vraagt of dit anderszins expliciet aan de orde komt in het persoonlijke gesprek.

Met het ministerie van SZW heeft het UWV afgesproken dat met alle WW'ers met een score van 50% of lager een gesprek plaatsvindt *binnen drie maanden* na de eerste dag dat de WW-uitkering is vastgesteld/ toegekend. In dat gesprek kan de adviseur dienstverlening aanbieden, waarbij het aanbod is afgestemd op de berekende belemmerende factoren.

Voor WW'ers met een hogere kans op werkhervatting dan 50% wordt een klantbeeld gemaakt. Een adviseur kan op basis van deze informatie alsnog besluiten de WW'er zo snel mogelijk te spreken. WW'ers met een hogere score dan 50%, die ook niet op basis van klantbeeld worden uitgenodigd, worden allemaal uiterlijk na zes maanden uitgenodigd voor een persoonlijk gesprek.

De WW'er zelf kan op elk moment in het proces aangeven dat deze een gesprek wil.

De applicatie is gebaseerd op onderzoek van en ontworpen door het kenniscentrum van het UWV in samenwerking met TNO, NOA en VUmc.[14] Het kenniscentrum is onderdeel van de directie strategie en beleid. Het UWV Werkbedrijf Noord-Holland Noord is eindverantwoordelijk voor de werking van de Werkverkenner. Werkbedrijf Dienstverlening levert onder andere de projectleider en business analist. De afdeling Werkbedrijf Businessadvies & Communicatie draagt zorg voor de algemene en interne communicatie en verzorgt de training van medewerkers/adviseurs in het gebruik van de Werkverkenner op de regiokantoren. Adviseurs die de Werkverkenner dagdagelijks gebruiken bevinden zich op de verschillende regiokantoren.

In het kader van dit onderzoek hebben we met vertegenwoordigers van deze afdelingen gesproken over de ontwikkeling en implementatie van de nieuwe versie (2.0). Betrokkenen die we niet gesproken hebben zijn onder meer de afdeling ICT/architect en de (deels externe) ontwikkelteams.

---

[13] https://www.werk.nl/werkzoekenden/uitkering-aanvragen/uwv-dienstverlening/ww/#paragraaf

[14] Dit betreft Werkverkenner 2.0. De eerste versie is ontwikkeld met het Universitair Medisch Centrum Groningen van de Universiteit Groningen.

Bevindingen van ons onderzoek

In dit hoofdstuk beantwoorden we voor zowel het inleidende deel van de conceptrichtlijnen als van elk van de richtlijnen op zich de twee onderstaande onderzoeksvragen.

- Wat is de reden dat een richtlijn wel of niet toegepast of toepasbaar is?
- Wat zijn concrete suggesties om specifieke richtlijnen waar nodig te verbeteren ten behoeve van de praktische bruikbaarheid?

Richtlijnen: Inleiding

Een belangrijk deel van de richtlijnen is gewijd aan het categoriseren van typen, gebruik en inzet van algoritmen. Hoewel de zeven richtlijnen zelf niet verwijzen naar deze indelingen, lijkt de bedoeling van de opstellers te zijn handvatten te geven voor de toepassing van de richtlijnen, als een eerste stap richting een *impact assessment*.

We hebben op die manier met gesprekspartners het document doorlopen: kan men de eigen casus en algoritme plaatsen in de voorgestelde ordening en geeft dit richting voor het toepassen van de 7 richtlijnen?

De **Typologie** van algoritmen beschrijft een omgekeerd verband tussen oplopende technische complexiteit en afnemende betekenis van technische transparantie van "eenvoudige beslisboom" tot aan "deep learning". De geldigheid van deze indeling wordt in de gesprekken vaak betwist: "een ambtelijk proces gebaseerd op spreadsheets kan enorm complex zijn." Wanneer binnen één socio-technisch systeem de keus moet worden gemaakt tussen twee verschillende algoritmen (en alle andere omstandigheden gelijk blijven, dus er is ook geen merkbaar verschil in impact/resultaat), zou het in het algemeen zo kunnen zijn dat technische transparantie bij deep learning minder betekenis heeft dan bij de andere genoemde technologieën. Omdat in processen vaak meerdere systemen met meerdere algoritmes worden ingezet, bleek de analytische waarde van deze typologie beperkt.

Verderop (onder 'Uitlegbaarheid') staat: "Dit brengt als uitgangspunt mee dat overheidsorganisaties in beginsel geen algoritmes mogen hanteren die te complex zijn om te kunnen worden uitgelegd." Dit uitgangspunt gekoppeld aan de typologie zou vertaald kunnen worden naar een subsidiariteitsbeginsel: gebruik altijd zo eenvoudig mogelijke technologie. Bij de UWV Werkverkenner is daar ook bewust voor gekozen: een complexere modellering met random forest[15] methodiek bleek in de onderzoeksfase een iets betere voorspelling van kans op werkhervatting te bieden. Toch werd door de onderzoekers geadviseerd een eenvoudiger statistische methodiek te gebruiken met een (marginaal) verlies in voorspellende kracht tot gevolg. De extra complexiteit brengt het risico met zich mee dat de statistische uitkomsten in de uitvoering minder makkelijk begrepen kunnen worden en dat de uitlegbaarheid afneemt.

[15] https://en.wikipedia.org/wiki/Random_forest

Bij de ordening van de **Inzetgebieden** van algoritmen in 4 categorieën gaan gesprekspartners door de generieke toonzetting van de tekst algoritmen uit verschillende processen met elkaar vergelijken door ze in de vier categorieën in te delen. De opzet is echter om gebruikers te laten reflecteren op de keuze die zij hebben in de inzet van het eigen algoritme. Op het moment dat dit in het gesprek naar voren komt geven betrokkenen aan dat de impact van het algoritme afhankelijk is van meerdere factoren, waarvan deze er één is.

Iets verderop in de tekst worden de inzetgebieden nog verbonden met al dan niet inherente menselijke tussenkomst, ook weergegeven in een figuur. De algemene geldigheid en bruikbaarheid hiervan voor de toepassing van richtlijnen is in het onderzoek niet helder geworden.

Het onderscheid tussen **technische transparantie** en **uitlegbaarheid** wordt als belangrijk genoemd. Waarbij alle geïnterviewden aangeven dat in maatschappelijk opzicht een overheidsproces vrijwel altijd complex is. De vraag is daarom: uitlegbaarheid aan wie en op welk moment? En betekent uitlegbaar dat de overheid is uitgesproken of dat de burger de uitleg heeft begrepen?

In beide casussen geven de gesproken architecten aan dat zij juist hierom een actieve band hebben met de wetenschappelijke wereld: zij hebben hun onderliggend statistisch en technisch onderzoek open gepubliceerd en verhouden zich tot actueel academisch onderzoek. Zij willen technisch transparant zijn om zowel intern als extern uitlegbaarheid te faciliteren op meerdere niveaus. Vanuit de ervaringen van het onderzoek is daaraan toe te voegen dat uitlegbaarheid altijd ook transparantie van het proces en transparantie van de organisatie vergt.

De tekst over inzet van algoritmen sluit af met een **categorie-indeling**. Deze bleek in de onderzochte casussen geen verheldering op te leveren. Het is helder dat er bij de richtlijnen behoefte is aan een inschatting die richting geeft aan niveau en gelaagdheid van de toepassing van de richtlijnen. Een algoritme in een enkele verkeersregelinstallatie heeft immers een andere systemische en individuele impact als een landelijk dekkend, zelflerend systeem dat de surveillance-inzet van de politie aanstuurt. Onder het kopje 'impact' bij 'Bewustzijn risico's' zetten wij hiervoor een aantal suggesties op een rij.

De richtlijnen

In onderstaande paragrafen behandelen we de zeven richtlijnen stuk voor stuk, waarbij steeds de richtlijn en de bijbehorende toelichting in één keer wordt behandeld, gevolgd door puntsgewijze aanbevelingen.

Richtlijn 1. Bewustzijn risico's

> *"Juist de alledaagse bureaucratie leidt tot Kafkaëske situaties. Juist die bureaucratie ondersteunen we met AI. Als het daar misgaat hebben mensen juist daar moeite om te begrijpen waar het misgaat en hoe dat te beheersen."*

> *"Het voornaamste risico voor mij is 'vervreemding'. Ambtenaar en burger weten niet meer precies wat er gebeurt én het gehele proces is geautomatiseerd zodat er ook geen alternatieve routes meer zijn."*

Alle geïnterviewden waren zich zeer bewust van het feit dat er bijzondere risico's kleven aan de inzet van automatisering in het algemeen en algoritmes in het bijzonder. Vooral het risico dat systemen onbedoeld vooroordelen bevatten of een discriminerende werking kunnen hebben werd vaak genoemd. Men vindt het goed dat er met de richtlijnen een structuur is om deze risico's te adresseren en waardeert het dat deze richtlijnen specifiek op het overheidsdomein zijn gericht, waar andere (zoals die van de Europese Commissie) zich meer richten op commerciële toepassingen.

