Rapportage AI- & Algoritmerisico's Nederland

Editie 3, zomer 2024

Autoriteit Persoonsgegevens | Directie Coördinatie Algoritmes (DCA)

Periodiek inzicht in risico's en effecten van
de inzet van AI & algoritmes in Nederland

AUTORITEIT
PERSOONSGEGEVENS

Inhoudsopgave

Kernboodschappen

1. Het AI-risicobeeld vraagt onverminderd om waakzaamheid bij iedereéén – van ministers tot burgers en van CEO's tot consumenten – want (i) het is moeilijk in te schatten of AI-toepassingen voldoende beheerst zijn en (ii) AI-incidenten kunnen zich steeds vaker voordoen, zeker omdat AI zich steeds dieper verweeft in de samenleving.

Het blijven stormachtige tijden. Dat is ook begrijpelijk bij de opkomst van een nieuwe systeemtechnologie. Een jaar geleden was de observatie dat Nederland stappen moest zetten om grip te krijgen op algoritmes. Inmiddels gaat de onstuimige groei van AI-technologie verder. Daarbij geeft de opkomst van generatieve AI een prikkel om op grote schaal te experimenteren met nieuwe AI-toepassingen. De komende jaren zal AI steeds dieper verweven raken met elementen van de samenleving. Dit resulteert, zowel qua omvang als aard, in meer en nieuwe risico's die we nog moeilijk op waarde kunnen schatten. Ook de langetermijneffecten zijn nog niet volledig te overzien. Op hoofdlijnen is de internationale beleidsrespons tot nu toe daadkrachtig. Deze richt zich zowel op klassiek toezicht als op nieuwe vormen van testen en beheersen van bijvoorbeeld de veiligheid van AI-systemen, bijvoorbeeld in het bestrijden van nieuwe risico's op het gebied van cyberveiligheid. Tegelijkertijd is het opzetten van AI-regulering een proces van lange adem. Dit betekent dat organisaties – en de samenleving als geheel – zich onverminderd moeten blijven voorbereiden op AI-incidenten, zonder deze op voorhand te accepteren. En zolang organisaties nog twijfels hebben over hun risicobeheersing, vraagt dit om terughoudendheid bij de inzet van AI. Dat gaat zowel om systemen gebaseerd op simpele (statische) algoritmes als om complexe (zelflerende) AI, die de AP beide onder de noemer 'AI-systeem' schaart.

2. Veel nieuwe AI-systemen en (mogelijke) risico's springen in het oog. Van experimenteerdrift door bigtechbedrijven tot het volop inzetten van AI in situaties waarin mensen kwetsbaar zijn.

Het valt op dat bigtechbedrijven nieuwe toepassingen gebaseerd op generatieve AI zo snel mogelijk op de markt willen brengen. Dit is de afgelopen periode vaak gepaard gegaan met missers en kwetsbaarheden die leidden tot noodreparatie of zelfs het terugtrekken van systemen. Daarbij kunnen grondrechten zoals non-discriminatie en privacy in het geding zijn en spelen er rechtmatigheidsvraagstukken in relatie tot bestaande regelgeving zoals de bijvoorbeeld de AVG, auteursrecht en consumentenbescherming. Ondertussen verkennen Nederlandse organisaties volop de verdere mogelijkheden van klassiekere AI-systemen, bijvoorbeeld voor gedragsmonitoring via camera-analyse, aansturing van werknemers, risicoselectie in het sociale domein en advisering over het geschikte onderwijsniveau voor kinderen. Belangrijk is dat informatie over deze systemen steeds meer beschikbaar komt via publieke algoritmeregistratie. Dit is de basis voor proactieve transparantie en toezicht. Zie verder hoofdstuk 1.

3. Informatievoorziening is essentieel voor het functioneren van de democratie, maar staat onder druk door de inzet van AI-systemen. Dit geldt zowel voor moderatie en distributie van content als, meer recentelijk, voor contentcreatie met generatieve AI.

De inzet van AI-systemen beïnvloedt op grote schaal de online informatievoorziening. Generatieve AI maakt het voor kwaadwillenden mogelijk om op grote schaal desinformatie te genereren. Daarbij heeft generatieve AI inherente technologische zwakheden, die ook bijdragen aan misinformatie en discriminerende en stereotyperende content. Des- en misinformatie hebben een grote invloed op het publieke debat. Nederlanders maken zich hier grote

zorgen over. Het verifiëren van de herkomst en 'echtheid' van content is daarom een kritieke schakel in zowel het kunnen vertrouwen van content als het kunnen omgaan met de effecten ervan. Het informatieaanbod is simpelweg te overweldigend, wat het gebruik van filtering in het aanbod nodig maakt. Deze moderatie is voor een groot deel in handen van bigtechplatforms. Dat kan een divers informatieaanbod in gevaar brengen. De Europese Digital Services Act legt zeer grote online platforms onder andere op om openheid te geven over moderatie en om desinformatie aan te pakken. Daarbij valt op dat het aantal Nederlandstalige moderatoren bij deze platforms afneemt. Omdat de inzet van (generatieve) AI in de online informatievoorziening het publieke debat op grote schaal beïnvloedt, is een gemeenschappelijke 'informatiebasis' nodig om polarisatie tegen te gaan. In welke mate het functioneren van de democratie op dit moment (of in de toekomst) daadwerkelijk wordt geraakt door de rol van AI-systemen is moeilijk meetbaar, waardoor het belangrijk is om via actieve monitoring en analyse nadrukkelijk een vinger aan de pols te houden. Zie verder hoofdstuk 2.

4. Voorwaarden voor adequate democratische controle van AI-systemen zijn op dit moment onvoldoende ingevuld.

De vormgeving van het proces voor democratische sturing en controle van AI-systemen is bepalend voor de wijze waarop volksvertegenwoordigers – van Tweede Kamer tot gemeenteraad – grip kunnen hebben op AI-systemen die worden ingezet door de overheid. Deze sturing en controle moet mogelijk zijn tijdens elke fase van ontwikkeling, inzet en evaluatie van een AI-systeem. Deze rapportage verkent dit onderwerp aan de hand van de situatie in het lokaal bestuur. Binnen de publieke sector gebruiken decentrale overheden de meeste AI-systemen. Democratische controle van publieke AI-systemen gebeurt door de volksvertegenwoordigers, samen met de rekenkamer, de ombudspersoon en de media. Deze instanties hebben echter beperkte capaciteit en expertise tot hun beschikking. Dit bemoeilijkt hun controlerende rol. Enquêteresultaten tonen dat gemeentelijke organisaties nog maar beperkt overzicht hebben over hun AI-systemen, dat raadsleden twijfelen over de adequaatheid van hun AI-kennis en dat slechts enkele lokale rekenkamers sporadisch onderzoek doen naar AI-systemen. Landelijk zijn investeringen wenselijk in een ondersteunende infrastructuur voor landelijke én lokale actoren, bijvoorbeeld via een AI-coördinatiecentrum of via AI-expertisecentra. Zowel om de beheersing van de uitvoerende en controlerende taak op AI-terrein te versterken.
Zie verder hoofdstuk 3.

5. De aselecte steekproef is een waardevol instrument om de risico's bij profilerende en selecterende AI-systemen te verminderen.

Veel organisaties gebruiken algoritmes voor risicoprofilering of soortgelijke processen waarbij onderscheid tussen mensen wordt gemaakt. Dit brengt grondrechtenrisico's met zich mee. Deze rapportage verkent dit onderwerp aan de hand van voorbeelden op het gebied van fraudedetectie. Belangrijk is deze algoritmes altijd als onderdeel van een breder proces te zien. Vrijwel iedereen wordt op verschillende plekken in de samenleving onderworpen aan deze fraudedetectiealgoritmes. Naast rechtmatigheidsvraagstukken– mag een dergelijk algoritme in bepaalde situatie ingezet worden en mogen bepaalde indicatoren gebruikt worden – is het een essentieel aandachtspunt dat fouten in deze algoritmes grote gevolgen hebben. Discriminatie en een overmatig vertrouwen in het fraudealgoritme zijn twee belangrijke risico's. Om deze tegen te gaan, kan het inbedden van een aselecte steekproef in het fraudedetectieproces helpen om discriminatierisico's te monitoren. Ook draagt de aselecte steekproef bij aan het meten van efficiëntie en het verkennen van nieuwe soorten fraude. De inrichting en werking zullen per context verschillen, maar in veel gevallen is het een maatregel die het waard is om te overwegen bij gebruik van een AI-systeem voor profilerende en selecterende AI-systemen. Zie verder hoofdstuk 4.

6. De inwerkingtreding van de AI-verordening (begin augustus 2024) is een mijlpaal, met als punt van zorg (i) de lange overgangstermijn (tot en met 2030) voor bestaande AI-systemen met een hoog risico binnen de overheid en (ii) de vraag of stevige en werkbare productstandaarden tijdig gereed zijn.

Sommige bepalingen onder de AI-verordening treden al begin 2025 in werking, bijvoorbeeld voor verboden AI-toepassingen en AI-geletterdheid binnen organisaties. Hier is dus werk aan de winkel, waarbij opgemerkt dat AI-toepassingen

die straks worden verboden onder de AI-verordening ook nu al in strijd kunnen zijn met andere wetgeving, bijvoorbeeld de AVG. De AP benadrukt dat de productstandaarden onder hoge tijdsdruk moeten worden afgerond. Tijdigheid is daarbij van het grootste belang, maar mag niet ten koste gaan van de inhoud. De productstandaarden zijn allesbepalend voor de daadwerkelijke effectiviteit en uitvoerbaarheid van de AI-verordening. Ondertussen werken toezichthouders in Nederland aan de voorbereiding op nieuwe toezichtstaken onder de AI-verordening. Dit heeft mede geleid tot eerste adviezen aan het kabinet. Zie verder hoofdstuk 5.

7. Bij de verdere uitwerking van het regeerprogramma adviseert de AP om onverminderd prioriteit te geven aan algoritmeregistratie door overheidsorganisaties en na te denken over registratie door semi-publieke organisaties.

Het hoofdlijnakkoord bevat belangrijke bepalingen over algoritmes en AI die het huidige beleid kunnen versterken. Bijvoorbeeld over het gebruik van een wetenschappelijke standaard voor het gebruik van modellen en algoritmes. Het is belangrijk deze eisen als onderdeel te zien van de bepalingen uit de AI-verordening en de verdere uitwerking in productstandaarden. Bijvoorbeeld om wildgroei in standaarden te voorkomen (zie volgend punt). Op korte termijn is het belangrijk dat algoritmeregistratie onverminderd een prioriteit blijft. De AP blijft er voorstander van dat het snel verplicht wordt voor overheidsorganisaties om algoritmes te registeren. Ook benadrukt de AP dat de scope van zo'n register breed genoeg moet zijn en dat het in samenhang moet worden gezien met de Europese registratieplicht voor hoog-risico AI-systemen onder de AI-verordening.
De afweging is niet altijd simpel of een bepaald AI-systeem (of algoritme) al dan niet een hoogrisicosysteem is. Aandachtspunt is het gebruik van AI-systemen door organisaties in bijvoorbeeld de gezondheidszorg, het onderwijs, de volkshuisvesting en het openbaar vervoer. Het gaat hierbij om essentiële dienstverlening, maar het zicht op het gebruik van AI-systemen in deze sector is troebel. Zie verder hoofdstuk 5.

8. De AP spant zich in voor het vergroten van de beheersing van AI-systemen, waarbij (i) een wildgroei aan kaders moet worden voorkomen en (ii) een herijking van de nationale AI-strategie kan bijdragen aan het verdere ecosysteem voor ontwikkeling en controle van AI-systemen.

Er zijn diverse beleidsontwikkelingen in binnen- en buitenland die bijdragen aan het in kaart brengen en verminderen van AI-gerelateerde risico's. Bijvoorbeeld de oprichting van AI-veiligheidsinstituten, samenwerking tussen toezichthouders en beleidsmakers en het opzetten van AI-adviesraden. Dit kan bijdragen aan tijdig en geharmoniseerd handelen als nieuwe risico's zich voordoen. En het geeft houvast voor verantwoorde ontwikkeling en inzet van AI-systemen. Tegelijkertijd bestaat het risico dat er een overvloed is aan initiatieven en kaders. Het levert verwarring op of kan zelfs schadelijk zijn – bijvoorbeeld via (onbedoelde) ethics washing – als kaders niet concreet genoeg of op meerdere manieren te interpreteren zijn. In aanvulling op rechtmatigheidsvraagstukken spant de AP zich vanuit de coördinerende taak in om houvast te geven in de beheersing van AI-systemen. Juist om kansrijke en verantwoorde toepassingen te ondersteunen. Hierbij is specifiek aandacht voor de aankomende AI-verordening. Het kabinet kan een bijdrage leveren aan versterking van het gehele AI-ecosysteem door de nationale AI-strategie uit 2019 te herijken, met behoud van het goede en tegelijkertijd aandacht voor de nieuwe uitdagingen die de complexere, moderne AI-systemen met zich meebrengen. Zie verder hoofdstuk 5.

