Autoriteit
Consument & Markt

Position Paper

# Toezicht op algoritmes

Muzenstraat 41
2511 WB Den Haag

www.acm.nl
070 722 20 00

Kansen & keuzes voor bedrijven en consumenten

# Inhoudsopgave

# 1 Inleiding

### *Algoritmes zijn overal*

Algoritmes en de toepassing ervan zijn al eeuwenoud.[1] In abstracte zin is een algoritme niets meer dan een set aan stappen c.q. instructies die worden uitgevoerd om een bepaald doel te bereiken. Zo zou je een recept voor een gerecht kunnen beschouwen als een algoritme: op basis van het recept (de 'functie' of het 'model') voert men bepaalde handelingen uit met de ingrediënten (de invoer) die uiteindelijk leiden tot een bepaald gerecht (de uitvoer). Algoritmes zijn echter veelal bekend om hun toepassing in de wiskunde en informatica.

Binnen de automatisering spelen algoritmes een belangrijke rol. Denk bijvoorbeeld aan relatief eenvoudige taken, zoals het omzetten van fysieke muisbewegingen naar cursorbewegingen op het beeldscherm. Door de continue toename in rekenkracht en de hoeveelheid beschikbare data, heeft de inzet van algoritmes in allerlei domeinen een grote vlucht genomen. Met behulp van (zelflerende) algoritmes is het inmiddels mogelijk om op basis van een constante stroom (en veelal ongestructureerde) gegevens, bijvoorbeeld voorspellingen te doen over het weer of klimaat of patronen in beelden te herkennen die voor de mens niet of nauwelijks zichtbaar zijn.

Als men tegenwoordig over algoritmes spreekt bedoelt men veelal softwaretoepassingen, waarvan algoritmes onderdeel uitmaken, die op geautomatiseerde wijze voorspellingen doen, besluiten nemen of advies geven. In die betekenis opereren algoritmes veelal binnen een bepaalde context, waarbij de gebruikte gegevens, de betrokken personen en de implementatie in een organisatie allemaal relevante elementen zijn voor de uitwerking van algoritmes. Men zou daarom beter kunnen spreken van 'algoritmische toepassingen' in plaats van 'algoritmes' als zodanig.

### *Toezicht op algoritmische toepassingen*

Inmiddels is het begrip gegroeid dat algoritmische toepassingen steeds vaker een (beslissende) rol spelen bij activiteiten die een direct effect hebben op mensen, bedrijven, organisaties en de samenleving als geheel. Bedrijven maken in toenemende mate gebruik van algoritmische toepassingen in hun bedrijfsvoering. Bijvoorbeeld voor een optimale inrichting van productieprocessen, routering van bezorgers, geïndividualiseerde aanbiedingen, het dynamisch aanpassen van aanbod of bijvoorbeeld dynamische prijszetting. Dit levert veel voordelen op voor de maatschappij. Maar er zijn ook veel maatschappelijke zorgen over algoritmische toepassingen die bedrijven gebruiken om hun producten en diensten aan te bieden aan consumenten.[2] Bijvoorbeeld omdat bedrijven online steeds beter in staat zijn consumenten te sturen in hun keuzes en aankopen. Zij zetten allerlei tactieken in om het online gedrag van consumenten te beïnvloeden. Waar gaat verleiden over in misleiden? De Autoriteit Consument & Markt heeft dit recent verduidelijkt door grenzen te stellen aan online beïnvloeding in de Leidraad bescherming van de online consument.[3]

Er is momenteel veel politieke en maatschappelijke aandacht voor de gevolgen van algoritmische toepassingen voor mensen, bedrijven en de maatschappij als geheel. In dat verband spreekt men vaak over kunstmatige intelligentie c.q. artificiële intelligentie (AI). In 2019 heeft het kabinet het Strategisch Actieplan voor Artificiële Intelligentie [4] gepresenteerd, waarin haar voornemens omtrent AI-beleid staan beschreven. Eén van de sporen van dit plan beschrijft hoe publieke belangen geborgd moeten blijven bij AI-ontwikkelingen. In 2020 reageerde het kabinet in afzonderlijke Kamerbrieven op de initiatiefnota van het

[1] Een van de bekendste algoritmes, het algoritme van Euclides, stamt uit de klassieke oudheid en de naamgever van het concept algoritme, Al-Chwarizmi, die een grote bijdrage leverde aan de algoritmiek, leefde rond 800 in Perzië.
[2] Uiteraard zijn de zorgen over de inzet van algoritmische toepassingen door bedrijven breder dan de zorgen die direct de taken van de ACM raken, zoals zorgen over ongelijke behandeling en discriminatie.
[3] https://www.acm.nl/sites/default/files/documents/2020-02/acm-leidraad-bescherming-online-consument.pdf.
[4] https://www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie

Kamerlid Middendorp[5] en op een drietal onderzoeken naar algoritmes.[6] In deze brieven wordt gesproken over de kansen en risico's van AI-ontwikkelingen bij bedrijven, maar ook over het toezicht op algoritmes.

Niet alleen in Nederland bestaan zorgen over de gevolgen van algoritmische toepassingen voor mensen en bedrijven. Er zijn verschillende internationale initiatieven om normenkaders te ontwikkelen voor algoritmische toepassingen. De Europese Commissie heeft bijvoorbeeld in het kader van haar strategie voor data en AI[7] een Witboek over AI gepresenteerd met daarin een kader voor excellente en betrouwbare AI.[8] Dit strategische EU-kader op basis van fundamentele waarden moet ervoor zorgen dat mensen AI kunnen vertrouwen en bedrijven aanmoedigen AI-oplossingen te ontwikkelen.