De term risico wordt bij de verschillende teams verschillend begrepen. Hoe 'eerder' in de keten, hoe meer systemisch het risico wordt ingeschat. De ontwerper van het algoritme dat burgers assisteert bij het invullen van de online aangifte, noemt het risico dat alleen nog "domme" criminelen worden vervolgd, omdat die aangiftes de meeste aanknopingspunten voor opsporing bieden. Voor het selecteren van zaken is dus menselijke interventie nodig, die bijvoorbeeld ook patronen kan zien tussen aangiften die op zichzelf weinig lijken voor te stellen.

Aan de meer uitvoerende kant van de keten is het risicobewustzijn in algemene zin aanwezig maar hebben mensen moeite om deze zich concreet voor te stellen of ontbreekt een kader waarin de impact van algoritme op proces en uitvoer begrepen wordt. Zo werd gezegd: "*Het algoritme is getoetst op bias en dat zit er niet in*". Het helpt dan om met extreme(re) scenario's vragen op scherp te stellen, of een vergelijking te maken met processen die technisch veel minder complex zijn.

In het geval er bij individuele beslissingen iets mis zou gaan, verwijst men naar reguliere processen waar burgers zich kunnen melden. In beide applicaties kunnen gebruikers de (voorgestelde) beslissing negeren. Bijvoorbeeld door alsnog aangifte te doen, of het UWV actief te vragen om een persoonlijk gesprek. Meermaals werd in gesprekken benadrukt hoe belangrijk het is dat burgers de impact van algoritmen rechtstreeks kunnen temperen. Als het vaststellen van ongewenste, systemische effecten van algoritme/omliggende proces ter sprake wordt gebracht, gaat dat pas leven als onderzoekers een voorbeeld geven en vragen hoe de geïnterviewden en hun organisatie dan zou reageren. Dit leidt steeds tot een geanimeerd gesprek over risico's die zijn verbonden aan ketenverantwoordelijkheid en veel vragen over waar de verantwoordelijkheden (zouden moeten) liggen. Bij het kopje '5. Verantwoording' gaan we hier verder op in.

Bij zowel UWV als Politie is het algoritme ontworpen door een afdeling die zich actief verhoudt tot en publiceert in academische context. Vanuit die werkhouding adresseren zij de specifieke aandachtspunten en voelen zij zich ook verantwoordelijk voor de operationele bewaking hiervan. Andere betrokken teams en divisies verwijzen hier ook naar. Dit is belangrijk omdat in deze teams de kennis en middelen beschikbaar zijn om op systemisch risico te monitoren en bij individuele gevallen (ondersteunende) tekst en uitleg te geven.

In de casus van de politie wordt een belangrijk deel van de risicobeheersing 'aan de voorkant' georganiseerd: zo is online aangifte doen van ernstige zaken onmogelijk.

Impact

Specifiek hebben betrokkenen in de casus behoefte om als onderdeel van de toepassing van richtlijnen de impact van het algoritme te bepalen. Er dient daarbij onderscheid gemaakt te worden tussen de

impact van de beslissing (op de burger of maatschappij als geheel) en *de impact van het algoritme óp de beslissing*.[16]

Wanneer *de impact van het algoritme* op de beslissing hoog is, is er op dat moment geen of beperkte menselijke tussenkomst. Dat stelt hogere eisen aan procesbewaking en het bieden van mogelijkheden om processen stil te kunnen leggen of terug te draaien. Wanneer er in het proces altijd sprake is van menselijke tussenkomst, bijvoorbeeld omdat *de impact van de beslissing* hoog is, moet deze menselijke tussenkomst georganiseerd én gekwalificeerd worden. Hoe de richtlijnen worden toegepast is dus mede afhankelijk van een *impact assessment*.

Een deelnemer aan de werksessie zei het als volgt:

> *"Bij ons zit veel intelligentie in de systemen, zodat het grootste deel van de uitvoering gedaan kan worden door een brede groep aan medewerkers. Als die systemen output geven die statistisch genuanceerd moet worden, is het niet vanzelfsprekend dat de medewerkers geëquipeerd zijn de uitkomsten ook zo te begrijpen en daarop beslissingen te nemen."*

> *"In de opzet van een model voor opsporing van Bijstandsfraude is afgesproken dat er voor terugvordering een minimumbedrag geldt in een geval een 'hit' te laten zijn. Bij dat soort bedragen gaat het meestal om kleine administratieve fouten, zou je daar het model op trainen dan vervuild dat teveel. Maar in de uitvoer merken we dat handhavers die gevallen toch handmatig als 'hit' willen toevoegen. Het is dan moeilijk om op conceptueel niveau uit te leggen dat dit niet wenselijk is. Je moet het dan eerst fout laten gaan, waarna je kunt laten zien dat het model minder precies is geworden."*

Deze zorgt wordt versterkt door bevindingen die vanuit onderzoek naar maatwerk bij de overheid naar voren komen:

> *"Opvallend is ook dat veel van de onderzochte organisaties het maatwerk op rechten en plichten in de praktijk in veel gevallen invullen via bezwaar en beroep. Daarmee vindt maatwerk meer aan de achterkant van de dienstverlening en de klantreis plaats. Maatwerk in de vorm van persoonlijk contact staat op veel plaatsen vanuit kostenoverwegingen onder druk. Dat leidt tot tekortkomingen in de dienstverlening aan bepaalde groepen klanten, tot lange klantreizen en maatwerk dat uiteindelijk alleen geboden wordt aan burgers met veel bureaucratische vaardigheden en uithoudingsvermogen."*[17]

Uitkomst is beslissend voor impact

Of een beslissing veel impact heeft, kan afhankelijk zijn van wat de beslissing is. Als voorbeeld uit de casus van het UWV: De Werkverkenner bepaalt (in veel gevallen) geautomatiseerd wanneer een WW'er

---

[16] Maranke Wieringa, 'Approaching Algorithmic Accountability' (in voorbereiding).

[17] Regels en Ruimte Verkenning Maatwerk in dienstverlening en discretionaire ruimte: https://www.rijksoverheid.nl/documenten/rapporten/2020/01/16/bijlage-rapport-abdtopconsult-maatwerk-dienstverlening

een persoonlijk gesprek krijgt aangeboden. Hoewel het inhoudelijk niet helemaal correct is om te spreken over het 'onterecht aanbieden' of 'onthouden' van een gesprek, zou je hier statistisch wel van kunnen spreken. Het gebruikte model kan er in de voorspelling immers naast zitten, waardoor er iets anders gebeurt dan het UWV eigenlijk wenst:

- Als de voorspelling van het algoritme ervoor zorgt dat een gesprek *"onterecht" aangeboden* wordt, is de negatieve impact beperkt tot de energie en tijd die van beide kanten aan het gesprek wordt besteed.
- Als de voorspelling van het algoritme ervoor zorgt dat een gesprek *onterecht onthouden* wordt, is de mogelijke negatieve impact veel groter: zowel onderzoekers als adviseurs geven aan dat de het ondersteunen van werkzoekenden in de eerste maanden vele malen effectiever is. Voor dit geval zijn daarom twee vangnetten: de WW'er kan zelf om een gesprek vragen en het UWV doet een check op klantbeeld.

Categorisering van impact

Tot slot kwamen we bij ons bureauonderzoek een indeling tegen van systemen voor kunstmatige intelligentie die behulpzaam kan zijn bij het categoriseren en typeren van impact en risico's. Onderzoeker Arvind Narayanan noemt drie categorieën:[18]

- **Systemen die perceptie automatiseren:** zoals gezichtsherkenning, medische analyse van scans, fotoherkenning etc.
  - Maatschappelijke risico's ontstaan *omdat* techniek goed werkt: publieke ruimte kan geautomatiseerd onder 100% surveillance worden geplaatst, mensen kunnen kennis opdoen over onbehandelbare kwalen die ze liever niet gehad zouden hebben. Tamelijk onzichtbaar kan een informatiepositie en handelingsperspectief worden opgebouwd, dat zich onttrekt aan toezicht en democratische controle.
- **Systemen voor het automatiseren van beslissingen:** spam- en uploadfilters, automatische aanbevelingen.
  - Maatschappelijke risico's ontstaan doordat systemen *meestal* redelijk goed werken maar onvermijdelijk fouten maken. In dit soort processen bestaat er niet altijd een universeel geaccepteerde beslissing, of de acceptatie daarvan over tijd verandert terwijl een algoritme als tijdscapsule blijft functioneren.
- **Systemen die sociologische voorspellingen doen:** zoals *predictive policing*, risico op recidive, matchen van cv's bij vacatures.
  - Maatschappelijk risico ontstaat omdat het voorspellen van de toekomst onmogelijk is, terwijl een systeem (of de verkopers daarvan) dit voor de hand liggende, fundamentele inzicht verbergt en er toch onomkeerbare beslissingen worden genomen.