AI-systeem als brede definitie

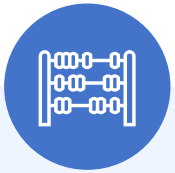
In de afgelopen periode is consensus ontstaan over de betekenis van de term 'AI-systeem'. De term AI-systeem is opgenomen in de AI-verordening en is gebaseerd op de mondiale aanbeveling vanuit de OECD. Beknopt geformuleerd kan een AI-systeem– voor expliciete of impliciete doelstellingen –uit ontvangen input afleiden hoe output te genereren. Zoals voorspellingen doen, inhoud genereren, aanbevelingen doen of beslissingen nemen. Deze output heeft daarbij invloed op de fysieke of virtuele omgeving. AI-systemen variëren in hun mate van autonomie en aanpassingsvermogen bij inzet (na de ontwikkelingsfase).

Op basis van huidige inschattingen heeft de term AI-systeem betrekking op een brede waaier aan toepassingen, van simpele (statische) algoritmes tot complexe (zelflerende) AI. In een recente toelichting bij de herkalibratie van de definitie legt de OECD bijvoorbeeld uit dat het model dat de basis vormt van een AI-systeem (dit kunnen ook meerdere modellen zijn), zowel (eenmalig) handmatig gebouwd kan zijn door menselijke programmeurs als automatisch tot stand gekomen, bijvoorbeeld door niet-begeleide, begeleide of zelfversterkende machine learning-technieken. De OECD benadrukt in de toelichting dat modelaanpassingen vaak onderdeel zijn van de ontwikkelingsfase en dat het model meestal gefixeerd is tijdens de inzet.[1] Verschillende AI-systemen variëren in hun niveau van autonomie en aanpassingsvermogen bij de inzet. Traditionele en eenvoudige softwareontwikkeling, waarbij mensen volledig hebben bepaald welke regels er worden toegepast, valt buiten de reikwijdte van de term.

De term AI-systeem omvat zowel toepassingen die 'slechts' aanbevelingen doen aan mensen als toepassingen die zelfstandig ingrijpende beslissingen kunnen nemen. Het zijn vaak de context en de inbedding in een bredere taak of doelstelling die bepalend zijn voor de vorm van de output van een AI-systeem. Een voorbeeld is een ondersteuningssysteem voor autobestuurders. Zo'n systeem kan 'voorspellen' dat er een andere auto nabij is, om vervolgens de bestuurder 'aan te bevelen' en te waarschuwen om te remmen. Maar het systeem kan ook zo ingezet worden dat het zelfstandig 'beslist' om te remmen. In beide gevallen is sprake van een AI-systeem. De kennis die nodig is om verantwoord om te gaan met deze systemen, bewust te zijn van de risico's en effecten ervan en om verantwoord besluiten te nemen over de inzet hiervan, noemen we 'AI-geletterdheid'.

Richtsnoeren van de Europese Commissie zullen verdere verduidelijking bieden. Er komen richtsnoeren die zullen ingaan op de praktische toepassing van de definitie van een AI-systeem. Dit zal zorgen voor concretisering, bijvoorbeeld via voorbeelden.

Overzicht Risicobeeld AI & Algoritmes zomer 2024



Welke AI-systemen zijn opgevallen?

- **Tech-innovaties.** Onstuimige en incidentrijke lanceringen van generatieve AI-systemen door bigtech-bedrijven.
- **Gedragsmonitoring.** Gericht op klanten en bezoekers via camera's in supermarkten, sportscholen en het openbaar vervoer.
- **Arbeidsaansturing.** Een systeem voor de aansturing van wegbeheerders bij Rijkswaterstaat.
- **Wonen.** Inzet van AI en scraping voor detectie van woonfraude door woningcorporaties.
- **Onderwijstoetsing.** De doorstroomtoets in het primair onderwijs, met belangrijke rol voor adaptieve toetsing.
- **Publieke diensten.** Een systeem voor het filteren van bijstandsaanvragen via machine learning.

Welke risico's springen in het oog?

- **'Rat race' in tech.** Bigtechbedrijven streven naar snelle marktdominantie in AI. Kwaliteitsstandaarden en risico-beheersing staan onder druk.
- **Misbruik van generatieve AI.** De technologie is als wondermiddel voor kwaadwillenden. Risico's zoals voor cybersecurity nemen sterk toe.
- **Informatievoorziening onder druk.** De inzet van AI-systemen in de productie, moderatie en consumptie van online informatie heeft invloed op diversiteit en betrouwbaarheid – mogelijk met grote invloed op het publieke debat.
- **Democratische controle op AI.** Overheden zijn onvoldoende uitgerust om de inzet van AI-systemen in de publieke sector te controleren. Incidenten worden daardoor mogelijk te laat (of niet) opgemerkt.
- **Discriminatie bij selecterende AI.** Profilerende en selecterende systemen, bijvoorbeeld voor fraudedetectie, liggen nog altijd onder een vergrootglas en het detecteren van discriminatie is vaak moeilijk.
- **Tijdigheid van detailregelgeving.** Het is de vraag of heldere en solide productstandaarden onder de AI-verordening op tijd komen.
- **Lange overgangstermijnen.** Er is een overgangstermijn tot 2030 voor bestaande AI-systemen met een hoog risico bij de overheid.
- **Door de bomen het bos niet zien.** Er is een wildgroei aan kaders en standaarden.

Wat moet er gebeuren?

- **Europese aanpak generatieve AI.** Doorpakken met standaarden voor generatieve AI en streven naar mondiale convergentie.
- **AI-veiligheidsinstituut.** Verkennen of en hoe dit opgezet kan worden, in aansluiting op bestaande (toezicht)taken.
- **Bezint eer gij begint.** In brede zin onzorgvuldige experimenteerdrift tegengaan. Verantwoorde inzet van AI vraagt om zorgvuldigheid.
- **AI-geletterdheid.** Relevant voor iedere werknemer en burger. De basis om AI-systemen te begrijpen en bewust te zijn van de effecten en risico's ervan.
- **Traceerbaarheid van informatie.** In een wereld met AI moet de herkomst van informatie bekend zijn om deze te kunnen verifiëren.
- **Investeren in democratische controle.** Ondersteuning van kennis, capaciteit en processen bij lokaal bestuur, zodat zij AI-systemen verantwoord inzetten en deze ook gecontroleerd kunnen worden.
- **Aselecte steekproeven.** Een bruikbare validatie-tool voor selecterende systemen die ingebed kunnen worden in het werkproces.
- **Verplichte algoritmeregistratie.** Als prioriteit behouden voor overheidsorganisaties en verplicht stellen, met mogelijke uitbreiding naar de niet-commerciële dienstverleningssector.
- **AI-strategie.** De onstuimige ontwikkeling van AI vraagt om herijking van de nationale AI-strategie.

1. Overkoepelende ontwikkelingen

SNEL NAAR DIT ONDERDEEL

1.1 Risicobeeld

De beheersing van de risico's van AI-systemen gaat niet in hetzelfde tempo als de ontwikkeling en inzet van de technologie. Dat is een realiteit die beleidsmakers, bestuurders en de samenleving onder ogen moeten zien. Dat betekent niet dat we alle toekomstige incidenten op voorhand moeten accepteren, maar wel dat we ons erop moeten voorbereiden.

Nieuwe technologie is het best bij te sturen als deze nog in de kinderschoenen staat. Zodra de technologie verder is ontwikkeld en al volop wordt ingezet, is het veel ingewikkelder om tot passende risicobeheersing te komen.
Net zoals het vooraf verplichten van veiligheids-voorzieningen in auto's, zoals een airbag, meer effect heeft dan dit achteraf te doen, als iedereen al een auto zonder airbag heeft. Dan vraagt het onevenredig veel tijd en kosten om bij te sturen én is de situatie tussentijds onveilige situatie.

Deze technologie is zo allesoverstijgend dat er op zijn minst binnen Europa, maar het liefst wereldwijd, consensus moet zijn over risicobeheersing. Zonder heldere principes, regelgeving, standaarden en maatschappelijke normen zal risicobeheersing steeds moeilijker worden.

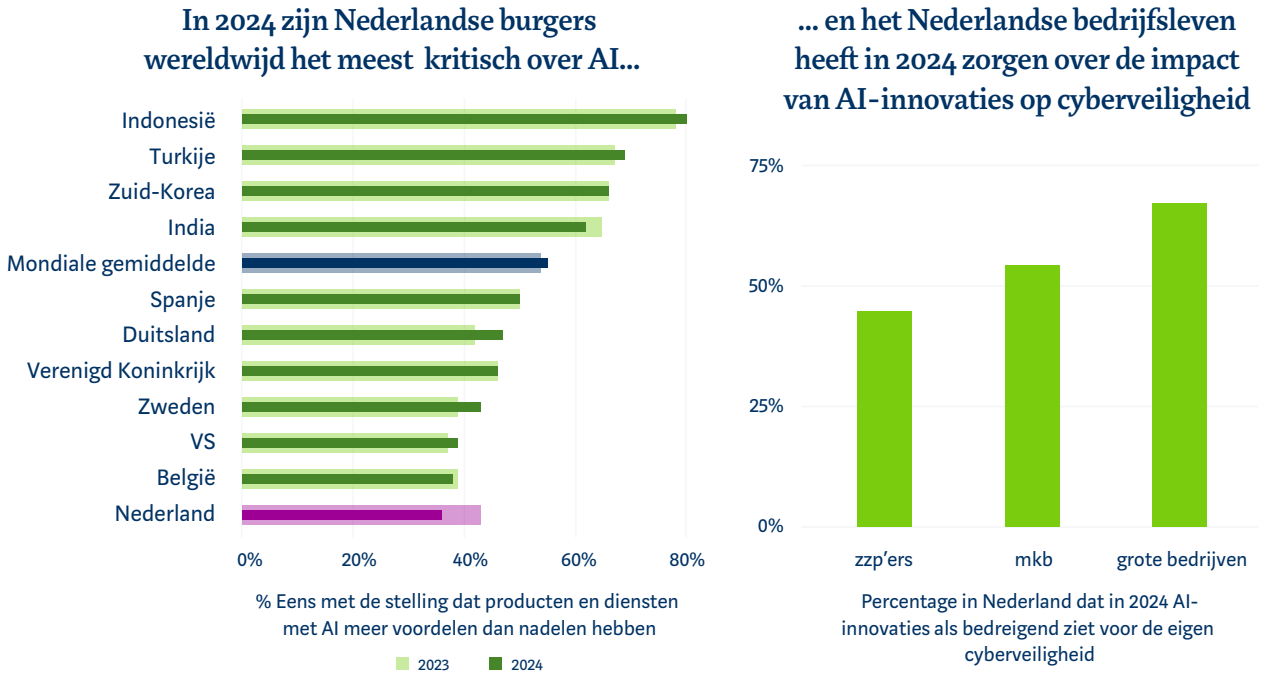
Het vertrouwen in AI-systemen onder Nederlandse burgers is verhoudingsgewijs laag en zorgen over sommige AI-risico's in het bedrijfsleven nemen toe. Wereldwijd reageert ruim de helft van de mensen positief op de stelling dat producten en diensten met AI meer voordelen dan nadelen hebben. In Nederland stokt dit aantal op 36%. Een jaar geleden was dit nog 43%. Daarmee staat Nederland onderaan een lijst van 32 landen.

Dit blijkt uit een jaarlijkse mondiale AI- monitor van Ipsos. Een zorgpunt dat nadrukkelijk meer aandacht krijgt binnen het Nederlandse bedrijfsleven is de impact van AI-innovaties zoals, generatieve AI-technologie, op cyberveiligheid. Uit onderzoek van ABN AMRO, in samenwerking met MWM[2], blijkt dat meer dan 50% het Nederlandse bedrijven deze zorgen heeft. Een jaar geleden was dit nog minder dan een kwart. De grootste zorgen zitten bij grote bedrijven.
Zie ook figuur 1.1.