Het bestaande normenkader verder invullen voor toepassing in de digitale economie is op zichzelf niet voldoende om de maatschappelijke zorgen weg te nemen. In een steeds verder digitaliserende economie moeten toezichthouders zoals de Autoriteit Consument en Markt (ACM) de normen waar zij op toezien ook kunnen handhaven wanneer bedrijven algoritmes gebruiken om hun marktgedrag te bepalen. Alleen dan kunnen mensen en bedrijven erop vertrouwen dat digitale markten goed voor hen blijven werken.

Dit position paper is een vertrekpunt van waaruit de ACM het toezicht op algoritmische toepassingen verder wil ontwikkelen. Het biedt algemene handvatten voor onderzoeken naar overtredingen waarbij algoritmische toepassingen een rol spelen. Het paper beschrijft eerst waarom algoritmische toepassingen relevant zijn voor de ACM (hoofdstuk 2) en wat algoritmes nu precies zijn (hoofdstuk 3). Vervolgens wordt beschreven hoe de ACM in praktijk algoritmische toepassingen kan onderzoeken (hoofdstuk 4).

[5] https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2020Z07093&did=2020D15106
[6] https://www.rijksoverheid.nl/documenten/kamerstukken/2020/11/20/tk-kabinetsreactie-op-drietal-onderzoeken-naar-algoritmen
[7] Zie https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/shaping-europe-digital-future_nl.
[8] Witboek over kunstmatige intelligentie — een Europese benadering op basis van excellentie en vertrouwen, zie https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_nl.pdf.

# 2 Wanneer zijn algoritmische toepassingen relevant voor de ACM?

De ACM draagt bij aan een gezonde economie door markten goed te laten werken voor mensen en bedrijven, nu en in de toekomst. In goed functionerende markten concurreren bedrijven eerlijk met elkaar en benadelen zij niemand met oneerlijke praktijken. Algoritmische toepassingen van marktpartijen kunnen ervoor zorgen dat markten minder goed functioneren. Voor de ACM zijn algoritmische toepassingen relevant wanneer ze een rol spelen bij activiteiten die de toezichtsdomeinen van de ACM raken. Het gaat dus om de concrete toepassing ervan bij activiteiten die een effect hebben op consumenten of marktpartijen. Algoritmische toepassingen die bijvoorbeeld prijzen bepalen, vraag en aanbod sturen op de energiemarkt, of het aanbod richting consumenten personaliseren zijn relevant voor de ACM. Denk bijvoorbeeld ook aan algoritmische toepassingen die leiden tot prijsdiscriminatie of kartelvorming tussen marktpartijen. Of tot een zodanige inrichting van de online keuzearchitectuur bij de aankoop van een product dat de consument, tegen zijn eigen economische belangen, een beslissing over een transactie neemt die hij anders niet had genomen. Algoritmische toepassingen die bijvoorbeeld sturen welke berichten in een bepaalde volgorde op een interne pagina van een bedrijf worden getoond, zullen niet snel relevant zijn voor de ACM.

De meest voorkomende functies waarvoor marktpartijen algoritmische toepassingen in de digitale economie inzetten zijn:[9]

- **Zoekfunctionaliteit**: het tonen en rangschikken van informatie op basis van bepaalde input.
- **Aggregatie**: het verzamelen, categoriseren en herschikken van informatie uit verschillende bronnen. Denk bijvoorbeeld aan het verzamelen van prijsinformatie van producten van concurrenten.
- **Observatie (surveillance):** het observeren van gedrag en patronen om afwijkingen te identificeren. Denk aan netwerkdetectie of fraudedetectie bij transactiedata.
- **Voorspellen**: het voorspellen van toekomstig gedrag of scenario's.
- **Filters**: het veelal op de achtergrond filteren (blokkeren) van informatie of data. Denk bijvoorbeeld aan spamfilters of filters om (vermeend) inbreukmakend materiaal te weren.
- **Aanbevelingssystemen**: het aanbevelen van bepaalde informatie of producten veelal op basis van (gedrags)data over de gebruiker, het product en/of andere parameters.
- **Scoren en rangschikken**: het scoren c.q. rangschikken van informatie, producten, bedrijven en/of consumenten. Denk aan online reviewscores en aan kredietscores van consumenten.
- **Informatieproductie**: produceren van informatie. Denk aan geautomatiseerde nieuwsberichten, zoals geautomatiseerde berichten over aandelen- en beurskoersen of sportwedstrijden.
- **Communicatie**: het geautomatiseerd communiceren met consumenten en/of bedrijven. Denk aan de communicatie tussen consument en chatbot of de virtuele asissent die ten behoeve van de consument communiceert met derden.
- **Allocatie**: het geautomatiseerd uitvoeren van transacties en verdelen en toewijzen van vraag en aanbod. Denk aan de geautomatiseerde verkoop van online advertentieruimte (real-time bidding) of het koppelen van een klant met een beschikbare taxi.

[9] Onderstaande typologie is grotendeels overgenomen uit: Latzer, M. & Festic, N. (2019). A guideline for understanding and measuring algorithmic governance in everyday life. Internet Policy Review, 8(2).