[18] https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf (acc. 28 nov. 2019)

Aanbevelingen bij richtlijn 1

- Risico: na implementatie en inrichting komen ontwerpers, architecten op afstand te staan van het proces waardoor risico's niet op tijd worden vastgesteld en juist geadresseerd. Benoem dit risico en adresseer het met een evaluatiecyclus/feedbackloop.
- Maak onderscheid tussen de impact van de beslissing en de impact van het algoritme op de beslissing.
- Ondersteun gebruikers van de richtlijnen door het bieden van een *impact assessment*.

Richtlijn 2. Uitlegbaarheid

Ook uitlegbaarheid wordt in de gesprekken spontaan en als fundamenteel benoemd. Geïnterviewden vinden dat zij een bijzondere verantwoordelijkheid hebben naar burgers/klanten en willen graag goed voorbereid zijn op (kritische) vragen die zij verwachten uit de maatschappij. Zij vulden in de interviews de richtlijn aan met een aantal maatregelen die zij in hun eigen casus nodig achten. Ondermeer wordt een onderscheid gemaakt aan wie en op elke moment uitleg gegeven moet worden.

Publieke uitlegbaarheid

De tekst richt zich expliciet op het uitleggen van beslissingen voor individuele burgers maar zegt niks over de systemische impact van de inzet van algoritmen. Zoals Frissen et al. (2019) schrijven:

> *"In de uitvoering waar alleen mensen beslissingen nemen, hebben overheidsmedewerkers binnen het wettelijk voorgeschreven kader de ruimte om beslissingen te nemen en hun eigen oordeel te vormen. (Evans en Hupe 2019, 4). Dit wordt de discretie genoemd. Als de overheid algoritmen inzet om de taken uit te voeren wordt deze ruimte (de beslissingspraktijk) ingevuld door algoritmen."* [19]

In beide casussen wordt door ontwerpers en betrokkenen bewust op systemische impact gereflecteerd en hierover is voor de buitenwereld ook documentatie en beleid beschikbaar.

Collegiale uitlegbaarheid

Transparantie over het gebruik van algoritmes vereist dat een organisatie intern zicht en regie heeft op het volledige socio-technische systeem: kunnen collega's, teams, afdelingen aan elkaar uitleggen wat zij doen en van elkaar verwachten? Een goede inrichting hiervan is een randvoorwaarde (en maakt het vele malen eenvoudiger) om het gebruik van algoritmen aan burgers uit te leggen en wanneer nodig extern te laten toetsen.

In onze interviews blijkt dat gesprekspartners regelmatig op eigen initiatief op zoek moeten naar de verwachtingen, voortgang en resultaten bij andere teams en daarbij niet kunnen terugvallen op een gedeeld proces. De interpretatie van statistische uitkomsten en foutmarges bij ontwerpende teams wordt heel precies geformuleerd. 'Verderop in de keten' worden in gesprekken dezelfde getallen

[19] https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2019Z26117&did=2019D53638

gebruikt, maar zonder statistische nuance en zonder dat altijd duidelijk is of men onderscheid maakt tussen effectiviteit van het proces zelf of het onderliggende onderzoek.

Collegiale *zichtbaarheid* speelt hier ook een rol: hoewel bijna iedere overheidslaag tegenwoordig een ICT-organisatie is, merken we in interviews dat afdelingen die verantwoordelijk zijn voor ICT zó weinig tast- en zichtbaar zijn, dat er over de werking van applicaties door medewerkers soms in bijna mystieke termen wordt gesproken. Dat de eigen organisatie een actieve, ontwerpende rol heeft in ICT-ontwerp en onderhoud, waar medewerkers iets over kunnen vinden en waar veel denk- en menskracht naar toe gaat is geen actieve kennis. Ook daarom is een voorwaarde voor het toepassen van transparantie van algoritmen dat ontwerpbeslissingen en code intern "leesbaar" worden gemaakt.

Specifieke maatregelen

Vaststellen of de specifieke maatregelen die hier in de richtlijnen worden genoemd zijn toegepast valt buiten de scope van het onderzoek (code georganiseerd in modules, functionele testen, etc.). Gesprekspartners gingen er vanuit dat door de betreffende afdelingen de daarvoor gebruikelijke standaarden voor (overheids)automatisering worden gebruikt. Er is in deze casussen niet in voorzien dat teams volledig toegang/inzicht hebben in elkaars documentatie, beslissingen en code. Wanneer beslissingen over features, specificaties, ontwerp, bouw en tests verdeeld zijn over meerdere teams kunnen er in de overdracht ongemerkt en onbedoeld interpretatieverschillen ontstaan. Transparantie en uitlegbaarheid komen dan in gevaar.

Aanbevelingen bij richtlijn 2

- Voeg onderscheid toe naar drie niveaus van uitlegbaarheid:
  - Publieke uitlegbaarheid: leg uit waarom specifiek dit algoritme wordt ingezet, op welke wetgeving en beleid dat is gebaseerd en waar algoritme en systeem voor worden geoptimaliseerd.
  - Collegiale uitlegbaarheid en verantwoording: leg uit welke borging in het operationele proces nodig is om de juiste werking van het algoritme te borgen. Waar moeten bijvoorbeeld (beslis)ambtenaren rekening mee houden bij het monitoren of begrijpen van uitkomsten van algoritmen?
  - Uitlegbaarheid van een specifieke beslissing: uitleg in begrijpelijke taal over hoe een algoritme tot een beslissing is gekomen. Biedt deze uitleg spontaan aan en verwijs daarbij ook naar de documentatie behorende bij de 'hogere' niveaus van uitlegbaarheid, zodat de burger al dan niet met behulp van ingeschakelde experts een gelijkwaardige informatiepositie heeft. Koppel dit aan laagdrempelige manieren om in contact te komen: in bezwaar gaan is een grote stap, even langskomen (of bellen) al een kleinere. Naast terugredeneren (waarom is deze beslissing nu voor mij genomen?) moet ook zoveel mogelijk vooruit redeneren worden gefaciliteerd ('als ik dit in mijn leven aanpas, heb ik eventueel wel een recht').

- De specifieke maatregelen m.b.t. evaluatie, testen, meten zijn wellicht beter op hun plek onder kopjes verantwoording/toetsing. Het is te overwegen daar expliciet te verwijzen naar standaarden en praktijken die in de betreffende domeinen/industrie in gebruik zijn.

- Het goed documenteren en publiceren van code is een voorwaarde voor transparantie, hergebruik, doorontwikkeling en onderhoud, inclusief het geïntegreerd documenteren van softwarecode, beslisregels, beleid en wetgeving. Een goed voorbeeld hiervan is de Standard for Public Code (een Nederlands initiatief) die hiervoor een standaard heeft ontwikkeld. Naast functionaliteit zijn in de documentatie ook onderlinge rollen, verantwoordelijkheden en risico's inzichtelijk voor zowel bestuurders, beleidsmakers als ontwikkelaars. Onderzoek of dit model bruikbaar kan zijn bij het toepassen van de richtlijnen.
- Naar het voor burgers bruikbaar, begrijpelijk en geautomatiseerd motiveren van computerbesluiten wordt praktisch onderzoek gedaan door Matthijs van Kempen bij de Belastingdienst. Dit biedt een model om gelaagd te traceren hoe een specifiek besluit bestaat uit verschillende beslissingen, gebaseerd op beslisregels en rechtsbronnen. Dan wordt onder meer vermeden dat een burger simpelweg wordt verwezen naar 'de wet'.

Richtlijn 3. Gegevensherkenning

De in de casussen gebruikte trainingsgegevens zijn niet open, maar wel beschikbaar voor bijvoorbeeld interne of rechterlijke toetsing. Bij het UWV gaat het om persoonsgegevens en bij de politie gaat het daarnaast ook om operationele en vertrouwelijk informatie m.b.t aangiften. De parametrisering is gedocumenteerd in publiek toegankelijke publicaties.