1.2 Onstuimige groei van AI-technologie

Veel investeringen in AI-technologie zorgen de komende jaren voor verdere ontwikkeling en verspreiding van AI. Vooral bij generatieve AI was er het afgelopen jaar een explosieve stijging te zien van durfkapitaalinvesteringen. Het zwaartepunt hiervan bevindt zich in de Verenigde Staten. Er ontstaan nieuwe bedrijven, maar met name gevestigde organisaties dragen eraan bij dat het brede publiek nieuwe AI-systemen verder omarmt. Zo werkt Microsoft aan de integratie van taalmodellen in veelgebruikte Officepakketten,

FIGUUR 1.1: PERCEPTIE VAN AI-RISICO'S ONDER NEDERLANDSE BURGERS EN BEDRIJVEN



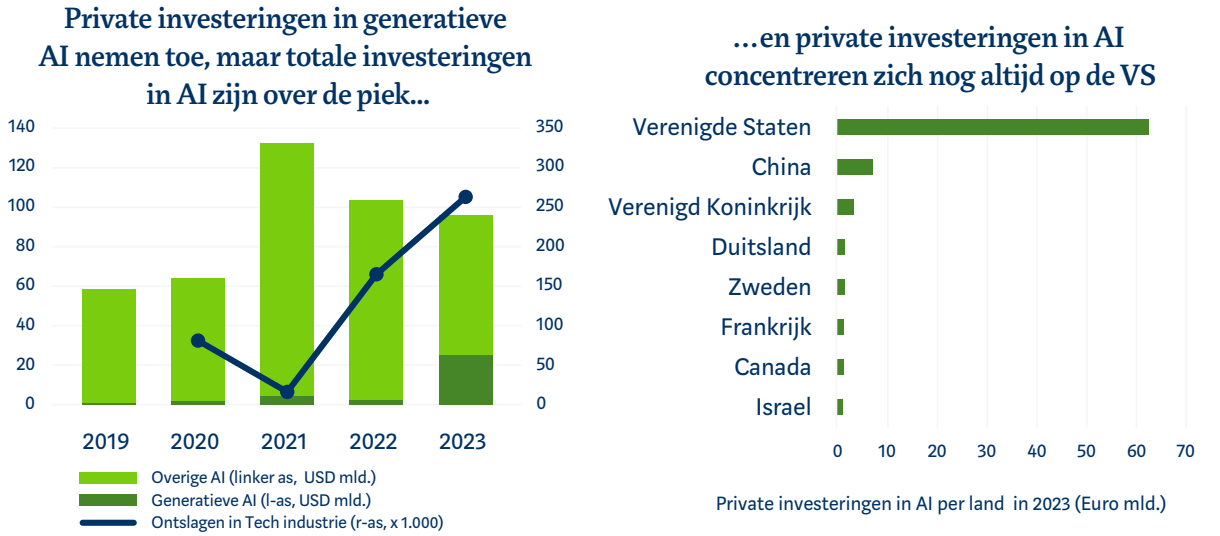
BRON (LINKS): IPSOS AI MONITOR (N = 23.685, 32 LANDEN)
BRON (RECHTS): ABN AMRO (2024) IN SAMENWERKING MET ONDERZOEKSBUREAU MWM2

gooide Google onlangs de meestgebruikte zoekmachine op de schop om meer AI in te zetten, en werkt Apple samen met OpenAI aan de integratie van taalmodellen in de besturingssystemen van Apple. Volgens voorspellingen van de Amerikaanse bank Morgan Stanley zullen AI-pc's al in 2028 het merendeel van de markt vormen.[2]

Maar niet alle verwachtingen van groei worden waargemaakt. Onderzoek van adviesbureau BCC laat zien dat het gebruik van generatieve AI de prestaties van werknemers kan verbeteren, maar ook kan verminderen. Dit is afhankelijk van de taak.[3] Waar een jaar geleden soms het beeld heerste dat generatieve AI overal bij kan helpen, ontstaat er nu een realistischer beeld van kansen in specifieke toepassingsgebieden. Daarnaast zijn de totale investeringen in AI afgelopen jaar verder gedaald. De piek was in 2021. Toen kreeg de technologiesector een impuls door verschillende omstandigheden, waaronder de versnelde digitalisering door de coronapandemie. Het einde van de pandemie viel in 2023 samen met het bijstellen van te hoge verwachtingen binnen de technologiesector. Dit uitte zich onder meer in veel ontslagen in die sector.

In de concurrentiestrijd tussen grote techbedrijven wordt nieuwe technologie vaak halsoverkop gelanceerd. Voor Google komt meer dan de helft van de inkomsten voort uit reclame bij zoekopdrachten. Ontwikkelt OpenAI een gebruiksvriendelijkere zoekmachine, bijvoorbeeld door samenwerking met de Microsoft Bing zoekmachine? Dan staat voor Google de bedrijfsvoering op het spel. Daarom hebben techbedrijven haast en lanceren zij nieuwe producten snel. Hierbij zien zij blootstelling van het product aan grote delen van de samenleving als experiment. Zo lanceerde Google onlangs de nieuwe zoekmachine AI-Overviews. Toen die veel kritiek kreeg, reageerde Google hier een aantal weken later op in een blog. De zoekmachine gaf onder andere advies over het aantal stenen dat een persoon per dag zou moeten eten.[4] Microsoft kondigde een grootschalige introductie aan van Recall, een tool die alles wat je op het scherm hebt gezien kan terughalen. Na ophef over de beveiliging van deze tool, draaide Microsoft de lancering terug.[5] OpenAI haalde onlangs de stem van Sky offline. Deze stem was een week eerder gelanceerd voor de communicatie met ChatGPT. Het stemgeluid van Sky leek te veel op de stem van actrice Scarlett Johansson, met wie OpenAI gehaast contact bleek te hebben gehad in de dagen voor de nieuwe release.[6]

FIGUUR 1.2: MARKTINDICATOREN VOOR INVESTERING EN ONTWIKKELING AI-INDUSTRIE

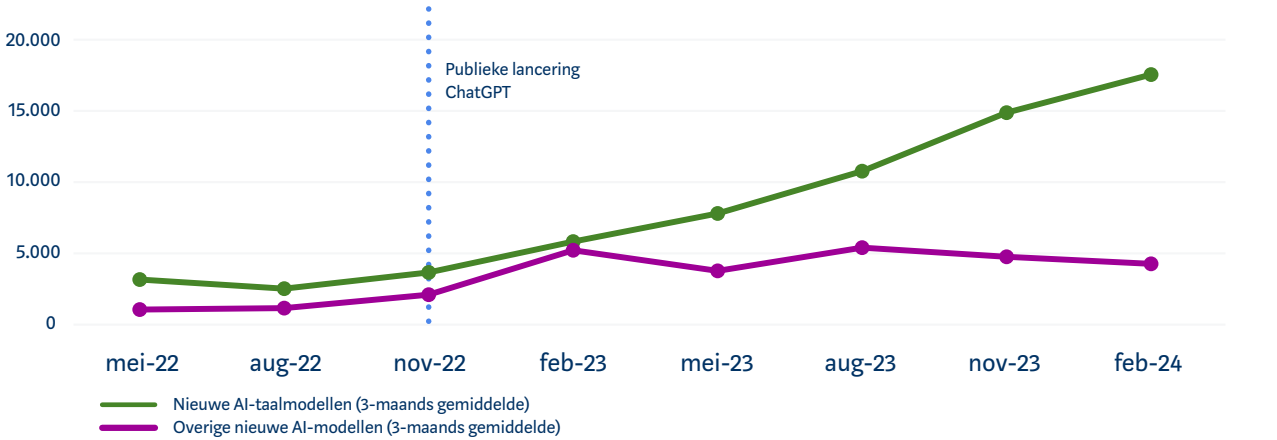


BRONNEN: STANFORD UNIVERSITY, 2024 AI INDEX REPORT EN LAYOFFS

1.3 Ontwikkelingen in general-purpose AI

Veel aandacht van ontwikkelaars richt zich op het afdekken van fundamentele zwaktes van taalmodellen. De belangrijkste basis van taalmodellen is de transformer-architectuur.[7] Deze modelvorm kan tekst aanvullen met volgende woorden op basis van patronen die het model heeft leren kennen uit bestaande teksten. State-of-the-art modellen genereren hiermee tekst die compleet natuurlijk klinkt.

FIGUUR 1.3: SINDS BEGIN 2023 GROEIT HET AANBOD VAN AI-TAALMODELLEN VEEL SNELLER DAN DAT VAN OVERIGE AI-MODELLEN



BRONNEN: OECD.AI (2024) OP BASIS VAN DATA VAN HUGGING FACE (22 APRIL 2024)

De modellen redeneren echter niet met alle kennis over de echte wereld. Dit brengt fundamentele uitdagingen met zich mee. Taalmodellen kunnen bijvoorbeeld tekst genereren met een compleet onlogische betekenis, terwijl het model zeker is over de woordencombinatie. Veel inspanningen van ontwikkelaars zijn er nu op gericht om de taalmodellen in de beantwoording gebruik te laten maken van extra informatie, die verifieerbaar correct is.

Een methode om verifieerbare informatie te gebruiken is het finetunen van modellen op basis van menselijke kennis. Dit kan bijvoorbeeld door met zorg opgestelde vraag-antwoordcombinaties te laten zien aan een model. Of door een model te laten leren van scores die mensen geven aan gegenereerde uitkomsten.[8] Een andere methode om de uitkomsten te verbeteren is meer context meegeven aan de vraag. Door in een vraag ook de inhoud van een gezaghebbende bron mee te geven, kan een taalmodel bijvoorbeeld gestuurd worden om antwoorden op die inhoud te baseren. Een vervolgstap hierin is het automatiseren van het opzoeken en toevoegen van de juiste gezaghebbende bronnen als context voor een vraag. Dit noemen we een Retrieval Augmented Generation (RAG)-systeem.

Nieuwe AI-modellen, zoals voor generatieve AI, zijn een zogeheten frontiertechnologie en daarmee moeilijk voorspelbaar. Frontiertechnologieën zijn technologieën die zich bevinden op het snijvlak van wetenschappelijke doorbraken en implementatie in de samenleving. AI is het paraplubegrip voor zo'n technologie die is voortgekomen uit de wetenschap en die na een aantal doorbraken toepassing vindt in de praktijk, zoals recentelijk generatieve AI. De snelle ontwikkeling en adoptie bieden ons nog maar een eerste glimp van wat deze technologie mogelijk gaat maken en gaat betekenen in de samenleving. Zoals met veel ontwikkelingen in technologie of de samenleving, ligt de focus in eerste instantie op toepassingen waaruit de kracht van technologie blijkt. Bijvoorbeeld hoe generatieve AI realistisch materiaal kan genereren. Maar de focus ligt ook op toepassingen die de mens en de samenleving gemak brengen of die bijzonder waardevol zijn, zoals in de medische sector. Ook zijn er verrassende toepassingen, die vooraf niet te voorzien of voorspellen waren.

AI-systemen worden ook ingezet voor malafide praktijken, zoals oplichting door social engineering. Door in te zetten op bepaalde vertrouwenselementen, kunnen oplichters gemakkelijker te werk gaan. Waar berichten of mailverkeer nog te wantrouwen zijn, vertrouwen veel mensen wel op hun zintuigen. Een beeld van een persoon of een digitaal gesprek vertrouwen mensen al snel, mits het beeld geloofwaardig genoeg is. Dit laat bijvoorbeeld een recent incident zien waarbij een multinational meer dan 25 miljoen dollar naar fraudeurs overmaakte nadat een medewerker werd uitgenodigd voor een videogesprek met een deepfake van de CEO. Bonafide toepassingen kunnen gereguleerd worden op basis van welwillende actoren. Malafide toepassingen daarentegen zijn veel moeilijker om grip op te krijgen. Generatieve AI is al veelvuldig hiervoor ingezet, bijvoorbeeld door deepfakes te maken voor oplichting of (wraak) pornografie. De kansen van frontiertechnologie kunnen zich ook in bedreigingen uiten. CEO-fraude is hiervan een goed voorbeeld.