# 3 Wat is een algoritme en welke varianten zijn er?

Een algoritme is in abstracte zin niets meer dan een set aan stappen of instructies die worden uitgevoerd om een bepaald doel te bereiken. Voor de betekenis van de term 'algoritme' sluiten we aan bij de definitie die de WRR hanteert: "een geautomatiseerde reeks stappen die inputdata in outputdata omzet".[10] Algoritmes zijn op verschillende manieren te categoriseren.[11] Dit hoofdstuk beperkt zich tot een categorisering aan de hand van enkele veel voorkomende werkwijzen van algoritmes.

De wijze waarop algoritmes werken, is te onderscheiden in verschillende varianten. Zo kan een algoritme bestaan uit een eenvoudige beslisboom waarvan de beslisregels vooraf zijn bepaald. Die beslisregels worden vervolgens veelal geautomatiseerd uitgevoerd, maar het is ook mogelijk dat een persoon deze regels uitvoert. Er zijn ook varianten die op basis van trainingsdata patronen en verbanden identificeren op basis waarvan vervolgens een model (wat zelf ook weer een algoritme is) wordt gegenereerd of aangepast. Vervolgens kan het model ingezet worden om op basis van inputdata een bepaalde output te genereren, bijvoorbeeld een voorspelling. Deze werkwijze wordt ook wel 'machine learning' genoemd. Binnen de categorie machine learning zijn er weer verschillende methoden te onderscheiden. Hieronder bespreken we de meest gangbare varianten.

## 3.1 Statische algoritmes

Statische algoritmes zijn algoritmes waarbij de regels (of de instructies) van het algoritme kennisgedreven, door mensen, vooraf worden geprogrammeerd. Het zijn veelal als 'X' dan 'Y' instructies op basis waarvan beslissingen worden gemaakt of voorspellingen worden gedaan. Een softwareontwikkelaar gebruikt bijvoorbeeld bepaalde kennis over kredietrisico's om de regels van een algoritme vast te stellen waarmee bepaald kan worden of iemand in aanmerking komt voor een lening en zo ja, tegen welk rentetarief. Bijvoorbeeld een algoritme dat op basis van een aantal datapunten (o.a. hoogte inkomen, schulden, BKR-registratie, leeftijd, looptijd lening, waarde onderpand, rentestanden) of en wat de maximaal te verstrekken lening is tegen welke rente. Dergelijke algoritmes kunnen als 'statisch' worden getypeerd omdat de regels van het algoritme, zonder menselijke tussenkomst, ongewijzigd blijven. Er is geen geautomatiseerd feedbackmechanisme aanwezig op basis waarvan het algoritme zelf 'leert' en zich aanpast.

## 3.2 Zelflerende algoritmes en machine learning

De term 'machine learning' wordt veelal gebruikt voor algoritmes die aan de hand van trainingsdata patronen en verbanden extraheren op basis waarvan een model wordt gegenereerd of aangepast. Dit model kan vervolgens ingezet worden om aan de hand van inputdata een bepaalde output te genereren, bijvoorbeeld om een voorspelling te doen. Het is een datagedreven leerproces. Bij zelflerende algoritmes past het model zichzelf (continu) aan tijdens de taakuitvoering op basis van ingebouwde feedbackmechanismen. Er zijn verschillende methoden van machine learning. Een aantal veelgebruikte methoden komen hieronder aan bod.

### *Supervised learnig*

Bij supervised learning wordt het model getraind aan de hand van voorbeelddata waarvan de input en de te verwachten output bekend is. De te verwachten output wordt ook wel 'ground truth' of gelabelde data genoemd. Tijdens de trainingsfase leert het algoritme welke eigenschappen van de input van invloed zijn op de output en past het model hierop aan. Het leert dus van de historische voorbeelddata. Vervolgens kan het model worden toegepast op nieuwe gegevens. Deze methode wordt daarom vaak ingezet om op basis van

[10] WRR-Rapport nr. 95, 'Big data in een vrije en veilige samenleving', 2016, p. 21. Zie ook het onderzoeksrapport dat in opdracht van het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties is uitgevoerd: Hooghiemstra & Partners, 'Toezicht op gebruik van algoritmen door de overheid', 2019.
[11] Zie voor een uitgebreide categorisering van algoritmes (functie/doel, werkwijze, inputdata, interpreteerbaarheid en aard ontwikkelaar): Bundeskartellamt, Autorité de la concurrence, 'Algorithms and Competition', November 2019, p.4-11.

historische gegevens toekomstige situaties te voorspellen. Een voorbeeld van supervised learning is geautomatiseerde spamdetectie waarbij een algoritme getraind wordt aan de hand van e-mails die door mensen gelabeld zijn als zijnde wel of geen spam.

### *Unsupervised learning*

Bij unsupervised learning wordt er geen gebruik gemaakt van voorbeelddata waarbij de input geclassificeerd, gelabeld of gecategoriseerd is. Er vindt dus geen sturing plaats aan de hand van voorbeelden. Het algoritme zal de inputdata dus zelfstandig systematiseren door bijvoorbeeld de data te clusteren op basis van gedeelde of vergelijkbare eigenschappen. Deze methode wordt bijvoorbeeld ingezet om consumenten bepaalde producten aan te bevelen.