In de werksessie werd nog aanvullend onderscheid belangrijk gevonden: gegevens kunnen afkomstig zijn van andere processen of organisaties en het resultaat van het algoritme kan zelf ook weer in andere processen of door andere organisaties worden hergebruikt. Voor het documenteren van gegevensherkenning en -gebruik moet dit onderscheid gemaakt kunnen worden.

Aanbevelingen bij richtlijn 3

- Maak bij het uitleggen en transparant maken onderscheid tussen gegevens die worden gebruikt:
  - Om een model te trainen of te ontwikkelen
  - Die als onderdeel van het proces worden verzameld/ingevoerd
  - Die worden opgevraagd uit andere processen of anderen bronnen
- Documenteer welke gegevens het algoritme oplevert, wie daar toegang tot heeft, in welke processen/organisaties deze uitkomsten worden gebruikt, wie daar dan voor verantwoordelijk is en welke impact dat heeft.

Richtlijn 4. Auditeerbaarheid

De richtlijn tracht onderscheid te maken tussen algemene en specifieke consequenties van algoritmen voor burgers en gaat vooral in op auditeerbaarheid in het laatste geval. Uit de gevoerde gesprekken komt naar voren dat in *beide* gevallen auditeerbaarheid van belang is. De onderverdeling die bij uitlegbaarheid werd genoemd (in drie niveaus) is hier wellicht ook zinvol.

Zowel vanuit oogpunt van collegiale verantwoording/ketenverantwoordelijkheid als het faciliteren van maatschappelijke democratie dienen algoritmes, code en documentatie open beschikbaar te zijn ("open tenzij"). Deze openheid is echter alleen zinvol als er ook sprake is van transparantie van het proces en transparantie van de organisatie.

Aanbevelingen bij richtlijn 4

- Auditeerbaarheid moet ook mogelijk zijn voor het evalueren van beleid/systemische impact van het algoritme.
- Afhankelijk van de vastgestelde impact (zie onder: 'Bewustzijn risico's') is wetenschappelijke validatie een vereiste (zie ook de volgende paragraaf).
- De organisatie en kwalificatie van menselijke tussenkomst en borging daarvan in het operationele proces zou ook onderdeel moeten zijn van audit-processen.

Richtlijn 5. Verantwoording

> *"Is verantwoording alleen op papier geregeld of ook in het echt? Je kunt iemand wel verantwoordelijk maken, maar die moet het ook kunnen dragen."*

De richtlijnen zijn summier over verantwoording terwijl uit onze gesprekken naar voren komt dat betrokkenen dit punt belangrijk vinden en het voor hen in bepaalde gevallen onvoldoende helder is, of niet is vastgelegd, wie (de eerst) verantwoordelijk(e) is.

Academische inbedding

In de onderzochte casussen zijn betrokkenen van mening dat zowel de impact van het ondersteunde proces als het innovatieve karakter van de applicatie vereisen dat over algoritme en het omliggende proces academisch verantwoording wordt afgelegd. Gesprekspartners geven aan dat dit zowel ten goede komt aan de uitlegbaarheid van een beslissing in een individueel geval als de systemische impact van het algoritme.

> *"Bij de politie is er een AI-Lab met promovendi, die zijn de helft van de tijd met hun promotie bezig en de andere helft met politiewerk. Het heeft als doel de politie een zelfredzame organisatie te maken op het gebied van A.I., maar is ook opgezet vanuit het oogpunt van transparantie. Het team ondersteunt promovendi bij het valoriseren van onderzoek en onderzoekt ook zelf ethische aspecten. Daarover is*

> *vooral van hogerhand een discussie gaande: het moet helder zijn hoe de politie dit soort technologie inzet. Als men niet snapt wat we aan het doen zijn krijg je anders een chilling effect. Het moet helder zijn waar we aan werken."*

> *"De politie is er niet om zoveel mogelijk mensen te arresteren of zo hè? We zijn er echt voor de veiligheid, en daar hoort ook het gevoel van veiligheid bij. En dat vereist dat je hele apparaat veilig overkomt, ook richting de burger."*

Ketenverantwoordelijkheid

In beide onderzochte casussen is het ontwerp van algoritmen en applicatie gebaseerd op recente wetenschappelijke inzichten en technische innovaties. De processen die daarmee worden ondersteund zijn, voor een complexe overheid als de Nederlandse, op zichzelf relatief eenvoudig. Ondanks de relatief geïsoleerde processen en relatieve eenvoud is er hier sprake van ketenverantwoordelijkheid, waarbij een diversiteit aan betrokken teams, vallend onder verschillende divisies of regio's, verantwoordelijk zijn voor afzonderlijke delen van het proces. Over het verschil tussen een proces dat met klassieke middelen wordt ondersteund ('Tanken zonder te Betalen') en een proces dat voorzien is van kunstmatige intelligentie (Keuzehulp):

> *"Het proces voor Tanken zonder te Betalen wil ik nog digitaliseren. Daar worden nu deels formulieren letterlijk met de hand overgetikt."*
> *Wie is nu voor dat proces verantwoordelijk?*
> *"Dat ben ik, daar is dat redelijk makkelijk. Bij de keuzehulp zit je toch met veel externen en complexe techniek."*

De applicaties in beide casussen zijn onderdeel van grote en gelaagde websites. Omdat digitale technologie van zichzelf al ondoorzichtig is, is het inzicht binnen een organisatie tussen teams over verantwoordelijkheden soms beperkt. Bijvoorbeeld welk team verantwoordelijk is voor wat, waar en wanneer voor de bezoeker zichtbaar is.

Kortom: de impact van een algoritme wordt bepaald door een keten van (ontwerp)beslissingen in meerdere teams, die in tijd en plaats van elkaar gescheiden zijn. Het risico is dat teams ervan uitgaan dat verantwoordelijkheid bij een ander team of afdeling ligt, of simpelweg naar een bovenliggend managementniveau verwijzen. Uiteindelijk komt de verantwoording zo hoog in een organisatie te liggen dat de operationele betekenis ontbreekt, of zelfs tussen organisatie in ligt en niemand meer aanspreekbaar is.

De richtlijnen kunnen daarom niet goed worden toegepast zonder ook (het grotere) vraagstuk rondom ketenverantwoordelijkheid en geautomatiseerde besluiten te adresseren.[20] Het onderzoek dat Maike Klip bij DUO doet naar het visualiseren en inzichtelijk maken van ambtelijk begrip en ketenverantwoordelijkheid in het leveren van digitale diensten is hier volgens ons toonaangevend.[21]

---

[20] Zie: Marlies van Eck, 'Geautomatiseerde ketenbesluiten & rechtsbescherming': https://pure.uvt.nl/ws/portalfiles/portal/20399771/Van_Eck_Geautomatiseerde_ketenbesluiten.pdf

[21] Voor een introductie zie deze presentatie: https://noti.st/maikeklip/sv6yQM/slides of dit blog, waarin een service blue print wordt gebruikt om de relatie tot de burger van verschillende teams inzichtelijk te maken: https://klipklaar.nl/begripvolle-ambtenaren/hessel-als-begripvolle-ambtenaar/.

Rollen

In de interviews spraken wij medewerkers die vanuit verschillende rollen een eigen betrokkenheid hebben bij de betreffende algoritmen. Iedereen die wij hebben gesproken geeft aan dat de richtlijnen voor hen relevant zijn, mede vanuit de behoefte een bevestiging te krijgen dat men 'het goed doet'.

De richtlijnen zouden voor de verschillende rollen specifieker en compact kunnen worden gemaakt. Met de aparte richtlijnen voor publieksvoorlichting is hiervoor een eerste aanzet gedaan.

Uit de gesprekken blijkt dat betrokkenheid van de architect/ontwerper van het algoritme een voorwaarde is voor het zinvol toepassen van de richtlijnen. Vooral als een applicatie is ontwikkeld als onderdeel van een innovatietraject, is er een risico dat betrokkenheid van teams of ontwerpers/architecten niet duurzaam is ingericht. Bij ontwerp, bouw en implementatie hebben zij rekening gehouden met de operationele context waarin het algoritme wordt gebruikt. In de interviews tonen zij groot eigenaarschap en geven zij aan zich verantwoordelijk te voelen voor zowel het algoritme als de juiste afstemming met het proces waarvan het algoritme onderdeel is. Deze verantwoordelijkheid geven zij ook actief vorm. Zo zijn er regelmatig demo-momenten waarop vertegenwoordigers van teams bij elkaar komen om nieuwe issues en feature te bespreken. De betrokkenheid van de architecten betekent ook dat andere teams reden hebben om de werking van het algoritme actief te volgen en met vragen of inzichten op regelmatige basis terecht kunnen.