1.4 Opkomst van AI-veiligheidsinstituten

In reactie op de bewustwording van risico's van generatieve AI zijn in het afgelopen half jaar verschillende AI-veiligheidsinstituten van start gegaan. Onder andere in het Verenigd Koninkrijk en de Verenigde Staten is een AI Safety Institute opgericht. In Europa is er het AI Office, dat voortvloeit uit de Europese AI-verordening. Een belangrijke doelstelling van deze AI-veiligheidsinstituten is gevaren van AI-(taal)modellen bloot te leggen, door deze op verschillende manieren te testen en te evalueren.[9] Soms vinden deze testen plaats voordat AI-(taal)modellen publiekelijk beschikbaar komen. Het werk van deze instituten heeft onder andere tot doel om onderzoek te doen naar de mogelijkheden waarop kwaadwillende gebruikers geavanceerde AI-technologieën kunnen misbruiken. De beleidsaanpak en strategieën voor dit soort instituten zijn nog volop in ontwikkeling. Omdat geavanceerde en multifunctionele AI-modellen direct wereldwijde impact hebben, is het noodzakelijk dat AI-veiligheidsinstituten zoveel mogelijk samenwerken. Een belangrijke stap daarbij is dat een grote groep landen recentelijk heeft aangegeven te willen werken aan interoperabiliteit van het werk van AI-veiligheidsinstituten.[10] Bijvoorbeeld door het ontwikkelen van onderling uitwisselbare testsystemen en vergelijkbare beoordelingskaders. Dit maakt kennisdeling en een gezamenlijk fundament voor testen mogelijk.

Grote taalmodellen zijn nog altijd kwetsbaar voor de meeste basale omzeilingstechnieken. Dit blijkt uit een eerste onderzoek door zo'n AI-veiligheidsinstituut naar de beschermingsmaatregelen in grote taalmodellen. Het doel van deze beschermingsmaatregelen is te voorkomen dat kwaadwillende gebruikers het systeem kunnen misbruiken om gevoelige en schadelijke informatie te krijgen of genereren (omzeiling). Zoals vertrouwelijke informatie en/of persoonsgegevens. Maar ook gevoelige informatie over bijvoorbeeld cybersecurity of terreurvraagstukken. Een eerste test door het Britse AI Safety Institute naar vier prominente taalmodellen laat zien dat deze allemaal zeer kwetsbaar zijn voor simpele omzeilingstechnieken. En dat het met geavanceerdere technieken zelfs bijna altijd lukt om minimaal een keer per vijf pogingen de modelbegrenzingen te omzeilen.[11] Het model geeft dan antwoorden op vragen waarvoor het instructies heeft gekregen om die niet te beantwoorden.

AI-veiligheidsinstituten moeten ook kijken naar meer risico's dan alleen misbruik. Bijvoorbeeld door aandacht te schenken aan de techno-sociologische impact. Een bevinding van het Amerikaanse National AI Advisory Committee is dat het beoordelen en testen van AI-veiligheid verder moet gaan dan alleen het beoordelen van (technische) modelkwetsbaarheden, omdat AI-technologie diep is ingebed in de samenleving. AI-systemen zijn onderdeel van bredere processen en worden door mensen gebruikt. De veiligheid van AI-technologie moet dus mede vanuit het brede techno-sociologische perspectief bekeken worden.[12] Denk aan hoe artsen, ambtenaren, onderwijzers of rechters in de praktijk zaken beoordelen en voortbouwen op suggesties die (mede) door generatieve AI worden gedaan. Dit soort evaluaties gebeurt echter nog maar beperkt. Tot nu toe hebben de meeste evaluaties van geavanceerde AI-taalmodellen een technische inslag. De suggestie is dat techno-sociologische evaluaties ruimte bieden om ook te kijken naar hoe mensen met dit soort AI-systemen omgaan en bijvoorbeeld hoe bias en discriminatie worden ondervangen. Dit vraagt mede om pilots, gefaseerde invoering en impactstudies. Het Amerikaanse AI Safety Institute geeft in de onlangs vastgestelde strategie van dit instituut invulling aan deze brede interpretatie.[13]

Ook in Nederland kan worden nagedacht over een beleidsaanpak waarin vanuit de publieke taak AI-veiligheid actief wordt getest – dit hangt nauw samen met het toezicht op de AI-verordening. Het toezicht op de AI-verordening waarborgt compliance van individuele systemen met productstandaarden. Vanuit een AI-veiligheidstaak kan breder en vergelijkend onderzoek gedaan worden, bijvoorbeeld door te testen. Dit brengt risico's in kaart en geeft richting aan de doorontwikkeling van standaarden en aan de regels in de AI-verordening. Het inrichten van een AI-veiligheidstaak op nationaal niveau draagt ook bij aan de samenwerking die nodig zal zijn met het Europese AI Office en AI-veiligheidsinstituten in andere landen. Als Nederland deze kennis en vaardigheden heeft, draagt dit bovendien bij aan een ecosysteem met internationale aantrekkingskracht voor AI-ontwikkelaars.[14]

Box 1.1

AI biedt kansen voor mensen met een beperking

Met de opkomst van AI zijn er steeds meer initiatieven om mensen met een beperking zelfstandig mee te laten doen in de samenleving. Een voorbeeld is een AI-bril die de omgeving voor een slechtziende gebruiker kan beschrijven. Zoals een obstakel op straat of wat er op een verpakking staat in de supermarkt. Een recent onderzoek van de Organisation for Economic Cooperation and Development (OECD) geeft een overzicht met 142 van dit soort AI-toepassingen voor mensen met een beperking.[15]

AI-toepassingen bieden kansen voor de inclusie van mensen met een beperking. Zo heeft het UWV in samenwerking met verschillende werkgevers onder meer een AI-app getest voor mensen met stemproblemen en een spraakherkenningssysteem (met *machine learning*) voor doven en slechthorenden. Het UWV stelt op basis van de pilots dat de technologie helpt om de arbeidsparticipatie van mensen met een beperking te verhogen, evenals hun werkplezier en autonomie.[16]

AI biedt ook kansen om mensen met een beperking te helpen om deel te nemen aan het democratische proces. Hiertoe is de EU bijvoorbeeld het onderzoeksproject iDEM gestart, met als doel AI-taalmodellen te ontwikkelen die informatie over publieke zaken begrijpelijker maken. Daarbij wordt ook onderzocht of taalmodellen mensen met een verstandelijke beperking kunnen ondersteunen om hun mening te uiten.[17]

Tegelijkertijd lopen mensen met een beperking vaak het risico gediscrimineerd te worden door AI-systemen. Hiervoor waarschuwt de speciaal rapporteur voor mensen met een beperking van de Verenigde Naties. Voorbeelden zijn gezichtsdetectiesoftware die mensen met een afwijking in het gezicht minder goed herkent en bancaire AI-systemen die incorrect hoofletttergebruik in de schriftelijke aanvraag voor een lening als indicator zien voor slecht betaalgedrag – dit laatste zal overwegend mensen met dyslexie vaak treffen. De boodschap is daarom alert te zijn op de risico's van AI-systemen voor mensen met een beperking. Een aanbeveling is om bij de ontwikkeling van AI expliciet rekening te houden met beperkingen, bijvoorbeeld door mensen met een beperking actief te betrekken bij het ontwikkelproces.

Toegankelijkheidseisen zijn ook onderdeel van de AI-verordening. Aanbieders van AI-systemen met een hoog risico moeten ervoor zorgen dat deze systemen voldoen aan toegankelijkheidseisen. Dit gaat over de manier waarop informatie wordt verstrekt, maar ook over de gebruikersinterface en functionaliteit.

1.5 Binnenlandse lessen en ontwikkelingen

Generatieve AI doet nadrukkelijk zijn intrede in diverse publieke organisaties, bijvoorbeeld in de zorg. Volgens de Rijksbrede visie op generatieve AI wil de Rijksoverheid bijvoorbeeld "kennisdeling faciliteren over de mogelijkheden voor veilig gebruik van generatieve AI door het delen van kennis en praktijkervaring."[18] De publieke sector lijkt generatieve AI vooralsnog vooral in experimenten en pilots te gebruiken. Dat geldt in elk geval voor de zorgsector.[19] Voorbeelden uit die sector tonen ook hoe divers generatieve AI wordt toegepast. Zorginstellingen gebruiken generatieve AI voor administratief werk, communicatie met patiënten en het samenvatten van patiëntendossiers.

Elk toepassingsgebied vergt voor verantwoord AI-gebruik contextspecifieke regulering naast generieke regulering. Elk toepassingsgebied kent immers eigen bestaande normen. In de financiële sector mag de inzet van AI de financiële soliditeit en de integriteit van financiële instellingen bijvoorbeeld niet in gevaar brengen.[20] Verder vergt de bescherming van fundamentele waarden en grondrechten in verschillende contexten verschillende handelswijzen. Sectorspecifieke normen kunnen algemene wetten aanvullen die grondrechten beschermen, zoals de AI-verordening en de AVG. In de financiële sector zijn er volgens DNB en AFM nog weinig sectorspecifieke normen voor AI-gebruik. Er is echter wel een groeiende behoefte aan sectorspecifieke regulering van AI-gebruik.[21]

Er wordt veel geëxperimenteerd met camerasystemen om ongewenst gedrag te herkennen in bijvoorbeeld winkels, sportscholen of het openbaar vervoer. Dat is relevant omdat gedragsmonitoring kan raken aan emotieherkenning. Het verwerken van persoonsgegevens voor deze toepassingen is in veel gevallen al onder de AVG verboden. Vanaf februari 2025 is het op de markt brengen en gebruiken van dergelijke toepassingen op de werkplek en in het onderwijs onder de AI-verordening ook verboden. In andere contexten is emotieherkenning een hoogrisicotoepassing binnen de AI-verordening. Veel huidige experimenten gebruiken met AI toegeruste camera's echter vooral om bezoekers of gebruikers te monitoren. Zo kondigde een supermarktketen aan een camerasysteem in te zetten dat mogelijke diefstal zou moeten kunnen herkennen. Een sportschoolketen is van plan een systeem te gebruiken dat agressief gedrag, noodgevallen, drukte en sporten zonder lidmaatschap zou kunnen detecteren.[22] Ook binnen het openbaar vervoer loopt momenteel een pilot met een camerasysteem dat agressief gedrag zou moeten herkennen op Amsterdam Centraal.[23] Dergelijke systemen gelden mogelijk als emotieherkenningssystemen met een hoog risico onder de AI-verordening, maar zijn in veel gevallen ook onder de AVG al onrechtmatig.

Nederland heeft een leidende rol in het ontwikkelen en inzetten van verantwoorde AI-systemen én bijbehorend toezicht. Door grote incidenten in de afgelopen jaren is onder Nederlandse beleidsmakers het risicobewustzijn hoog bij het ontwikkelen van verantwoorde inzet van AI. Volgens de Global Index on Responsible AI[24] staat Nederland bovenaan de lijst van verantwoorde inzet van AI. Tegelijkertijd komen er regelmatig incidenten, risico's en negatieve effecten in de keten aan het licht. Het lerend vermogen na eerdere problemen lijkt nog niet overal op orde. Dit laat zien dat de weg omhoog is ingezet, maar ook dat er aanhoudend aandacht voor nodig is om risico's van AI-systemen te beheersen en kansen te benutten. Bijvoorbeeld voor het assisteren van mensen. Nieuwe producten voor mensen met een beperking gebruiken AI om de levenskwaliteit te verbeteren (zie box 1.1).

De effecten van en de verplichtingen bij de inzet van een AI-systeem zijn zelden eendimensionaal. Dit is bijvoorbeeld zichtbaar bij het systeem dat Rijkswaterstaat gebruikt om de inzet van wegbeheerders bij incidenten te optimaliseren. Dit systeem adviseert de meldcentrale over beschikbare wegbeheerders en de kortste aanrijtijden voor deze hulpdiensten bij een incident. Ook adviseert het systeem waar wegbeheerders het beste gepositioneerd kunnen zijn voor optimale dekking en minimale aanrijtijden.
Onder medewerkers heeft dit systeem echter tot ophef geleid, omdat het monitoring van gedrag mogelijk zou maken en het afbreuk zou doen aan de autonomie en kennis van wegbeheerders.