### *Semi-supervised learning*

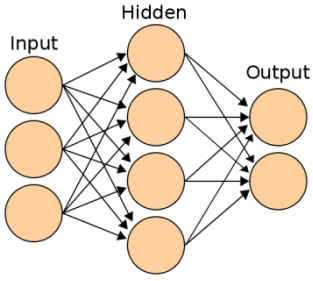
Deze variant is een combinatie van supervised en unsupervised learning. Voor het trainen van het algoritme wordt gewerkt met zowel gelabelde data (inputdata met de daarbij behorende outputdata) als niet gelabelde data. Dit is een variant die in de praktijk regelmatig ingezet wordt omdat men vaak over grote datasets beschikt die grotendeels nog niet gelabeld zijn. Het door mensen laten labelen van data is tijdrovend en kan vrij kostbaar zijn. Bij deze methode wordt eerst aan de hand van de gelabelde data het model getraind. Bijvoorbeeld een model voor spraakherkenning wordt eerst getraind op basis van gesproken tekst waarvan de betekenis al bekend is (gelabelde data) Vervolgens wordt het model verder getraind met nog niet gelabelde spraakdata.

### *Reinforcement learning*

Bij reinforcement learning wordt het algoritme getraind aan de hand van ‘trial and error’. Acties worden beloond of bestraft naargelang er wel of geen stap wordt gezet richting de gewenste uitkomst. Het is een iteratief leerproces waarbij het algoritme leert door de beloning te maximaliseren en straf te minimaliseren. Het grote verschil met supervised en unsupervised learning is dat bij deze methode het algoritme niet getraind wordt aan de hand van (gelabelde) trainingsdata. Deze methode werd bijvoorbeeld ingezet om het algoritme van het computerprogramma ‘AlphaGo’ te trainen waardoor het in staat was om de beste menselijke spelers van het bordspel Go te verslaan.[12] Een groot voordeel van deze methode is dat het niet afhankelijk is van historische data. Kwalitatief goede historische data die bijvoorbeeld nauwkeurig gelabeld is, is immers niet altijd voorhanden.

### *Neurale Netwerken en Deep Learning*

Een aparte categorie algoritmes, waarbij de ontwikkelingen snel gaan, zijn de algoritmes waarbij sprake is van ‘deep learning’. Bij deze variant van machine learning wordt gewerkt met de structuur van neurale netwerken. Deze werkwijze is geïnspireerd op de wijze waarop hersenen werken. Net als bij de hersenen bestaat een neuraal netwerk uit een netwerk van neuronen. Bij een neuraal netwerk is er een inputlaag van neuronen, een of meerdere ‘verborgen’ lagen van neuronen, en een outputlaag van neuronen. De neuronen van de ene laag kunnen een signaal (outputdata) afgeven aan de volgende laag (inputdata) die vervolgens weer een signaal kunnen afgeven aan de volgende laag, enzovoort. De statistische output van de ene laag vormt de input voor de volgende laag. Of de output van de vorige laag ook wordt gebruikt in de volgende laag, hangt onder andere af van de drempelwaardes die worden toegepast.




**Figuur 1 - Schematische weergave neuraal netwerk**

[12] https://deepmind.com/research/case-studies/alphago-the-story-so-far.

De term ‘deep learning’ wordt gebruikt voor neurale netwerken waarbij er meerdere verborgen lagen zijn. Deze methode is vooral nuttig voor patroonherkenning in ongestructureerde data. Zo wordt deze methode ingezet voor toepassingen zoals spraak- en beeldherkenning. Een belangrijk kenmerk van deze methode is dat door de complexiteit de werking en het gedrag van het algoritme minder transparant zijn.

# 4 Hoe kan de ACM algoritmische toepassingen onderzoeken?

## 4.1 Onderzoeksbevoegdheden

De ACM kan onderzoek doen naar de inzet van algoritmische toepassingen door marktpartijen om informatie of bewijs te vergaren over een mogelijke overtreding. De omvang van dit onderzoek zal afhangen van de elementen die bewezen moeten worden en de informatie die dit onderzoek kan opleveren.

De ACM kan de volgende bevoegdheden inzetten om algoritmische toepassingen te onderzoeken:[13]

- Het betreden van plaatsen (bedrijfsbezoeken) (art. 5:15 Awb)
- Het vorderen van inlichtingen (art. 5:16 Awb)
- Het vorderen van inzage in zakelijke gegevens en bescheiden (art. 5:17 Awb)

Wanneer de ACM op grond van artikel 5:17 Awb inzage vordert in digitale gegevens, neemt zij daarbij de waarborgen van de ACM Werkwijze voor onderzoek in digitale gegevens 2014 in acht.[14] Deze werkwijze is ook van toepassing wanneer de ACM onderzoek doet naar algoritmische toepassingen en daarbij digitale gegevens vordert.

## 4.2 Ervaring met digitaal onderzoek

De ACM heeft ruime ervaring met het inzetten van deze onderzoeksbevoegdheden ter vergaring van digitaal bewijsmateriaal. De ACM heeft in dat verband ook ervaring met het analyseren van algoritmische toepassingen of het bestuderen van het gebruik daarvan. De ACM heeft dergelijk onderzoek in het verleden bijvoorbeeld verricht in het kader van onderzoeken naar vermoedelijke overtredingen van de mededingingswet, reguleringstoezicht, misleiding van consumenten en overtredingen van de telecommunicatiewet. Hierbij zijn onder andere onderstaande onderzoeksmethoden ingezet.

### *Administratief onderzoek*

Aan de hand van documentatie, interviews en communicatie zijn bedrijfsprocessen onderzocht waarbij ook is gecontroleerd hoe algoritmische toepassingen werken, hoe deze tot stand zijn gekomen en wie waarvoor welke verantwoordelijkheid had.