Ontwerpers/architecten hebben formeel echter beperkt zicht op de daadwerkelijke implementatie (bijvoorbeeld door machine-controleerbare validatie). Of en hoe zij nog betrokken zijn als de applicatie geen pilot meer is maar een reguliere lijnactiviteit is voor hen niet altijd duidelijk. Als dan de operationele context verandert of het algoritme (technisch ongewijzigd) wordt toegepast in een heel ander proces, is niet noodzakelijkerwijs helder wie de verantwoordelijkheid heeft om de richtlijnen opnieuw toe te passen en of daarbij de benodigde capaciteit beschikbaar is.

> *"Als onze applicatie succesvol is wil men deze ook in elders inzetten, een totaal ander domein."*
> *"Ben jij daar dan als architect nog bij betrokken?"*
> *"Dat weet ik niet."*

> *"(...) Dus dat heeft te maken met je eigen normen en waarden maar ook met de verandering van de technieken en andere mogelijkheden. En die dialoog, dat is denk ik ook wat ze hier bedoelen, dat is belangrijk inderdaad dat je dat blijft doen. Nou in zoverre denk ik dat dat ook wel in die sprint reviews aan de orde komt. Dat mensen toch op gegeven moment zeggen van ja eigenlijk doen we het al maanden nu op deze manier maar misschien is het wel helemaal niet de goede manier. En is dat moment voor reflectie en feedback wel ingebouwd in het onderhoudstraject?"*

> *"Het valt me op dat de overgang naar agile werken hier veel in heeft veranderd. Helemaal als het gaat om diensten of besluiten die door meerdere teams worden gemaakt. Ieder team focust zich op z'n eigen stukje, maar de dienst en het besluit is de hele legpuzzel. Soms wordt er dan wel een klassieke projectleider/ manager opgezet maar die heeft eigenlijk geen macht meer omdat product owners met hun teams zelf mogen prioriteren. Ja, iedereen is dan verantwoordelijk en dus niemand."*

Aanbevelingen bij richtlijn 5

- Definieer voor wie de richtlijnen zijn en maak ze voor rollen specifiek en compact.
- Ken in de richtlijnen een expliciete rol, betrokkenheid en verantwoordelijkheid toe aan de architect(en) van systemen en algoritmen.
- Benoem en beleg bij complexe processen met grotere impact expliciet de regie over het gehele proces, om ook verantwoordelijkheid te dragen voor transparantie van het proces en de transparantie van de organisatie.
- Bied gebruikers van de richtlijnen handvatten voor het in kaart brengen en inzichtelijk maken van (keten)verantwoordelijkheden.
- Onderzoek of het zinvol is in de richtlijnen aandacht te besteden aan de verschillende fasen van volwassenheid van een systeem/algoritme (bijvoorbeeld ontwerp/pilot/ implementatie of bestaand algoritme in nieuwe context).
- Richtlijnen 4, 6 en 7 (auditeerbaarheid, validatie en toetsbaarheid) zijn te beschouwen als specifieke operationalisering van verantwoordelijkheden.

Richtlijn 6. Validatie

> *"Waar wij heel vaak mee te maken hebben is niet zozeer een nul-meting maar dat onze normen en waarden gewoon veranderen. Dat we gewoon een andere mening krijgen over wat een kwetsbare groep is, terwijl we dat gisteren misschien niet vonden. Zijn we nog met zijn allen het juiste aan het doen? Dat soort discussies zijn altijd super moeilijk want meestal klinkt het veroordelend richting optreden van mensen in het verleden, terwijl gewoon de maatschappij verandert."*

Zoals deze richtlijn nu geformuleerd is bood deze in de interviews geen nieuwe inzichten, omdat de genoemde punten eerder in de richtlijnen al aan bod waren gekomen.

We denken dat het expliciet noemen van validatie zinvol is, zodat de wetenschappelijke en statistische validiteit van het algoritme is geborgd. Ook validatie van de juistheid van beslisregels tegen wet- en beleid door juristen (of materiedeskundigen) valt hieronder.

Men moet zich er bewust van zijn dat de impact van een (technisch) rigide systeem kan fluctueren in de tijd omdat die impact ook afhankelijk is van maatschappelijke veranderingen. Hetzelfde geldt voor

dynamische, 'zelflerende' systemen. Validatie vindt dus plaats door middel van voortdurende procesbewaking.

Zo geeft de Werkverkenner van het UWV elke WW'er een score (percentage) dat weergeeft wat de voorspelde kans is dat de aanvrager binnen een jaar weer betaald werk heeft. Welke beslissing er vervolgens wordt genomen op basis van dit percentage is vooral afhankelijk van andere factoren: afspraken met het ministerie van SZW (die kunnen veranderen), het aantal beschikbare adviseurs en de economische conjunctuur: zijn er meer WW'ers, dan wordt de werkdruk hoger en kan het UWV relatief minder dienstverlening bieden.

Aanbevelingen bij richtlijn 6

- De kwalificatie van menselijke tussenkomst in het proces is onderdeel van de validatie. Het kan dan gaan om begrip van statistiek, impact, uitval. In de brief van staatssecretaris Dekker worden onder het kopje 'vii. Kwaliteitseisen aan het gebruik van statistiek' een opzet genoemd die in de richtlijnen kan terugkeren.
- Validatie vereist ook een expliciete koppeling met (en documentatie van) relevante wet- en regelgeving, beleid en inrichting van het proces. En routinematige toetsing van de validiteit van deze verbanden.

Richtlijn 7. Toetsbaarheid

> *"Het onderliggende systeem bestaat uit 46 regels en 26 observable facts. Kan je die laten zien?"*
> *"Dat kan ik jullie niet laten zien want dat wordt gezien als operationele informatie. Maar in een rechtszaak zou dat kunnen, ook in overleg met de officier van justitie."*

Tijdens de interviews was de praktische betekenis van het onderscheid tussen toetsing en auditeerbaarheid niet altijd helder.

Aanbeveling bij richtlijn 7

- Maak voor de doelgroep van de conceptrichtlijnen een helder onderscheid tussen een (gerechtelijke) toetsing en een (externe) audit en geef praktische handvatten voor dit onderscheid.

Richtlijnen: bijlage publieksvoorlichting data-analyses

Deze bijlage bij de richtlijnen richt zich vooral op een specifieke categorie algoritmen: data-analyses op basis van persoonlijke gegevens, zoals applicaties die burgers profileren of sociologische voorspellingen doen. De overheid kan dat doen ten bate van opsporing of, zoals bij de UWV Werkverkenner, om beperkte middelen voor ondersteuning gerichter in te zetten.

Dit deel van de richtlijnen verhoudt zich ook expliciet tot de Algemene verordening gegevensbescherming (AVG). In ons onderzoek, vooral tijdens de (meerdere uren durende) interviews, ontbrak de tijd om daar wezenlijk aandacht aan te besteden. We hebben dit deel van de richtlijnen aan bod laten komen voor zover passend bij de onderzoeksopdracht.

Transparantie

De algemene inzichten die hier in de richtlijnen staan worden door de geïnterviewden onderschreven. Verder zijn de belangrijkste bevindingen op dit punt hierboven vermeld bij de bespreking van richtlijn '2. Uitlegbaarheid'.

Zoals vermeld bij uitlegbaarheid, is vorm en diepte van documentatie en transparantie altijd afhankelijk aan wie er op welk moment gecommuniceerd moet worden. We denken dat de toepassing van de richtlijnen wordt gediend door publieksvoorlichting op te nemen als intrinsiek onderdeel. Dat reflecteert ook dat transparantie meerdere doelen tegelijkertijd dient: het ondersteunen van interne, collegiale uitlegbaarheid naast het faciliteren van interne en externe verantwoordelijkheid en toetsing (onafhankelijk van of dat nu een burger, rechter of auditor is).

Gaming the system vs. transparantie

Hoewel er geenszins sprake is van opsporing of fraudebestrijding wordt er bij de UWV Werkverkenner toch rekening gehouden met het risico van *gaming the system*: de online communicatie met aanvragers *nudged* om de bijbehorende vragenlijst in te vullen. Er wordt vervolgens beperkt informatie gegeven over de werking van de applicatie en de gevolgen daarvan: zowel om aanvragers niet te overladen met informatie als om het geven van wenselijke antwoorden enigszins te voorkomen.