Het AI-systeem van Rijkswaterstaat scoort volgens de Algemene Rekenkamer onvoldoende op beheersing. Specifiek op (model)kwaliteit, maatregelen en privacywaarborgen oordeelde de Algemene Rekenkamer dat het systeem niet voldoet aan gestelde eisen of criteria.
Het oordeel is dat Rijkswaterstaat onvoldoende zicht heeft op de accuraatheid van het model. Hierdoor weet de organisatie niet hoeveel de inzet van het AI-systeem bijdraagt aan snellere afhandeling van een incident.[25] Opvallend is dat – naast onvoldoende privacywaarborgen, die onder de AVG verplicht zijn – de punten die een onvoldoende scoren grotendeels overeenkomen met de punten uit aanvullende regelgeving voor hoogrisicosystemen die de AI-verordening binnenkort verplicht stelt. Daarom is het des te belangrijker om de beheersing van dit AI-systeem op orde te krijgen.

Veel organisaties zetten in op nieuwe vormen van fraudedetectie met algoritmes. De risico's daarvan zijn in het bijzonder onder de aandacht door de toeslagenaffaire.[26]
In het maatschappelijk domein zetten bijvoorbeeld woningcorporaties algoritmes in om woonfraude op te sporen.[27] Door het belang van huisvesting is dit een toepassing met een hoog risico voor de grondrechten van het individu. Een aandachtspunt is dat een dergelijk fraude-algoritme een hoogrisicosysteem is onder de AI-verordening. Dit brengt specifieke eisen aan beheersing en transparantie met zich mee. Hoofdstuk 4 gaat verder in op de risico's van fraudealgoritmes en een maatregel om deze te beheersen.

De nieuwe doorstroomtoets voor leerlingen uit groep 8 van de basisschool heeft dit jaar tot ophef geleid. Daarbij valt op dat veel van deze toetsen AI-systemen zijn.[28] Tien jaar geleden zijn in Nederland digitale, adaptieve toetsen geïntroduceerd. Scholen kregen hierbij de keuze tussen verschillende toetsaanbieders.[29] Bij een adaptieve toets krijgt elke leerling een geïndividualiseerde toets. De vragen worden dan gaandeweg moeilijker of makkelijker, totdat het AI-systeem voldoende zekerheid heeft om tot een advies te komen. Deze nieuwe doorstroomtoets is dit jaar op een nieuwe manier toegepast. Een nieuwe normering voor de verschillende toetsen moet deze onderling vergelijkbaar maken, door verplichte standaardvragen. Tegelijkertijd zijn scholen sinds dit jaar verplicht om zich aan de toetsuitslag te houden wanneer het advies op basis van de doorstroomtoets (bijvoorbeeld vwo) hoger is dan het voorlopig schooladvies (bijvoorbeeld havo).

Scholen mogen nu nog alleen gemotiveerd van het advies afwijken. De koepelorganisatie in het basisonderwijs (PO-raad) heeft naar aanleiding van de eerste doorstroomtoets aangegeven dat scholen de toetsuitslagen niet konden plaatsen en dat de uitslagen niet bleken te corresponderen met de resultaten in het leerlingvolgsysteem.[30] Het ministerie van OCW heeft sindsdien aangegeven dat dit jaar onderlinge verschillen tussen de toetsen aan de oppervlakte zijn gekomen, waarop eerder minder goed zicht was. Daarom doet het ministerie samen met het College voor Toetsen en Examens onderzoek naar de oorzaak van deze verschillen. Hierbij wordt ook aandacht wordt besteed aan het adaptieve karakter van de doorstroomtoetsen.[31]

Een actueel aandachtspunt bij dit soort adaptieve toetsen is de mate waarin wordt voldaan aan (toekomstige) vereisten voor AI-systemen. AI-systemen om het passende onderwijsniveau te bepalen, inclusief waartoe kinderen toegang krijgen, zijn hoogrisicotoepassingen onder de AI-verordening. Dit brengt verplichtingen met zich mee, bijvoorbeeld voor hoe transparant de toetsresultaten moeten zijn. Hierbij valt op dat sommige adaptieve doorstroomtoetsen slechts beperkt informatie verstrekken aan kinderen. Bij sommige toetsen krijgen leerlingen bijvoorbeeld niet te zien welke vragen zij goed of fout hebben beantwoord. Ook het vermogen om zelf regie te houden is bij sommige adaptieve toetsen beperkt. Bijvoorbeeld doordat leerlingen niet terug kunnen gaan naar eerdere vragen als ze die willen verbeteren. Andere relevante vereisten zijn bijvoorbeeld dat het systeem voldoende nauwkeurig voorspelt en dat het niet discrimineert. Ook moet het AI-systeem zodanig ingezet worden dat een menselijke controleur, zoals een docent, kan besluiten om de uitkomst van het AI-systeem niet te gebruiken. Relevant is de vraag hoe dit zich verhoudt tot het verplichtende karakter dat de toets heeft gekregen bij afwijkingen omhoog.

1.6 Stappen in kwaliteit en risicobeheersing

AI-systemen raken steeds verder verweven in steeds meer processen en systemen in onze samenleving. Het bewustzijn groeit dat de inzet van deze systemen vaak invloed heeft buiten de enkele toepassing ervan. Bijvoorbeeld bij het vervangen van een menselijke beoordeling in een proces. Of bij het verder automatiseren van controles met een AI-systeem. Dit heeft veelal effecten die veel verder strekken dan enkel de beoordeling of enkel de controlehandeling. Het beïnvloedt andere elementen van een proces of keten. Dit zijn veranderingen of effecten die vaak niet vooraf in kaart zijn gebracht en die zelden actief worden opgemerkt.

Het is begrijpelijk dat er wordt gezocht naar houvast om deze veranderende processen of ketens te kunnen beheersen. Innovatie moet immers ook gepaard gaan met verantwoordelijkheid. Wet- en regelgeving biedt steeds meer handvaten om de beheersing in te richten. Maar zoals de effecten divers zijn, is het begeleidend wettelijk kader dat vaak ook. Maar zelden heeft een organisatie die AI-systemen inzet één wettelijk kader om zich aan te houden. Vrijwel altijd zal het een samenspel zijn van algemene en specifieke of sectorale wetgevingskaders. Maar ook de rollen zijn diverser dan op het eerste gezicht vaak wordt gedacht. Ontwikkelaars zijn niet altijd een en dezelfde partij. Bijvoorbeeld als een ontwikkelaar toepassingen op een foundation model van een derde partij baseert. Tegelijkertijd kan een organisatie die een AI-systeem inzet, ook zelf de ontwikkelaar zijn of worden. En meerdere rollen tegelijk hebben.

Een sluitend model voor adequate beheersing van alle AI-systemen is niet mogelijk of nuttig, gezien de snelheid van ontwikkeling en innovatie. Wel is duidelijk dat er stappen worden gezet door organisaties die investeren in innovatie én beheersing. De huidige toepassingen zullen enkel groeien in complexiteit, zowel op technologisch vlak als in de veelvoud van actoren en de schaal van toepassing. Toepassingen die nu zichtbaar zijn, zijn nog maar het topje van de ijsberg. De uitdagingen voor beheersing van huidige systemen moeten snel worden opgepakt, om te voorkomen dat de achterstand in beheersing nieuwe toepassingen onmogelijk maakt. Dit vraagt om een gezamenlijke en holistische benadering, vanuit meerdere wettelijke kaders. Maar ook vanuit toezicht, bedrijfsleven en overheid. En vooral vanuit de gemeenschappelijke wens om de kansen die AI-systemen bieden op een verantwoorde wijze te kunnen benutten.

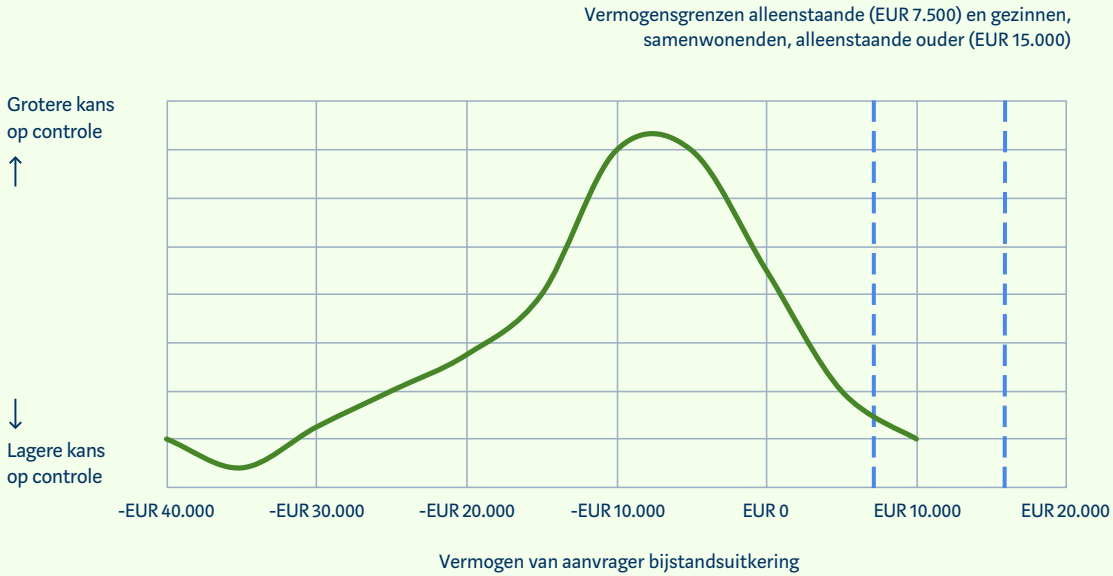
Box 1.2

Uitdagingen om machine learning rechtmatig, eerlijk, transparant en zonder willekeur in te zetten

Praktische initiatieven om het goed te doen, maken ook duidelijk hoe moeilijk het écht is om AI-systemen verantwoord in te zetten. Een voorbeeld is de Slimme Check van de gemeente Amsterdam. Dit is een algoritme dat bij aanvragen voor een bijstandsuitkering aangeeft of deze een onderzoek op onrechtmatigheid waard zijn.[32] In het Amsterdamse algoritmeregister is te lezen hoe dit algoritme de onderzoeken gelijkwaardiger en effectiever zou laten verlopen. De technische documentatie, procesoverzichten, maatschappelijke overwegingen en bias-analyse zijn openbaar beschikbaar. De onzekerheid over de eerlijkheid van het algoritme bleek echter te groot. Daarom heeft de wethouder besloten het algoritme uit voorzorg niet in te zetten.[33]

De complexiteit en ethische dilemma's bij een dergelijk algoritme worden duidelijk als je kijkt naar het belang van de indicator 'totaal vermogen' bij de onderzoekswaardigheid van een aanvraag. Mensen met te veel geld hebben geen recht op een bijstandsuitkering. Uit de gegevens blijkt echter dat iemand met een schuld van zo'n 8000 euro volgens het algoritme meer onderzoekswaardig is dan iemand zonder schulden. Mogelijk is hier een statistische verklaring voor, maar het is waarschijnlijk moeilijk te accepteren voor de mensen die onderzoek krijgen. Door voortschrijdend inzicht door dit soort praktische initiatieven, ontstaan steeds meer standaarden en kwaliteitsraamwerken. Hoofdstuk 5 gaat hier verder op in.

FIGUUR 1.4: DE INZET VAN *MACHINE LEARNING* IN DE PRAKTIJK: OP BASIS VAN DE WET IS ER EEN MAXIMAAL VERMOGEN OM BIJSTANDSGERECHTIGD TE ZIJN, MAAR MACHINE LEARNING STUURT AAN OP HET ONDERZOEKEN VAN NEGATIEVE VERMOGENS



Toelichting: *Deze grafiek is een versimpelde en gestileerde weergave van de documentatie over het 'Onderzoekswaardigheid Algoritme Slimme Check', een AI-systeem dat als pilot is opgezet door de gemeente Amsterdam. Via machine learning op basis van historische data kent het risicomodel, als alle overige omstandigheden gelijkblijven, een grotere kans op onderzoekswaardigheid toe aan bijstandsaanvragers met een negatief vermogen (oftewel, een schuld) van rond de 10.000 euro, dan aan bijstandsaanvragers met een positief vermogen of een negatief vermogen van bijvoorbeeld 20.000 of 30.000 euro.*

BRON: ALGORITMEREGISTER GEMEENTE AMSTERDAM

Daarbij zorgt de inzet van machine learning voor extra uitdagingen. In het genoemde voorbeeld is, na gedocumenteerde afweging, het vermogen van de bijstandsaanvrager als geschikte risico-indicator geselecteerd. De motivatie hiervoor is dat dit een 'kernfeit' is voor de bijstandsaanvraag en dat "[een aanvrager] als het vermogen te groot is niet bijstandsgerechtigd is".[34]

Na kalibratie van het risicomodel via machine learning, geeft het algoritme op hoofdlijnen echter een grotere onderzoekswaardigheid aan negatieve vermogens dan aan positieve vermogens. Zie figuur 1.4 voor een versimpelde weergave van de marginale bijdrage van deze indicator aan de risicoprofilering. De uitwerking is contra-intuïtief, want het is juist een te groot vermogen dat een bijstandsuitkering verhindert.