### *Simulatie en code-onderzoek*

Er is onderzoek gedaan naar de wijze waarop systemen binnen een bedrijf en/of tussen meerdere bedrijven met elkaar communiceren en welke gegevens hierbij worden uitgewisseld door middel van simulaties of broncode-onderzoek. Bij simulaties heeft de ACM het gedrag van algoritmische toepassingen onderzocht door deze te draaien in een afgezonderde en veilige (virtuele) omgeving. Bij broncode-onderzoek heeft de ACM de werking van algoritmische toepassingen die niet konden worden gesimuleerd, geanalyseerd aan de hand van de broncode.

## 4.3 Onderzoek naar rol, gedrag en werking van algoritmische toepassingen

Bij onderzoek naar algoritmische toepassingen kunnen drie belangrijke onderzoeksvragen worden onderscheiden:

1. Wat is de rol van een algoritme bij de activiteit die onderzocht wordt en welke processen zijn hierbij gevolgd (procedurele transparantie)?

[13] Bedrijven zijn op grond van artikel 5:20 Awb verplicht om mee te werken aan onderzoeken van de ACM.
[14] https://www.acm.nl/sites/default/files/old_publication/publicaties/12594_acm-werkwijze-digitaal-onderzoek-2014-02-06.pdf.

2. Wat is het gedrag van het algoritme (uitlegbaarheid)?
3. Wat is de werking van het algoritme (technische transparantie)?

De ACM zal bij onderzoeken naar activiteiten van bedrijven waarbij algoritmische toepassingen een rol spelen, niet in alle gevallen de werking of het gedrag daarvan daadwerkelijk hoeven te onderzoeken. Andere informatie, zoals (interne) correspondentie, documentatie en verklaringen, kunnen voldoende bewijs opleveren voor een overtreding. Dergelijk bewijsmateriaal zal vooral vergaard worden bij het onderzoeken van de rol van een algoritmische toepassing. Maar ook bij onderzoek naar het gedrag en de werking van een algoritmische toepassing kunnen reguliere onderzoeksmethoden, zoals het houden van interviews, vorderen van inlichtingen en zakelijke gegevens en onderzoek op locatie, geschikt zijn voor het vergaren van bewijsmateriaal.

### 4.3.1 Onderzoek naar de rol van een algoritme

Een onderzoek naar algoritmische toepassingen start vaak met de vraag wat de rol van een algoritme is bij de activiteiten die worden onderzocht en welke processen daarbij zijn doorlopen. Met dit onderzoek naar de **procedurele transparantie** van een algoritme verkrijgt de ACM een beter algemeen beeld waarvoor het algoritme is ingezet, wat de onderliggende uitgangspunten, doelen en belangen zijn, welke data hierbij gebruikt worden, welke keuzes zijn gemaakt, welke risico's zijn geïdentificeerd en hoe die worden gemitigeerd, en tot slot wie hierbij betrokken waren en wat hun specifieke rollen en verantwoordelijkheden waren.[15] Bij beantwoording van deze vraag zal ook snel duidelijk worden of het gaat om een relatief eenvoudige algoritmische toepassing (bijvoorbeeld een eenvoudig aanbevelingssysteem) of dat het gaat om een complexe toepassing waarin mogelijk wel honderden verschillende algoritmes (mogelijk deels van verschillende partijen) en/of datastromen een rol spelen. Hieronder volgen enkele vragen die relevant kunnen zijn bij het onderzoeken van de organisatorische context waarin algoritmische toepassingen worden ingezet.

#### *Doel, uitgangspunten en wijzigingen algoritme*

- Welke doelen wil het bedrijf bereiken met het algoritme?
- Welke uitgangspunten, belangen en voorkeuren liggen ten grondslag aan het ontwerp en/of de implementatie van het algoritme in de organisatie? Welke keuzes zijn hierbij gemaakt?
- Waarom is specifiek voor deze oplossing gekozen? Waren er ook alternatieven en zo ja, welke en waarom is niet voor een van die alternatieven gekozen?
- Welke gegevens zijn gebruikt voor de training van het algoritme? Waar komen die gegevens vandaan? Hetzelfde geldt voor de gegevens die als inputdata dienen.
- Hoe is het algoritme getest? Wat waren de uitkomsten? Heeft dit geleidt tot wijzigingen? Waarom wel of niet?
- Welke wijzigingen in het algoritme hebben zich over tijd voorgedaan? Waarom hebben wijzigingen plaatsgevonden? Is dit vastgelegd?

#### *Rol bij bedrijfsprocessen en activiteiten*

- Voor welke bedrijfsprocessen wordt het algoritme toegepast?
- Welke activiteiten of besluiten worden ondersteund door de inzet van het algoritme? Welke rol speelt het algoritme daarin?

#### *Risico-identificatie en -mitigatie*

- Welke methode/methodes is /zijn gehanteerd om risico's te identificeren?
- Welke risico's zijn er geïdentificeerd?
- Welke maatregelen zijn er genomen om deze risico's te mitigeren?

[15] De term 'procedurele transparantie' is ontleend aan de concept inkoopvoorwaarden voor eerlijke en transparante algortimes die de Gemeente Amsterdam in samenwerking met Pels Rijcken en KPMG heeft opgesteld. Zie: https://www.amsterdam.nl/wonen-leefomgeving/innovatie/digitale-stad/grip-op-algoritmes/ (laatst bezocht op 10 april 2020).