Uit oogpunt van transparantie en democratische controle is de werking van de applicatie wel volledig online vindbaar, als men daar naar zou zoeken. Dat is te beschouwen als een vorm van getrapte transparantie.

Aanbevelingen bij bijlage publieksvoorlichting

- Overweeg om de inzichten in deze bijlage te integreren met de richtlijnen zelf.
- De opmerkingen over *qualifiers* zouden onderdeel moeten zijn van de richtlijnen zelf.

Aanvullende onderzoeksvragen

Voor zover deze vragen niet beantwoord werden in het vorige hoofdstuk, gaan we hier nog kort in op de volgende onderzoeksvragen:

- Zijn de richtlijnen toepasbaar in de context van de doelstelling waar binnen het algoritme is ingezet? Waarom wel of niet?

Algemeen gesproken kunnen de richtlijnen in beide casussen goed worden toegepast. Waar geïnterviewden een richtlijn, of onderdeel niet goed kunnen toepassen is dat veelal omdat zij vanuit hun rol niet precies weten wie voor dat onderdeel verantwoordelijk is of zou moeten zijn. Zie verder hierboven de opmerkingen per richtlijn.

- In hoeverre helpen de richtlijnen om risico's, zoals aansprakelijkheid en bias, te beheersen?

We merkten in ons onderzoek dat het actief en stap voor stap bespreken van de richtlijnen de aanzet vormt tot een geëngageerd gesprek en meermaals ook een dialoog tussen teamleden. Geïnterviewden namen, soms aanzienlijk, meer tijd dan gepland voor het gesprek, maakten aantekeningen en gaven na afloop aan dat ze op specifieke onderdelen van plan waren intern een aantal vragen te stellen.

Wel mist men een proces waarin de toepassing van de richtlijnen helder en deelbaar vastgelegd kan worden, zodat men kan weten en kan laten zien "dat we het goed doen". De richtlijnen werken op dit moment vooral als kader voor een gestructureerde dialoog maar zijn nog geen zelfstandig bruikbaar instrument.

We willen er hier expliciet voor pleiten om de doelgroep van de beslissing/het algoritme in dit proces een volwaardige rol te geven. Zowel vanuit oogpunt van betrokkenheid, legitimiteit als het borgen van diversiteit en kwaliteit.

- In hoeverre zijn de richtlijnen bij deze casus toegepast?

De richtlijnen zijn in de loop van 2019 door het ministerie vastgesteld op het moment dat beide applicaties al grotendeels klaar waren en konden door de betrokken organisaties daarom niet expliciet worden toegepast. Vanaf het begin van het ontwerp hebben betrokken teams oog gehad voor onderdelen die benoemd worden in de richtlijnen en daar ook interesse gehad. Bij beide organisaties waren ontwerpers van de algoritmen tijdens bouw en ontwikkeling op de hoogte dat op rijksniveau aan richtlijnen werd gewerkt en zij hebben eerdere versies wel eens gezien. Omdat andere richtlijnen eerder al definitief beschikbaar waren, zoals de "Ethics guidelines for trustworthy AI" van de Europese Commissie, werd die in een geval wel toegepast.

Aanbevelingen

Op basis van ons onderzoek zetten we hier een aantal aanbevelingen voor de richtlijnen in het algemeen op een rij. In het vorige hoofdstuk wordt specifiek ingegaan op de afzonderlijke richtlijnen en de relatie met de onderzochte casuïstiek. Hiermee geven we ook antwoord op de onderzoeksvraag:

- Wat zijn concrete suggesties om specifieke richtlijnen waar nodig te verbeteren ten behoeve van de praktische bruikbaarheid?

Aanbeveling 1-3 gaan inhoudelijk over de richtlijnen zelf. Aanbeveling 4 en 5 zijn maatregelen die het inbedden en toepassen van de richtlijnen bij overheden kunnen faciliteren.

1. Verscherp de definitie (waar gaan de richtlijnen over?)

In enge of meest precieze zin wordt 'algoritme' begrepen als 'rekenschema' of 'rekenwijze'. Maar het woord algoritme is zelf al gebaseerd op misverstanden: in de achtste eeuw schreef Al-Chwarizmi in het arabisch een verhandeling over decimale rekenkunde. Zijn boek "Opmerkingen van Al-Chwarizmi over de kunst van het rekenen met Indiase cijfers" kreeg in de twaalfde eeuw in Europa bekendheid onder de titel "Algorismi de numere Indorum". 'Algorismi' was de Latijnse verbastering van Al-Chwarizmi maar werd al snel verstaan als 'rekenwijze'. De latere spelling algoritme is te wijten aan verwarring met arithmos, Grieks voor 'getal'.[22]

De verwarring rondom de term algoritme zet zich in ons onderzoek vrolijk voort. Analytisch en technisch onderlegde gesprekspartners zien ze overal: een bureaucratisch proces is voor hen ook een algoritme.

> *"Een werkvoorbereider levert een projectvoorstel op. Dat is een PV met gevorderde gegevens en een verhaal van hoe de oplichter te werk is gegaan. In dat verhaal kan de werkvoorbereider zijn creativiteit en menselijkheid en ervaring kwijt. Voor de rest is het proces zelf al heel algoritmisch."*

Andere geïnterviewden lijken de term algoritme gelijk te stellen aan "puur de techniek" of kunstmatige intelligentie.[23] Omdat een technische stack op iedere laag bestaat uit meerdere algoritmen (begrepen als rekenschema's) maakt dit de reikwijdte van de richtlijnen ondoorzichtig.

De risico's die de richtlijnen adresseren doen zich voor in de interactie tussen algoritmes en de systemische context waarin deze worden ingezet. Het primaire onderwerp is dus het snijvlak van algoritmen en ambtenaar. Om te voorkomen dat de richtlijnen te beperkt of juist te breed worden opgevat adviseren we in de richtlijnen de definitie aan te passen en daarbij gebruik te maken van het onderzoek dat Maranke Wieringa hiernaar doet. Zij definieert een algoritmisch systeem als een socio-

[22] http://www.etymologiebank.nl/trefwoord/algoritme

[23] Onderzoeker Maranke Wieringa vertelde ons dat zij vertegenwoordigers van gemeenten wel eens hoort zeggen dat er bij hen "geen algoritmen worden gebruikt". Het woord wordt dan kennelijk gebruikt voor dingen die je niet begrijpt.

technische verzameling bestaande uit een combinatie van technische onderdelen, sociale praktijken en (organisatie)cultuur.[24] De overheid gebruikt deze socio-technische verzamelingen om beslissingen te nemen voor of over burgers. De richtlijnen kunnen niet goed worden toegepast zonder ook (het grotere) vraagstuk rondom ketenverantwoordelijkheid en geautomatiseerde besluiten te adresseren en expliciet te maken hoe deze richtlijnen zich hiermee verhouden.[25]

2. Benoem rollen

Definieer voor wie de richtlijnen zijn en maak ze voor deze rollen specifiek en compact. Het ligt voor de hand dat een programmamanager of proceseigenaar verantwoordelijk is voor de toepassing van de richtlijnen. De toepassing van de richtlijnen zal voor een belangrijk deel bestaan uit het organiseren van interne dialoog en afstemming over het treffen van de juiste maatregelen bij ontwikkeling en gebruik van algoritmen. Wij adviseren op zijn minst de volgende zes:

- Programmamanager, proceseigenaar, afdelingshoofd: verantwoordelijk voor de beslissingen die met het algoritme worden genomen.
- Architect/onderzoeker: hoofdontwerper van het algoritmisch systeem, begrijpt innerlijke werking van het algoritme, inclusief de relevante domeinkennis.
- Business/regel-analist, projectleider, applicatiebeheerder: vertaalt specifieke onderdelen van wet, beleid, en regels naar beslisregels, *requirements* en andere ontwerp instructies.
- Softwareontwikkelaar: vertaalt *requirements* naar code, verantwoordelijk voor versiebeheer (en documentatie), tests, (afstemming met) technisch en systeembeheer en beveiliging.
- Communicatieafdeling: communicatie met de burger aangaande de afdeling of processen waarbij het algoritme wordt gebruikt.
- Beslismedewerker: medewerker die voor de burger waarop het besluit betrekking heeft het eerste aanspreekpunt is en onderliggende motivatie biedt.