De ontwikkelaars erkennen dit ook in de toelichting bij deze indicatoren en dragen verschillende argumenten aan. Bijvoorbeeld dat schulden een complicerende factor zijn bij het bepalen van de rechtmatigheid van een bijstandsuitkering en dat schulden alleen van het vermogen mogen worden afgetrokken als schulden moeten worden terugbetaald. Tegelijkertijd is het zo dat ook iemand met een totaal vermogen dat positief is óók schulden kan hebben.

Bij een dergelijke indicator kan (i) uitlegbaarheid en (ii) het voorkomen van willekeur problematisch zijn. Juist omdat het model is getraind met historische gegevens (over onderzoekswaardigheid), moet rekening gehouden worden met de mogelijkheid dat deze indicator – in hoe deze is gekalibreerd – een proxy is voor iets anders. Of dat deze een statistisch significant verband waarneemt bij iets wat niet relevant is. Als er geen duidelijke logica is waarmee de daadwerkelijke bijdrage aan de risico-inschatting van een indicator uit te leggen is, ontstaan problemen met uitlegbaarheid en het voorkomen van willekeur.
De AI-verordening gaat hoge eisen stellen aan uitlegbaarheid: iedereen die een met AI genomen besluit heeft gekregen, heeft recht op inhoudelijke uitleg over de rol van het AI-systeem en de belangrijkste elementen van het genomen besluit. We mogen ervan uitgaan dat individuele risico-indicatoren onderdeel zijn van deze belangrijkste elementen. Zie de bijlage voor een verdere bespreking van de uitdagingen bij het beheersen van AI-systemen.

1.7 Nationale AI-agenda

AI-systemen bieden kansen voor de samenleving, die actief onderzocht en benut moeten worden. Niet alleen door de overheid óf het bedrijfsleven, maar gezamenlijk binnen de kaders en waarden van de samenleving. Wet- en regelgeving biedt relevante kaders, maar voor de invulling en sturing van kansen is een nationale strategie nodig.
De AP heeft in de RAN Editie 2, Najaar 2023[35] opgeroepen om met een deltaplan voor beheersing van risico's te komen. Maar ook voor het benutten van kansen is een (strategisch) deltaplan essentieel. Het huidige Strategische Actieplan Artificiële Intelligentie[36] is in 2019 verschenen. Relevante ontwikkelingen in technologie en kennis over toepassingen en effecten ontbreken in dit actieplan door het hoge tempo waarin ontwikkeld wordt. TNO Vector heeft op het jaarsymposium 'Strategische autonomie in een open economie' vier papers gepubliceerd[37] die inzicht geven in de huidige rol van technologieën in nationale veiligheid, energieleveringszekerheid, kennis en innovatie, maar ook in de kosten en baten van digitale autonomie. Hierin wordt aangegeven dat gerichte beleidsmaatregelen nodig zijn, maar dat die ontbreken of dat inzicht in de effecten ervan ontbreekt.
Dit ondersteunt de noodzaak om naast beleid en regelgeving een duidelijke strategie te hebben, die in woelige tijden houvast kan bieden om kansen te benutten maar ook burgers en de samenleving te beschermen. Dit vraagt om nationale keuzes over investeringen, focusgebieden en kernwaardes. Maar bij snel ontwikkelende technologieën ook om doorlopende aanpassingen en updates van de strategie.

Een overkoepelende strategie zal de verantwoorde ontwikkelingen op het gebied van AI in Nederland versterken. Verschillende Nederlandse initiatieven laten de afgelopen jaren zien dat Nederland deel uitmaakt van de beweging naar verantwoordere inzet van AI. Bijvoorbeeld de Fundamental Rights and Algorithms Impact Assessment (FRAIA). Er studeren hier relatief veel studenten af in AI. Ook zijn onderzoekers uit Nederland goed vertegenwoordigd op wereldwijde conferenties over verantwoorde AI.[38] Steeds meer ontwikkeling, zoals de hiervoor genoemde Slimme Check-pilot bij de gemeente Amsterdam, wordt transparant uitgevoerd en relevante toezichthouders denken samen na over effectieve vervolgstappen.[39] Door de Nederlandse ambities voor de komende jaren verder te formuleren, kunnen beleidsmakers deze positieve ontwikkelingen versterken. In de tweede editie van de RAN[40] staat onder het kopje 'Deltaplan Algoritmes & AI: Ambitie 2030' een aantal mogelijke thema's om terug te laten komen in de strategie. Dit loopt van 'menselijke regie' tot het 'nationaal ecosysteem en infrastructuur'.

2. Informatievoorziening in de democratie onder druk door AI-systemen

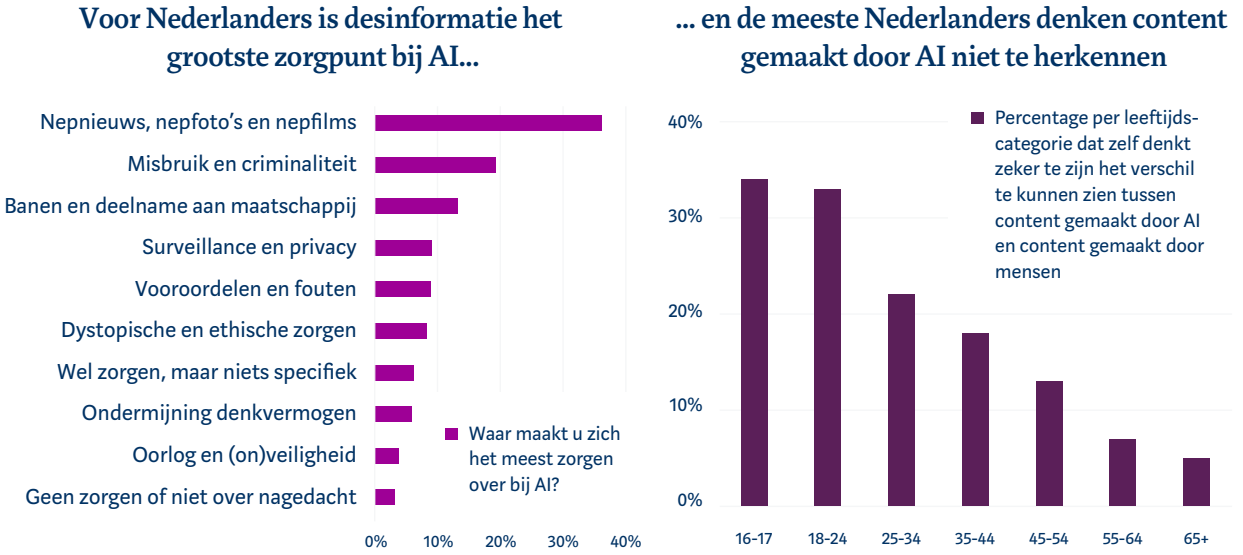
SNEL NAAR DIT ONDERDEEL

Nieuwsvergaring en -consumptie gebeuren steeds meer online. Vooral jongeren gebruiken sociale media als voornaamste nieuwsbron.[41] AI-systemen hebben een grote stempel gedrukt op de online informatievoorziening. Platforms hebben hierbij veel macht door de inzet van verslavende AI-aanbevelingssystemen. Met AI kan ook op grote schaal informatie worden gegenereerd. Een groot risico is het genereren en verspreiden van desinformatie en misinformatie. De macht van platforms en het verkeerd gebruiken van generatieve AI brengen de diversiteit van het informatielandschap in gevaar, waardoor democratische processen zoals verkiezingen onder druk komen te staan. In hoeverre het functioneren van het democratisch systeem daadwerkelijk wordt geraakt is op dit moment moeilijk meetbaar. Mede ingegeven door de verdere opkomst van AI in de komende jaren is het daarom belangrijk dit actief te monitoren.

Dit hoofdstuk is mede gebaseerd op publieke input. In het voorjaar van 2024 heeft de AP met een publieke oproep gevraagd om visies, inzichten en zorgen betreffende AI-systemen in de informatievoorziening. Hierop zijn een tiental reacties ontvangen die elk hele eigen inzichten, expertise en specifieke zorgen naar voren brachten. Deze kennis en inzichten zijn gebruikt bij de ontwikkeling van dit hoofdstuk, in de voorbereiding voor gesprekken die ten behoeve van dit hoofdstuk zijn gevoerd, en voor interne gedachtevorming.

Recente gebeurtenissen tonen aan dat AI een steeds grotere invloed heeft op het publieke debat. In januari 2024 ontvingen inwoners van New Hampshire in de Verenigde Staten een deepfake robocall van president Biden. Een AI-gegenereerde robotstem aan de telefoon imiteerde de stem van Biden en riep voorstanders van de Democratische Partij op om niet naar de stembus te gaan tijdens de voorverkiezingen.[42] In India werd het internet tijdens de verkiezingen in het voorjaar van 2024 overspoeld door deepfakes. Het gaat om allerlei soorten misleidende informatie: van humor en satire tot illegale, beledigende en schadelijke content.[43] Een ander voorbeeld is een AI-gegenereerd beeld

FIGUUR 2.1: MEESTE ZORGEN OVER DESINFORMATIE EN WEINIG VERTROUWEN IN HERKENNEN VAN AI-GEGENEREERDE CONTENT



BRON: WAAG FUTURELAB, APRIL 2024 (LINKS) EN ALGOSOC AI OPINION MONITOR 2024 (RECHTS)

van een tentenkamp in Rafah, dat in mei 2024 tientallen miljoenen keren gedeeld werd op sociale media met de tekst All Eyes on Rafah. De rol van AI hierin is prominent: een virtueel, 'schoon' beeld dat aan menselijk leed refereert zonder te choqueren, stimuleert gebruikers om zich op deze manier tegen de oorlog uit te spreken. Daarnaast glipt het door de geautomatiseerde moderatiesystemen, die bloedige foto's en kritische uitingen onderdrukken. De reacties op het viraal gaan van de post zijn wisselend. Enerzijds geeft de AI-gegenereerde afbeelding de ernst van de situatie in Rafah mogelijk niet juist weer. Anderzijds zorgt dit juist voor een groter bereik voor de boodschap die de post uitdraagt.[44]

Nederlanders maken zich zorgen over de risico's van (generatieve) AI op de informatievoorziening in onze samenleving. Wanneer Nederlanders wordt gevraagd naar hun grootste zorgen over AI, denken de meeste mensen aan de mogelijke invloed, verspreiding en misbruik van nepnieuws, -foto's en -filmpjes. Ook zijn er zorgen dat gegenereerde informatie niet van echt te onderscheiden is (zie figuur 2.1).[45] [46] Dit heeft grote invloed op het vertrouwen in de informatievoorziening. De AP analyseert de rol van AI in dit hoofdstuk aan de hand van de informatievoorziening-scyclus: creatie/productie, moderatie/distributie en consumptie. Om de verschillende stadia van de cyclus en de beïnvloeding via verschillende kanalen te beschouwen. En om te zien welke maatregelen genomen kunnen worden om de risico's zo veel mogelijk te verkleinen.

2.1 Creatie / productie

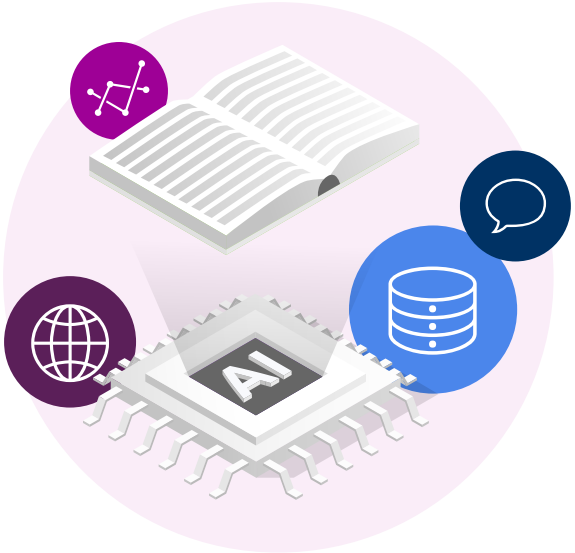
Het produceren en genereren van content is met generatieve AI makkelijker dan ooit. Hiermee is het mogelijk om zeer snel tekstuele en audiovisuele output te creëren die nauw aansluit bij een specifiek verzoek.[47] Problematisch is dat gegenereerde content steeds lastiger te herkennen is.