### *Rol en verantwoordelijkheden actoren*[16]

- Besluitvormers: Wie is betrokken bij en/of verantwoordelijk voor het vaststellen van het doel en de uitgangspunten (inclusief foutmarges) van het ontwerp en/of de implementatie van het algoritme in de organisatie?
- Ontwikkelaars: Wie zijn er betrokken bij de ontwikkeling van het algoritme, de implementatie en het onderhoud?
- Gebruikers: Wie in de organisatie maken er gebruik van het algoritme bij bedrijfsprocessen en activiteiten en wat is hun rol? Wat is de rolverdeling tussen de algoritmische toepassing en gebruikers bij beslissingen die effect hebben op consumenten en marktpartijen? In hoeverre is er menselijke tussenkomst bij deze beslissingen?
- Controleurs: Wie zijn er betrokken bij de normering van de ontwikkeling en inzet van algoritmische toepassingen en de naleving hiervan binnen17 en buiten de organisatie? Welke beslissingen hebben zij genomen en waarom?
- Externe partijen: Zijn externe partijen betrokken bij de ontwikkeling en/of implementatie van een algoritme? Wat is hun rol en welke afspraken zijn er gemaakt?

## 4.3.2 Onderzoek naar de werking en gedrag van een algoritme

### *Transparantie*

De mate van transparantie van een algoritmische toepassing is relevant voor het vaststellen van de werking en het gedrag van een algoritme. Dit hangt veelal af van de complexiteit van het algoritme. Naarmate de complexiteit toeneemt, wordt het redeneerproces van een algoritme veelal ook ondoorzichtiger.[18] In de literatuur is er veel discussie over wat precies moet worden verstaan onder ‘transparantie’ van algoritmes, maar er zijn op dit moment grofweg twee vormen van transparantie te onderscheiden: **technische transparantie** (werking) en **uitlegbaarheid** (gedrag).[19]

Bij technische transparantie gaat het om hoe het algoritme werkt. Daarvoor is onder andere informatie relevant over de werkwijze van een algoritme, de broncode, de (technische) documentatie, de gebruikte trainingsdata, en de toegepaste variabelen, parameters en drempelwaarden.

Bij uitlegbaarheid gaat het om hoe het algoritme zich gedraagt. Daarvoor is onder andere informatie relevant waarmee de output van een algoritme verklaard kan worden. Een verder onderscheid dat hierbij gemaakt kan worden is of de uitleg betrekking heeft op het algoritme danwel het toegepaste model in algemene zin ongeacht de input (*model-centric*), of betrekking heeft op een specifieke input-output relatie (*subject centric*).[20] Een voorbeeld van deze laatste benadering is bijvoorbeeld informatie over welke wijzigingen in de inputdata leiden tot een andere output of beslissing.

### *Onderzoek naar werking en gedrag*

Bij eenvoudige algoritmes is het gedrag veelal te verklaren aan de hand van de werking van het algoritme. Het kan onder omstandigheden nuttig zijn om “onder de motorkap” te kijken en de broncode zelf te analyseren. Er zitten echter wel beperkingen aan deze onderzoeksmethode. Allereerst is men voor effectieve codereview afhankelijk van de aanwezigheid van goede documentatie. Bij afwezigheid daarvan is codereview al snel tijdrovend en kostbaar. Dit geldt ook voor complexe code. Daarnaast is codereview bij complexe algoritmes minder snel van toegevoegde waarde, omdat het weinig inzicht geeft in het gedrag van het algoritme. Veel ontwikkelaars gebruiken broncode die door anderen zijn ontwikkeld en gedeeld

[16] De typologie ‘besluitvormers, ontwikkelaars en gebruikers’ is overgenomen uit: M. Wieringa, ‘What to account for when accounting for algorithms: A systematic literature review on algorithhmic accountability’, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 2020, p1-18.
[17] Binnen organisaties kan men denken aan bedrijfsjuristen, compliance-medewerkers, de privacy officer of ethicus.
[18] Zie ook de Richtlijnen voor het toepassen van algoritmes door overheden van het Ministerie van Justitie en Veiligheid.
[19] Zie o.a de twee rappporten die zijn opgesteld in opdracht van het European Parliamentary Research Service Panel for the Future of Science and Technology: R. Koene A. Clifton C. Hatada Y. Webb H. Patel M. Machado C. LaViolette J. Richardson and D. Reisman, ‘A governance framework for algorithmic accountability and transparency’, 2019; en C. Castelluccia, D. Le Métayer, ‘Understanding algorithmic decisionmaking: Opportunities and challenges’, 2019.
[20] L. Edwards, M. Veale, ‘Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For’, 16 Duke Law & Technology Review 18 (2017).

(bijvoorbeeld via platforms als Github). Met een zoekopdracht op basis van specifieke broncode kan in deze gevallen de originele bron gevonden worden waarin informatie over de werking te vinden is. Naast het achterhalen van de werking van een algoritme, kan codereview inzicht geven in de intenties van de ontwikkelaar(s). Zo kan het bijvoorbeeld inzicht geven in de wijzigingen (transformaties) die op de inputdata worden uitgevoerd voordat deze langs het algoritme gaan.