De mate waarin deze rollen betrokken en verantwoordelijk zijn verschilt: in ontwerp- en testfase zijn kleine hoeveelheden beslismedewerkers betrokken en houdt architect/onderzoeker dagdagelijks overzicht. Als het algoritme is geïmplementeerd en onderdeel is van lijnactiviteiten, is die verhouding omgekeerd. Afhankelijk van de fase (ontwerp, pilot, test, implementatie) hebben rollen een andere betrokkenheid met en verantwoordelijkheid voor de toepassing van de richtlijnen. Uitgangspunt is dat voor alle rollen in iedere fase de betrokkenheid geborgd moet zijn.

3. Vereenvoudig de opzet

In de tekst kan het onderscheid tussen richtlijnen (die voorschrijvend en stellig van aard zijn) en onderdelen die een verduidelijkende / verhelderende functie hebben worden aangescherpt.

---

[24] Maranke Wieringa, 'What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability', in: Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA.

[25] Zie: Marlies van Eck, 'Geautomatiseerde ketenbesluiten & rechtsbescherming': https://pure.uvt.nl/ws/portalfiles/portal/20399771/Van_Eck_Geautomatiseerde_ketenbesluiten.pdf

Op basis van ons onderzoek denken we dat de zeven richtlijnen kunnen worden gegroepeerd in vier categorieën.

1. **Wet en beleid:** Expliciet moet worden aangetoond dat het toegepaste algoritme en het omliggende proces geschikt en proportioneel zijn voor het doel zoals vastgelegd in wet, beleid en regels. Deze afweging moet publiek beschikbaar zijn. Dit omvat ook de richtlijnen voor validatie en gegevensherkenning. In de toepassing van de richtlijnen kan het nuttig zijn te verwijzen naar richtlijnen en standaarden voor specifieke domeinen of vakgebieden (Algemene beginselen van behoorlijk bestuur, de eerder genoemde Standard for Public Code[26], BurgerServiceCode[27]).

2. **Impact:** De context waarin het algoritme wordt toegepast is bepalend voor de manier waarop je de richtlijnen wil toepassen. Met daarbij het onderscheid:
    A. Impact van algoritme op de beslissing
    B. Impact van de beslissing (op de burger)

    In procesontwerpen worden (on)voorziene risico's vaak geadresseerd door menselijke monitoring/interventie. Zowel de kwaliteit als de kwantiteit hiervan moet geborgd worden.

3. **Verantwoording:** Wie is verantwoordelijk voor de toepassing van de richtlijnen? Hoe krijgt die verantwoordelijkheid vorm, wie is daar intern bij betrokken? Hoe wordt geborgd dat verantwoordelijkheid betekenisvol gedragen kan worden met voldoende domeinkennis, technologisch inzicht en overzicht over het proces? Hoe is dit auditeerbaar en toetsbaar?

4. **Uitlegbaarheid**: Maak onderscheid voor wie de uitlegbaarheid beschikbaar moet zijn, op welk moment en welke maatregelen daarvoor zijn genomen.
    A. Publieke uitlegbaarheid: leg publiek en transparant uit waarom dit algoritme wordt ingezet, op welke wetgeving en beleid dat is gebaseerd en waar algoritme en systeem voor worden geoptimaliseerd.
    B. Collegiale uitlegbaarheid, borging, bewaking en verantwoording: leg uit wat er in het operationele proces nodig is om de juiste werking van het algoritme te borgen. Waar moeten bijvoorbeeld (beslis)ambtenaren rekening mee houden bij het monitoren of begrijpen van uitkomsten van algoritmen?
    C. Uitlegbaarheid van een specifieke beslissing: spontaan aangeboden uitleg in begrijpelijke taal over de totstandkoming van een beslissing, gericht op een gelijkwaardige informatiepositie voor de burger. Gekoppeld aan laagdrempelige manieren om in contact te komen.

[26] https://standard.publiccode.net/

[27] https://www.noraonline.nl/wiki/BurgerServiceCode

4. Koppel toepassing richtlijnen aan dialoog, voorlichting en verantwoording

De richtlijnen zijn bedoeld als kader/instrument voor het (kunnen) toepassen van geschikte maatregelen. Daarnaast werken de richtlijnen op dit moment vooral als kader voor een gestructureerde dialoog maar zijn nog geen zelfstandig bruikbaar instrument. Daarvoor moet worden voorzien in tooling om ketenverantwoordelijkheid in kaart te brengen, interne dialoog te organiseren en resultaten daarvan vast te leggen. Dit komt ook tegemoet aan de behoefte van mensen om te weten wanneer men het goed doet. Hierbij kan het organisaties helpen als er een *impact assessment* beschikbaar is. In de onderzochte casussen wordt er vanuit gegaan dat de burger zelf in actie komt bij een ongewenste uitkomst. Afhankelijk van de uitkomst en impact van een beslissing kan het zijn dat dit onvoldoende borging biedt dat de corrigerende actie ook tijdig wordt ingezet.

De resultaten van dit soort processen kunnen vervolgens worden ingezet bij voorlichting en verantwoording.

5. Maak de richtlijnen zelf onderdeel van toetsing en doorontwikkeling

Over de impact en ontwikkeling van het grootschalig en ingrijpend inzetten van geavanceerde technologie bestaat veel onzekerheid, gevoed door angsten en hypes. Systemische en onvoorziene risico's gaan juist spelen als algoritmen op schaal worden ingezet en worden pas met de tijd zichtbaar. In de brief van minister Dekker wordt dan ook aangegeven dat het gaat om 'materie die relatief nieuw is'.[28]

Uit oogpunt van kwaliteit en borging in maatschappelijke democratie is het belangrijk om de doorontwikkeling van de richtlijnen zelf en de toepassing daarvan onderdeel te maken van een publieke dialoog. Bij dit proces horen niet alleen relevante overheidsorganisaties betrokken te worden maar ook wetenschap, professionals, maatschappelijke organisaties en burgers.

Betrek en onderzoek daarbij de verhouding en afstemming met belangrijke andere standaarden op dit gebied, ook in internationaal verband. Een aantal Nederlandse voorbeelden op dit gebied worden in dit onderzoek genoemd.

[28] https://www.rijksoverheid.nl/documenten/kamerstukken/2019/10/08/tk-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid

Bijlagen

Lijst van geïnterviewden en betrokkenen

Politie, casus Keuzehulp: Dick van Kuilenburg, Peter Laveman, Gijs van der Linden, Bas Testerink

UWV, casus Werkverkenner: Margot van Engen, Maurice Guiaux, Vincent van der Heiden, Samyaa el Jout, Joeri van Proosdij, Maud van Vuren, Martijn Wijnhoven.

In het kader van het bureauonderzoek en of de werksessie spraken wij daarnaast met: Remco Boersma (J&V), Fiora van de Bosch (DUO), Alex Corra (SVB), Sabine den Daas (Bluefield), Marlies van Eck (H&P), Marc Gerrard (BZK), Matthijs van Kempen (Belastingdienst), Maike Klip (DUO), Kübra Kul-Özbasi (UWV), Edwin Rijgersberg (NFI), Jason Schipper (Gem. Nissewaard), Maranke Wieringa (UU).

Namens Waag werden bijdragen geleverd door Tom Demeyer, Tiwánee van der Horst, Alain Otjens, Douwe Schmidt en Marleen Stikker.

Bronnen keuzehulp online aangifte bij online oplichting

Nieuwsartikel AD: https://www.ad.nl/nieuws/ieder-jaar-duizenden-onterechte-aangiftes-internetoplichting-politie-begint-tegenoffensief~a069e5ea/

Floris Bex et al., 'A.I. for Online Criminal Complaints: From Natural Dialogues to Structured Scenarios', http://www.florisbex.com/papers/AI4J2016.pdf

William Kos et al., 'Classification in a Skewed Online Trade Fraud Complaint Corpus', (2017), http://www.florisbex.com/papers/BNAIC_2017_Kos.pdf

Martijn Schraagen et al., 'Argumentation-driven information extraction for online crime reports', (2018), http://www.florisbex.com/papers/LeDAM2018.pdf

Martijn Schraagen et al., 'Evaluation of Named Entity Recognition in Dutch online criminal complaints', (2017), https://www.umiacs.umd.edu/~oard/desi7/papers/MS2.pdf

Martijn Schraagen et al., 'Evaluation of Named Entity Recognition in Dutch online criminal complaints', (2017), https://clinjournal.org/clinj/article/download/65/58/

Martijn Schraagen et al., `Extraction of semantic relations in noisy user-generated law enforcement data`, http://www.florisbex.com/papers/relation_extraction_icsc.pdf