Met generatieve AI wordt het mogelijk gemaakt om op grote schaal desinformatie te genereren. Desinformatie is onware, inaccurate of misleidende informatie die opzettelijk wordt verspreid om mensen te verwarren of manipuleren. Bijvoorbeeld om politieke of ideologische steun te krijgen, om andere ideologieën in een kwaad daglicht te stellen of om wantrouwen en polarisatie te creëren.[48] Dit kan gevolgen hebben voor democratische processen, de economie en de nationale veiligheid.[49]

Generatieve AI-systemen produceren daarnaast ook onbedoeld misinformatie. Misinformatie wordt niet-doelbewust verspreid, bijvoorbeeld omdat mensen niet in de gaten hebben dat het om valse of misleidende informatie gaat.[50] De effecten van misinformatie kunnen echter nog steeds schadelijk zijn. Omdat generatieve AI antwoorden 'verzint', en hierbij geen onderscheid kan maken tussen wat waar is en wat niet, ontstaan er al snel fouten. Een taalmodel als ChatGPT geeft regelmatig onjuiste informatie, terwijl het die informatie als feitelijk presenteert. Dit voorjaar ontstond er ophef omdat AI-systemen de verspreiding van desinformatie en het zaaien van angst aanbevolen als campagnestrategie voor de Europese verkiezingen.[51] Zo kan AI-gegenereerde misinformatie onbewust de wereld in worden geholpen. Zeker als ontwikkelaars, (overheids)organisaties en gebruikers dergelijke systemen te veel beschouwen als alwetend of daadwerkelijk intelligent.

AI-gegenereerde content neemt steeds vaker de plaats in van broncontent. Een voorbeeld is Google Overview, dat het 'oude' zoeken laat verdwijnen door als eerste resultaat geen websitelinks maar een AI-samenvatting aan te bieden.[52] De kwaliteit van de AI-zoekdienst laat nog te wensen over, maar in de komende jaren zullen dit soort AI-systemen een steeds grote rol gaan spelen op verschillende platforms.[53] Het gevolg is dat de klassieke zoekmachine, die op een navolgbare wijze de resultaten ordent, daardoor verdwijnt. Soms ontbreken bronnen, wat het lastiger maakt om te controleren waar de gegenereerde informatie vandaan komt. De betrouwbaarheid van de informatie is daardoor onbekend. Dit vergroot het risico op misinformatie.

Tegelijkertijd wordt onderzocht of AI-systemen ook geschikt zijn om te detecteren of er AI is ingezet. Speciale AI-detectietools kunnen mogelijk AI-teksten herkennen. Bepaalde woorden en zinconstructies worden bijvoorbeeld statistisch gezien veel vaker door AI gebruikt dan door menselijke schrijvers.[54]



Labelen of watermerken helpt om echt materiaal van gegenereerd materiaal te onderscheiden. De AI-verordening stelt het in 2025 verplicht om gegenereerde beelden te labelen. Dit houdt in dat iedereen die een deepfake maakt of verspreidt, openheid moet geven over de herkomst en de techniek die is gebruikt. Het doel hiervan is om transparantie over het gebruik van AI te stimuleren en daarmee manipulatieve content tegen te gaan. Ontwikkelaars en platforms geven steeds vaker aan content te gaan labelen.[55]

Meer transparantie over de herkomst van informatie helpt gebruikers om te beoordelen of informatie betrouwbaar is of niet. Een gezamenlijk initiatief van meerdere bedrijven is het Content Authenticity Initiative (CP2A).[56] Dit is een coalitie van mediabedrijven, technologiebedrijven en NGO's, die technische standaarden heeft ontwikkeld die het onder meer met cryptografie mogelijk maken om de herkomst van beelden te verifiëren. Bedrijven zoals de BBC, Microsoft en Google zijn aangesloten bij C2PA. In maart 2024 heeft de BBC voor het eerst gebruikgemaakt van C2PA, door onder een video meer informatie te geven over de herkomst van de video en toe te lichten hoe de video op echtheid is gecontroleerd.[57] Op grond van de AI-verordening zijn aanbieders van AI-tools om materiaal te genereren straks ook verplicht om dit soort technieken toe te passen om gegenereerde inhoud als zodanig herkenbaar te maken.

2.2 Moderatie en distributie van informatie

Het informatieaanbod is te groot om zelf in korte tijd te bepalen welke artikelen relevant, nuttig en betrouwbaar zijn: er is sprake van een information overload.[58] Vóór het digitale tijdperk selecteerden nieuwsredacties van kranten en televisie het aanbod voor het publiek. Door de grotere hoeveelheid online informatie is de poortwachtersfunctie van de traditionele media deels weggevallen. Burgers zijn in het digitale tijdperk niet alleen consumenten van informatie, ze produceren en verspreiden zelf ook informatie. Bijvoorbeeld met toegankelijke generatieve AI-tools, die op grote schaal beschikbaar zijn geworden in 2024. Om deze online informatiestromen behapbaar te maken, wordt er door AI-systemen, als een soort redactie, gefilterd op informatie die voor de consument persoonlijk interessant zou zijn. Waar nieuwsredacties inhoudelijke keuzes maken, zijn AI-systemen hier niet toe in staat. AI kan bijvoorbeeld waarheden en onwaarheden niet van elkaar onderscheiden, maar registreert wel wat voor content de gebruiker aanspreekt.

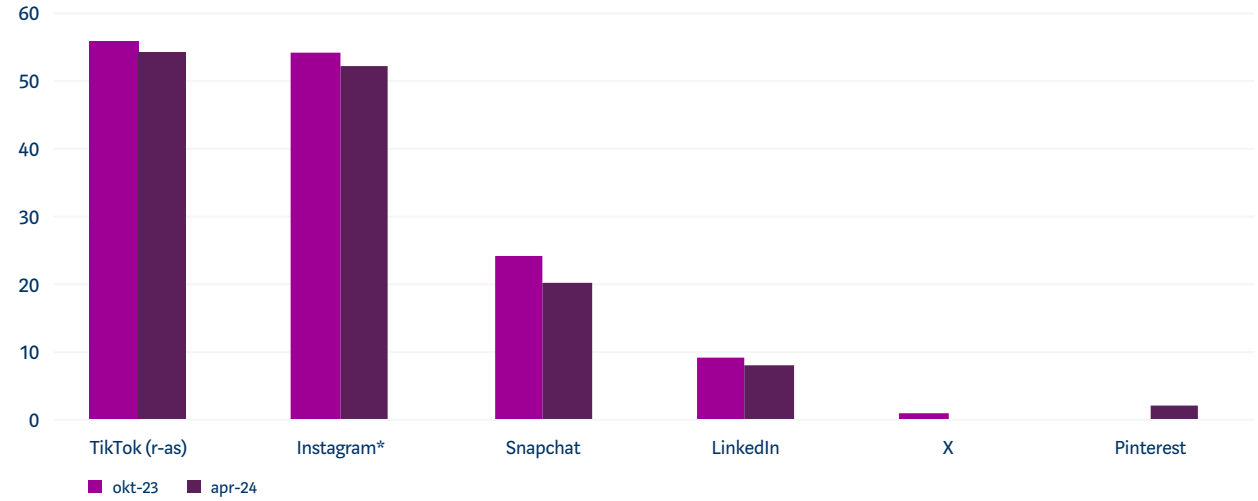
De online informatievoorziening ligt in handen van een paar bigtechplatforms. Wereldwijd gebruikt ruim 90 procent de zoekmachine van Google.[59] Veel jongeren gebruiken steeds vaker het videoplatform TikTok als zoekmachine. [60] Platforms hebben door deze macht een online poortwachtersfunctie.

Bigtechplatforms bepalen voor een groot deel ons 'informatiedieet': wat we te zien krijgen, maar ook wat we niet te zien krijgen. Een gepersonaliseerd AI-systeem moet ervoor zorgen dat gebruikers zo lang mogelijk op hun platform blijven, zodat ze zoveel mogelijk advertenties zien. Bigtechplatforms zijn voor hun inkomsten in belangrijke mate afhankelijk van advertenties. De gepersonaliseerde AI-systemen worden bewust verslavend gemaakt (addictive design). Dit zorgt er bijvoorbeeld voor dat er veel berichten getoond worden die emoties oproepen.[61] Omdat deze platforms vooral gefocust zijn op inkomsten en niet op een zo divers mogelijk informatieaanbod, kan dit de pluriformiteit, betrouwbaarheid en onafhankelijkheid van de informatievoorziening in gevaar brengen.

Sommige overheden zetten moderatie in om informatie op online platforms te censureren. Dit brengt de vrijheid van meningsuiting en van informatie in gevaar. [62] Volgens een rapport van Freedom House (2023) zetten minstens 22 onderzochte landen AI in om ongewenste politieke en religieuze uitingen te verwijderen van socialemediaplatforms.[63] In de Information Technology Act in India is opgenomen dat de centrale overheid en andere autoriteiten in noodsituaties bevelen kunnen uitvaardigen om socialemedia-accounts, video's, berichten of foto's te blokkeren en verwijderen wanneer de inhoud schadelijk wordt geacht voor de openbare orde, vrede, soevereiniteit en veiligheid van het land. Dit wordt breed geïnterpreteerd en veelvuldig ingezet om kritische content te blokkeren en verwijderen.[64] Socialemediaplatform X heeft meerdere kritische berichten van de oppositie onzichtbaar moeten maken voor de Indiase gebruikers van X.[65] Het mogelijke gevaar is dat met de inzet van deze regelgeving censuur makkelijker, onopgemerkt en op grote schaal kan worden toegepast.[66]

FIGUUR 2.2: AFNAME VAN NEDERLANDSTALIGE CONTENTMODERATORS BIJ GROTE SOCIALEMEDIAPLATFORMS

Aantal Nederlandstalige contentmoderatoren per platform



BRONNEN: DSA VLOP TRANSPARANCY REPORTS

Om de invloed van grote platforms op de informatievoorziening in te perken, moeten sinds augustus 2023 online platforms en zoekmachines zich houden aan de Digital Services Act (DSA). Deze wet dwingt zeer grote online platforms, die aangewezen zijn door de Commissie, om desinformatie aan te pakken (zie box 2.2). Ook moeten platforms meer openheid geven over inhoudsmoderatie. Moderatoren controleren of berichten in lijn zijn met de voorwaarden van het platform. Het is hun taak om illegale of schadelijke inhoud te verwijderen. Ondanks dat platforms aangegeven desinformatie zo veel mogelijk tegen te gaan, zien we bij bijna alle platforms een afname in (Nederlandstalige) moderatoren (zie figuur 2.2).

Platforms moeten transparant zijn over hoe zij informatie modereren. Ook moet er de mogelijkheid zijn om de keuzes van de moderatoren te betwisten als moderatie in strijd lijkt met de vrijheid van meningsuiting. In de DSA is vastgelegd dat onafhankelijke organisaties de status van betrouwbare *flagger* kunnen aanvragen. Betrouwbare flaggers zijn organisaties die illegale content opsporen en deze melden bij de platforms. Deze meldingen moeten platforms met voorrang behandelen. De flaggers staan onder toezicht van de ACM en moeten jaarlijks de meldingen publiceren.[67]

Gebruikers moeten meer invloed krijgen op het informatieaanbod dat ze te zien krijgen. In de DSA zijn goede eerste stappen gezet. Voor zeer grote online platformen, geldt een extra verplichting waarbij gebruikers de mogelijkheid krijgen om persoonlijke aanbevelingssystemen uit te schakelen. Meerdere zeer grote online platforms bieden momenteel deze optie.

De distributie van platforms moet gericht zijn op een divers informatieaanbod. Door het gebruik van verslavende aanbevelings-en filtersystemen kan de diversiteit van het media-aanbod afnemen. Dit is schadelijk voor het publieke debat. Aanbevelingssystemen kunnen ook worden ingezet om juist een divers informatieaanbod te tonen.
Om de kernwaarden van het mediabeleid te beschermen heeft het Commissariaat voor de Media, toezichthouder op de mediasector, recentelijk een verkenning uitgebracht die de effecten van AI hierop in kaart brengt. De box 'Hoe AI de kernwaarden van mediabeleid uitdaagt' geeft inzicht in deze effecten.