Er zijn ook methodes om het gedrag en de onderliggende logica en (causale) verbanden tussen de input en output beter in kaart te brengen zonder code review. Hiervoor zijn verschillende onderzoeksmethodes beschikbaar, van traditionele statistische en econometrische methodes zoals regressie-analyse tot meer complexe methoden waarbij het gedrag van het algoritme wordt benaderd met behulp van algoritme(s).[21]

### *Input-output analyse: historische data of zelf gedefinieerde data*

De werking van een algoritme kan onder meer onderzocht worden door input-output analyses. Bij dit type onderzoek kan de ACM werken met bestaande historische input- en outputdata of met zelf gedefinieerde input. Een voordeel van zelf gedefinieerde input is dat de ACM dan meer gecontroleerd kan onderzoeken wat verschillen in inputdata voor invloed hebben op de output. Bij zelflerende algoritmes dient de ACM er echter rekening mee te houden dat de inputdata zelf, het model en uiteindelijke gedrag kunnen aanpassen. Dit kan gevolgen hebben voor de repliceerbaarheid van een bepaalde uitkomst. Het is dus van belang om er rekening mee te houden dat het model zelf kan wijzigen en dat hier zicht op is, bijvoorbeeld door middel van vastlegging van de wijzigingen.

### *Input-output analyse: live omgeving of gecontroleerde omgeving*

Als men werkt met zelf-gedefinieerde input kan men ervoor kiezen om deze input te testen in een live-omgeving, dat wil zeggen in de context waarin het algoritme in de praktijk functioneert. Door (toegang tot) de bijbehorende output te vorderen kan het gedrag van het algoritme in de praktijk onderzocht worden. Een andere optie is om de zelf-gedefinieerde input te testen met behulp van een kopie van het algoritme dat draait in een gecontroleerde omgeving (sandboxing). In die situatie heeft de ACM meer controle over de werking van het algoritme. Dit kan vooral bij zelf-lerende algoritmes relevant zijn. Een nadeel van sandboxing is dat het vertoonde gedrag in meer of mindere mate kan afwijken van het gedrag van het algoritme in de praktijk. Dit kan gevolgen hebben voor de bewijswaarde van uitkomsten van analyses in een gecontroleerde omgeving. Bij analyses in een gecontroleerde omgeving is het dus goed om in kaart te brengen of en zo ja, welke afwijkingen er kunnen optreden ten opzichte van een live-omgeving.

### *Explainable AI*

Zoals eerder aangegeven, zijn er ook methoden waarbij algoritmes worden ingezet om het gedrag van algoritmes (bij benadering) te verklaren. Een bekend voorbeeld van zo'n methode is LIME.[22] Dergelijke oplossingen kunnen worden ingezet om het gedrag van complexe algoritmes, zoals algoritmes die gebruik maken van neurale netwerken, beter in kaart te brengen. Andere voorbeelden van technische methoden om het gedrag meer inzichtelijk te maken zijn Anchors[23], Sunlight[24] en TREPAN.[25]

Bij de keuze voor dergelijke oplossingen is het van belang om stil te staan bij wat men wil onderzoeken en goed te controleren of de gekozen methode hiervoor wel geschikt is. Zo zijn sommige methoden beter geschikt voor het verklaren van het gedrag van een algoritme bij specifieke input, terwijl andere methoden meer geschikt zijn om een algemeen beeld te krijgen van het gedrag ongeacht de input. Er is overigens discussie over de betrouwbaarheid van dergelijke methodes ('explainable AI') om het gedrag van

[21] Voor een overzicht van verschillende technische methodes zie: A. Adadi, M. Berrada, 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)', IEEE Access 6 (2018).
[22] M. Tulio Ribeiro, S. Singh, C. Guestrin, '"Why Should I Trust You?": Explaining the Predictions of Any Classifier', http://arxiv,org/abs/1602.04938.
[23] M. Tulio Ribeiro, S. Singh, C. Guestrin, 'Anchors: High-Precision Model-Agnostic Explanations', AAAI 2018: 1527-15355. https://homes.cs.washington.edu/~marcotcr/aaai18.pdf.
[24] M Lecuyer e.a., 'Sunlight: Fine-grained Targeting Detection at Scale with Statistical Confidence', CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, October 2015 p. 554–566. http://www.cs.columbia.edu/~djhsu/papers/sunlight.pdf.
[25] Deze voorbeelden zijn ontleend aan C. Castelluccia, D. Le Métayer, 'Understanding algorithmic decisionmaking: Opportunities and challenges', 2019, p. 48-50.

algoritmes nauwkeurig te verklaren.[26] Ze vormen immers slechts een statistische interpretatie c.q. benadering van wat het onderliggende model doet.

### *Relevante informatie voor onderzoek naar werking en gedrag*

Bij onderzoek naar de werking en/of het gedrag van algoritmes kan o.a. de volgende informatie relevant zijn:

- De relevante (technische) documentatie over de werking en onderliggende uitgangspunten van het algoritme, inclusief interne gebruikershandleidingen;
- de broncode, inclusief oudere en alternatieve versies (die beschikbaar zijn via het gebruikte versie-controle systeem zoals Git);
- de gebruikte inputdata en/of trainingsdata;
- de toegepaste variabelen, parameters en drempelwaarden;
- logfiles en andere diagnostische informatie zoals test- en debuggingrapporten, en documentatie over aangebrachte wijzigingen; en
- scrumborden, communicatie binnen samenwerkingsomgevingen die ontwikkelaars gebruiken (zoals Slack, Mattermost en Github) en andere communicatiekanalen (bijvoorbeeld e-mail en chat apps).

## 4.4 Uitdagingen bij onderzoek naar algoritmes

Hieronder worden enkele uitdagingen besproken waar de ACM mee te maken zou kunnen krijgen bij concrete onderzoeken naar algoritmische toepassingen. Ook de bedrijven die algoritmische toepassingen gebruiken kunnen te maken krijgen met deze uitdagingen in het kader van hun verantwoording naar interne en externe stakeholders.