Bas Testerink et al., 'A Method for Efficient Argument-based Inquiry', http://www.florisbex.com/papers/FQAS_ArgQA.pdf

Bas Testerink, 'Two use-case assessments with the EU HLEG ethics guidelines', (in voorbereiding, tussentijdse versie, 2019)

Bronnen UWV Werkverkenner

Inleidende documentatie

Guiaux, M., Wijnhoven, M. & Havinga, H. (2018). Werkverkenner 2.0: De wetenschappelijke doorontwikkeling van een model waarmee UWV de kansen op werk voorspelt van werkzoekenden met een WW-uitkering. UWV Kennisverslag, 2018-8, 3-9. https://www.uwv.nl/overuwv/Images/uwv-kennisverslag-2018-8.pdf

Wijnhoven, M. & M. Guiaux (2019). Evidencebased werken bij dienstverlening aan werkzoekenden: Over het ontstaan en gebruik van de Keuzehulp dienstverlening WW, UWV Kennisverslag 2019-6, 2-9. https://www.uwv.nl/overuwv/Images/ukv-2019-6-evidencebased-werken-bij-dienstverlening-aan-werkzoekenden.pdf

Wetenschappelijk onderzoek

Dusseldorp, E., Hofstetter, H. & Sonke, C. (2018). Landelijke doorontwikkeling van de UWV Werkverkenner: Eindrapportage. Leiden: TNO. https://www.uwv.nl/overuwv/Images/landelijke-doorontwikkeling-uwv-Werkverkenner-eindrapport.pdf

Aanvullend tabellenboek bij deze publicatie: https://www.uwv.nl/overuwv/Images/bijlages-bij-landelijke-doorontwikkeling-uwv-Werkverkenner.pdf

Rijksuniversiteit Groningen UMCG & UWV Kenniscentrum (2011): Eindrapportage Project voorspellers van Werkhervatting: Een onderzoek onder werklozen in Noord-Holland. Groningen/Amsterdam. www.rug.nl/research/gezondheidswetenschappen/tgo/rapporten/pdf-files/2011-pdf/brouwer-ea-2011-eindrapportage.pdf

Overige publicaties over de Werkverkenner

Brouwer, S., R.H. Bakker & J.M.H. Schellekens (2015): Predictors for re-employment success in newly unemployed, Journal of Vocational Behavior 89: 32-38. www.sciencedirect.com/science/article/pii/S0001879115000366

Havinga, H. (2014): Ontwikkeling en invoering van de Werkverkenner, UWV Kennisverslag 2014-3: 27-34. www.uwv.nl/overuwv/Images/20141209_UKV_2014_03%20DIG.pdf

Havinga, H. & W. Hijlkema (2012): Wat is de Persoonsverkenner? UWV Kennisverslag 2012-2: 44-55. http://www.uwv.nl/overuwv/Images/UKV%202012-II.pdf

Havinga, H., E. van Wijk & J. van Rijssen (2016): Van WW naar Ziektewet: Wat zegt de Werkverkenner? UWV Kennisverslag 2016-1: 3-11. www.uwv.nl/overuwv/Images/UKV%2001-2016%20DIG%20.pdf

Wijnhoven, M.A. & H. Havinga (2014): The Work Profiler: A digital instrument for selection and diagnosis of the unemployed, Local Economy 29 (6-7): 740-749. http://lec.sagepub.com/content/29/6-7/740.full.pdf+html

Wijnhoven, M.A.& H. Havinga (2015): Werkverkenner voorspelt kans op werk voor WW'ers, Sociaal Bestek 77 (3): 56-57.

Wijnhoven, M.A. & P. Hilbers (2014): Werkverkenner, dan krijg je wat, UWV Kennisverslag 2014-3: 35-41. www.uwv.nl/overuwv/Images/20141209_UKV_2014_03%20DIG.pdf

Overige bronnen

AlgorithmWatch, 'AI Ethics Guidelines Global Inventory', (bez. november 2019), https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

Gregory Barber, `Artificial Intelligence Confronts a 'Reproducibility' Crisis`, (2019), https://www.wired.com/story/artificial-intelligence-confronts-reproducibility-crisis/

Petra Bíró, 'Algorithm says no: ethische richtlijnen voor AI systemen', (2019), https://waag.org/nl/article/algorithm-says-no-ethische-richtlijnen-voor-ai-systemen

Ruben Boyd, ``We moeten veel meer uitleggen' Technische transparantie neemt algoritmeangst niet weg', (2019), https://ibestuur.nl/nieuws/we-moeten-veel-meer-uitleggen

Ministerie van BZK, `Maatschappelijke dialoog magazine` (2019), https://www.digitaleoverheid.nl/wp-content/uploads/sites/8/2019/09/BZK-Maatschappelijke-dialoog-Magazine-1-2019-T.pdf

Tom Demeyer, 'AI in Culture & Society', (2019), https://waag.org/nl/article/ai-culture-society

Marlies van Eck, 'Geautomatiseerde ketenbesluiten & rechtsbescherming', (2018), https://pure.uvt.nl/ws/portalfiles/portal/20399771/Van_Eck_Geautomatiseerde_ketenbesluiten.pdf

Foundation for Public Code, 'Standard for Public Code', v.0.1.4., (2019), https://standard.publiccode.net/

Valerie Frissen et al., 'Onderzoek Toezicht op het gebruik van algoritmen door de overheid', Hooghiemstra en Partners, (2019), https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2019Z26117&did=2019D53638

Jessica Fjeld et al., 'Principled Artificial Intelligence. A map of ethical and rights-based approaches (draft)', (2019), https://ai-hr.cyber.harvard.edu/primp-viz.html

Hannah Fry, `Algoritmes aan de macht: Hoe blijf je menselijk in een geautomatiseerde wereld?`, (2018)

H. Hilligoss, 'Introducing the Principled Artificial Intelligence Project', (2019), https://blogs.harvard.edu/cyberlawclinic/2019/06/07/introducing-the-principled-artificial-intelligence-project/

ICTU, 'BurgerServiceCode', (2008), https://www.noraonline.nl/wiki/BurgerServiceCode

Chris Julien, 'Biased by Default', (2019), https://waag.org/sites/waag/files/2019-06/Biased-by-default.pdf

Matthijs van Kempen, Motivering van automatisch genomen besluiten, (2019), http://www.knowbility.nl/wp-content/uploads/Motivering%20van%20automatisch%20genomen%20besluiten%20DefPub%20digitaal.pdf

Rob Kerstens, 'Regels en Ruimte. Verkenning Maatwerk in dienstverlening en discretionaire ruimte', (2019), Ministerie van BZK, https://www.rijksoverheid.nl/documenten/rapporten/2020/01/16/bijlage-rapport-abdtopconsult-maatwerk-dienstverlening

Maike Klip, blog over begrip, empathie bij ambtenaren en bouwers van digitale diensten op presentatie, (2019), https://klipklaar.nl/

Linda Kool et al, `Opwaarderen - Borgen van publieke waarden in de digitale samenleving', (2017), https://www.rathenau.nl/sites/default/files/2018-02/Opwaarderen_FINAL.pdf

Meijer, A. et al., 'Principes voor goed lokaal bestuur in de digitale samenleving. Een aanzet tot een normatief kader', (2019), Bestuurswetenschappen. 73(4), https://tijdschriften.boombestuurskunde.nl/tijdschrift/bw/2019/4/Bw_0165-7194_2019_073_004_003

The Oxford Internet Institute, 'Principles Are No Guarantee Of Ethical AI, Says Oxford Ethicist`, (2019), https://www.oii.ox.ac.uk/news/releases/principles-are-no-guarantee-of-ethical-ai-says-oxford-ethicist/

Jan Terpstra, 'De abstracte politie', (2018), Het tijdschrift voor de politie, https://www.websitevoordepolitie.nl/de-abstracte-politie/

Totta Data Lab, 'de 7 design principles van onze algoritmen', (2019), https://www.tottadatalab.nl/2018/09/24/design-principles-algoritmes/

Max Vetzo et al., 'Juridische onderzoeksrapport Algoritmes en grondrechten', (2018), Universiteit Utrecht, https://www.uu.nl/sites/default/files/rebo-montaigne-algoritmes_en_grondrechten.pdf

Maranke Wieringa, 'What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability', in: Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA.

WRR,'WRR-rapport nr. 97: Weten is nog geen doen. Een realistisch perspectief op redzaamheid', (2017), https://www.wrr.nl/publicaties/rapporten/2017/04/24/weten-is-nog-geen-doen