### 4.4.1 Vluchtigheid

Zelflerende algoritmes en/of de gegevens die een rol spelen bij het trainen en het functioneren daarvan kunnen vluchtig zijn. Zelflerende algoritmes passen zichzelf aan en kunnen dus van het ene op het andere moment zich anders gaan gedragen. Maar ook de context waarin zij opereren is vluchtig, zowel in termen van input voor het systeem in werking, als het gedrag dat gekoppeld wordt aan de uitkomst van het algoritme. De hoeveelheid gegevens kan dusdanig omvangrijk zijn dat (langdurige) opslag moeilijk of erg kostbaar is. Ook kan het bijvoorbeeld gaan om persoonsgegevens die wel geaggregeerd worden, maar om privacyredenen niet bewaard (mogen) worden. Als de ACM onderzoek doet naar een structurele overtreding die in het verleden is begaan of is begonnen, kan deze vluchtigheid van (zelflerende) algoritmes en/of de hieraan gerelateerde gegevens een uitdaging zijn.

### 4.4.2 Externe partijen / ketenproblematiek

Algoritmische toepassingen werken niet in isolement en zijn binnen een onderneming onderdeel van een ICT-omgeving met koppelingen naar (deel)systemen en datasets die toe kunnen behoren aan of vallen onder de verantwoordelijkheid van andere onderdelen van de onderneming of externe partijen. Juist waar het gaat om toepassingen die gebruik maken van complexe algoritmes, maken bedrijven regelmatig gebruik van algoritmische toepassingen van derden. Deze toepassingen kunnen dan – als reguliere software – lokaal draaien of worden op afstand aangesproken (cloudtoepassingen).

Het gebruik van algoritmische toepassingen van externe partijen betekent niet dat de ACM hier geen onderzoek naar kan doen. Ook derden zijn verplicht om mee te werken aan een vordering van de ACM om inlichtingen of inzage in zakelijke gegevens en bescheiden, waarbij dit wel proportioneel dient te zijn. Wanneer er meerdere derde partijen betrokken zijn bij de algoritmische toepassingen kan dit het onderzoek van de ACM bemoeilijken. Het uitgangspunt blijft echter dat de onderneming die ten behoeve van zijn

[26] C. Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', Nature Machine Intelligence 1, 206–215 (2019).

bedrijfsactiviteiten de diensten en producten van externen afneemt, zelf verantwoordelijk blijft voor zijn eigen bedrijfsactiviteiten.

### 4.4.3 Grensoverschrijdende aspecten

De ACM wordt bij onderzoeken met enige regelmaat geconfronteerd met gegevens die in het buitenland staan opgeslagen bij een externe partij of waarbij de onderzochte onderneming gebruik maakt van diensten die worden aangeboden door een externe partij uit het buitenland. De ACM gaat uit van het standpunt dat gegevens gekopieerd kunnen worden in een onderzoek als aan één van de volgende drie zaken, of een combinatie daarvan, is voldaan. De onderzochte onderneming:

1. Is eigenaar van de data;
2. beheert de data; en/of
3. is gebruiker van de data.

De praktijk wijst uit dat dit –afhankelijk van de situatie- goed werkt voor gegevens. Dit uitgangspunt kan ook worden toegepast bij onderzoek naar algoritmische toepassingen. Wanneer de algoritmische toepassing in zijn geheel als bestanden of als (virtueel) systeem kan worden gekopieerd, dan is dit te vergelijken met het kopiëren van een gegevensverzameling.

### 4.4.4 Privacy en noodzakelijke data voor onderzoek

Het is algemeen bekend dat er algoritmische toepassingen zijn waarbij grote hoeveelheden persoonsgegevens worden verwerkt. Voor onderzoek naar de werking en het gedrag van dergelijke algoritmische toepassingen kan het noodzakelijk zijn dat de ACM over (een deel van) deze persoonsgegevens kan beschikken. Dit kan betekenen dat de ACM de beschikking krijgt over grote hoeveelheden persoonsgegevens waar de nodige risico’s aan kleven. De ACM zal in dat geval als verwerkingsverantwoordelijke moeten voldoen aan de vereisten uit de AVG. Daarom dient de ACM vooraf in kaart te brengen of er bij een onderzoek naar algoritmische toepassingen persoonsgegevens verwerkt zullen worden en zo ja, welke maatregelen de ACM in dat concrete geval dient te treffen om dit op een rechtmatige wijze met voldoende waarborgen te doen.

### 4.4.5 ICT onderzoeksinfrastructuur ACM

Algoritmische toepassingen kunnen beperkt zijn tot een eenvoudige infrastructuur en gebruik maken van een of meerdere computersystemen van een onderneming. Het is dan waarschijnlijk mogelijk om een dergelijke omgeving in kopie mee te nemen en deze op een beschermde en afgeschermde wijze bij de ACM te activeren en te onderzoeken. Het gebruik door ondernemingen van clouddiensten bij derden, waaronder ook het gebruik van algoritmische toepassingen van derden, kan tot gevolg hebben dat de ACM een algoritmische toepassing alleen kan onderzoeken wanneer de ACM ook dezelfde clouddiensten van deze derden gebruikt. De ACM zou in dat geval dus moeten afwijken van de normale procedure waarin digitaal onderzoek wordt verricht binnen een afgeschermde omgeving.