2.3 Consumptie van informatie

Desinformatie en misinformatie kunnen zorgen voor toenemend wantrouwen in de media. Ook als de onjuiste informatie al ontkracht is. Niet iedereen die desinformatie te zien krijgt, zal naderhand meekrijgen dat het om onjuiste informatie ging. Door de toenemende hoeveelheid AI-gegenereerde content op online platforms neemt het wantrouwen ten opzichte van informatie toe. Veel Nederlanders denken ook niet te kunnen herkennen dat content door AI gegenereerd is. Dit kan ervoor zorgen dat zij legitieme informatie in twijfel trekken. Journalistiek bewijs kan dan bijvoorbeeld worden weggezet als deepfake, terwijl het wél om echte beelden of audio-opnames gaat.

Box 2.1

Hoe AI de kernwaarden van mediabeleid uitdaagt

Door: Commissariaat voor de Media

Voor een goed werkende democratie is het essentieel dat iedereen een eigen mening kan vormen. Het Commissariaat voor de Media (hierna: Commissariaat) draagt daaraan bij door een onafhankelijk, pluriform (divers), toegankelijk en veilig media-aanbod te bewaken en te stimuleren. Daartoe houdt het Commissariaat toezicht op de regels uit de Mediawet. Maar het Commissariaat agendeert ook ontwikkelingen die invloed hebben op de eerdergenoemde waarden. AI is een van deze ontwikkelingen. Er liggen veel kansen in het gebruik van AI, maar het gebruik van AI levert ook risico's op. Hierna bespreken we per waarde de belangrijkste kansen en risico's.

Onafhankelijkheid van het media-aanbod is nauw verwant aan een andere waarde: de **betrouwbaarheid** van het media-aanbod. Aan de ene kant ziet het Commissariaat dat de inzet van generatieve AI de betrouwbaarheid van informatie negatief kan beïnvloeden. Denk hierbij aan het kwaadwillig inzetten van AI om misleidende informatie te genereren. Generatieve AI-toepassingen zijn daarnaast vaak niet transparant over hoe ze werken. Dit noemen we ook wel een black box. Dat maakt het lastig te begrijpen hoe en door wie een afbeelding, tekst of video is gemaakt. Aan de andere kant ziet het Commissariaat dat de inzet van AI journalisten en de media ook kan ondersteunen. Maar dan moet AI verantwoord worden ingezet, bijvoorbeeld door richtlijnen. Media-instellingen kunnen deze richtlijnen zelf opstellen. Daarbij moeten ze aangeven welk gebruik van AI wel en niet is toegestaan. Een richtlijn kan ook aangeven hoe en in welke mate mensen worden geïnformeerd over de rol van AI in de totstandkoming van het media-aanbod dat ze tot zich nemen. Het Commissariaat merkt wel op dat meer transparantie juist ook kan leiden tot meer wantrouwen.

AI levert ook enkele risico's op voor de **pluriformiteit** van het media-aanbod. Het Commissariaat ziet bijvoorbeeld dat mensen door aanbevelings- en filtersystemen online mogelijk minder verschillende soorten media-aanbod te zien krijgen. Daarnaast hebben bedrijven die AI-toepassingen ontwikkelen veel macht in de markt. Zo is de infrastructuur van AI grotendeels in handen van een kleine groep bedrijven. Ook de 'opiniemacht' die deze bedrijven hebben neemt toe. Dit komt omdat big tech invloed heeft op de eerdergenoemde aanbevelings- en filtersystemen. En daarmee kan bepalen wat voor media-aanbod mensen te zien krijgen. Er zijn wel regels, bijvoorbeeld in de Digital Services Act en de Digital Markets Act, die ervoor moeten zorgen dat de gatekeepers minder macht krijgen.

Het Commissariaat ziet verder dat AI veel kansen biedt voor de **toegankelijkheid** van het media-aanbod. Met name mensen met auditieve of visuele beperkingen kunnen door AI makkelijker toegang krijgen tot de digitale wereld. Bijvoorbeeld door de inzet van op AI gebaseerde automatische ondertiteling, vertaalsystemen en audiodescriptie.

AI kan ook de online **veiligheid** van mensen vergroten. Bijvoorbeeld door te helpen bij het automatiseren van leeftijdsverificatie en het modereren van schadelijk media-aanbod. Maar het gebruik van AI kent ook bij deze waarde risico's. Deepfakes kunnen mensen misleiden met realistische, door AI gegenereerde bewerkingen van afbeeldingen, geluidsfragmenten of video's. Dit kan schadelijke gevolgen hebben voor het democratisch proces. Bijvoorbeeld als deepfakes worden gebruikt om verkiezingen te beïnvloeden. Zo werd in 2023 twee dagen voor de Slowaakse verkiezingen een deepfake-geluidsfragment van een politicus verspreid. Tot slot kunnen AI-toepassingen negatieve effecten hebben op de mentale gezondheid, vooral bij jongeren. Bijvoorbeeld door de verslavende algoritmen die socialemediabedrijven inzetten.[68]

Deze box is geschreven door het Commissariaat voor de Media, toezichthouder op de Mediawet. Zie voor meer informatie over dit thema de publicatie 'Tussen Bits en Principes: Hoe AI de kernwaarden van mediabeleid uitdaagt' van juni 2024.

Door de inzet van AI wordt er op grote schaal geautomatiseerd invloed uitgeoefend op het publieke debat. De AIVD stelt in het jaarrapport van 2022 dat verschillende landen bewust desinformatie in omloop brengen om de Nederlandse bevolking een positiever, maar onjuist, beeld te geven van (acties van) hun land.[69] Desinformatie wordt vaak verspreid door anonieme accounts. Steeds vaker zijn dit bots: accounts die volledig geautomatiseerd zijn en worden aangestuurd door een AI-systeem. In 2023 waren bots voor 49,6 procent verantwoordelijk voor al het internetverkeer.[70] Bots opereren vaak op grote schaal, zetten desinformatie uit en verstoren zo het maatschappelijke debat. Bots kunnen onder andere informatie creëren, berichten liken en delen en interactie hebben met gebruikers. Op die manier doen ze zich voor als echte personen en stemmers. Door veelvuldig gebruik van hashtags en het liken van bepaalde berichten sturen ze het aanbevelingssysteem, om zo invloed uit te oefenen op wat voor berichten wel en niet worden weergegeven.[71] [72]

Desinformatie is vaak gericht tegen gemarginaliseerde groepen die ook in de offline wereld achtergesteld zijn of gediscrimineerd worden.[73] Een voorbeeld is het toegenomen antisemitisme, dat online extra zichtbaar is. Veel antisemitische content bestaat uit desinformatie, variërend van conservatief-nationalistische berichtgeving tot complottheorieën. AI-systemen spelen een belangrijke rol in het blootstellen van gebruikers aan dit soort desinformatie, misinformatie en haat. Dit is ook te zien aan de grote hoeveelheid vrouwonvriendelijke content, die met name gericht is op jongens. Al na 5 dagen op het platform TikTok kan het aanbevelingssysteem een vervierdubbeling van dat soort content aanbevelen op iemands persoonlijke pagina, op basis van onschuldige interesses als mentale gezondheid of fitness. Op die manier kunnen jongeren in online 'echokamers' terecht komen waarin vrouwonvriendelijke retoriek genormaliseerd wordt.[74] Daarnaast worden vrouwen steeds vaker slachtoffer van seksuele deepfakes (gemanipuleerd beeldmateriaal).[75] Hoewel gekunsteld (pornografisch) materiaal al langer een probleem is, maken AI-tools het makkelijker om deze video's en foto's te maken. Ook in Nederland is dit een groeiend probleem onder vrouwelijke BN'ers en politici.[76]

Een gebrek aan een gemeenschappelijke 'informatiebasis' kan polarisatie in de hand werken. In de afgelopen jaren is de interesse in het volgen van nieuws gedaald, met name onder 18-34-jarigen.[77] Het nieuws waar zij wel mee in aanraking komen, krijgen ze vooral mee via sociale media. Daarnaast ervaren steeds meer mensen de hoeveelheid nieuws als vermoeiend. Dit geldt met name voor mensen die vaak het nieuws mijden. Dit is een groep die 8% van de Nederlandse bevolking beslaat (een verdubbeling ten opzichte van 2017). Ook voor hen zijn sociale media de voornaamste nieuwsbron.[78] Op sociale media is de kans groter dat de berichtgeving die hen wel bereikt onjuiste informatie bevat of manipulatief en eenzijdig is. Hoewel de omvang van het effect van deze 'filterbubbels' onzeker is, komen mensen hierdoor gemakkelijker in aanraking met extreme standpunten. Extreme content levert namelijk interactie met het sociale medium op. Zij lopen het risico om het zicht op de feiten en het geloof in een gedeelde werkelijkheid te verliezen. Als deze gedeelde werkelijkheid als gemeenschappelijke basis wegvalt, kan dit bijdragen aan een gepolariseerd politiek en sociaal landschap, waarin vooral de eigen gemeenschap wordt vertrouwd.[79] Sociale media en AI-systemen zijn niet eenduidig als oorzaak van deze ontwikkeling aan te wijzen, maar maken deze wel zichtbaarder.

Een combinatie van maatregelen is nodig om de risico's van AI voor de online informatievoorziening te verkleinen. Transparantie over de herkomst van digitale content is nodig om de betrouwbaarheid van informatie te kunnen vaststellen. Bronvermelding, labelen en watermerken kunnen hierbij helpen. Ook kan AI ingezet worden om gegenereerde content te herkennen. Veel AI-tools om te detecteren zijn nog relatief nieuw en zullen zich in de toekomst verder worden ontwikkeld. Daarnaast zijn mediawijsheid en algoritmische geletterdheid noodzakelijk om op de juiste manier te kunnen omgaan met online informatie.

AI-geletterdheid en mediawijsheid zijn nodig om mee te kunnen doen in de digitale informatiesamenleving. Mediawijsheid is het geheel van kennis, vaardigheden en mentaliteit waarmee burgers zich bewust, kritisch en actief kunnen bewegen in een digitale mediasamenleving.[80] Vanuit het Rijk zijn er diverse initiatieven om burgers weerbaar en mediawijs te maken.[81] Digitale geletterdheid wordt bijvoorbeeld een verplicht onderdeel van het onderwijscurriculum.[82] AI-geletterdheid moet een belangrijk onderdeel zijn van digitale geletterdheid. Kennis van AI-systemen en van de risico's en effecten hiervan is nodig om veilig mee te kunnen doen in de digitale informatiesamenleving. Het landelijk expertisecentrum voor het curriculum (SLO) heeft het doel van AI-geletterdheid opgenomen in de conceptkerndoelen van digitale geletterdheid.[83] Het is belangrijk dat niet alleen kinderen en jongeren, maar ook volwassenen zich bewust(er) worden van de invloed van AI-systemen op de informatievoorziening. De risico's van AI voor de informatievoorziening raken iedereen.

Box 2.2

De Digital Services Act

De Digital Services Act (DSA) verplicht zeer grote online platforms en zoekmachines maatregelen te nemen tegen systeemrisico's, zoals bedreigingen voor de democratie. Verspreiding van desinformatie kan en systeemrisico zijn onder de DSA. Zeer grote online platforms moeten hier maatregelen voor nemen. De DSA is geïnitieerd door de Europese Commissie (EC) en geldt in alle lidstaten van de Europese Unie.

De EC publiceert richtsnoeren voor grote online platforms en zoekmachines om negatieve effecten op verkiezingen te voorkomen.[84] De richtsnoeren kleuren de verplichtingen van platforms en zoekmachines onder de DSA verder in met concrete voorbeelden en best practices. Het volgen van de richtsnoeren is echter niet verplicht. Platforms en zoekmachines mogen ook op andere manieren dan omschreven in deze richtsnoeren aan de DSA voldoen door systeemrisico's te beperken en op te vangen. Het succes van de maatregelen is daarmee sterk afhankelijk van de bereidheid en interpretatie van de platforms en zoekmachines.

Een paar van de concrete voorstellen in de richtsnoeren zijn gerichte aanpassingen van aanbevelingssystemen om verkiezingsprocessen te beschermen. Zeer grote online platforms kunnen bijvoorbeeld maatregelen nemen om ervoor te zorgen dat hun aanbevelingssystemen geen aantoonbare desinformatie tonen bij verkiezingen.

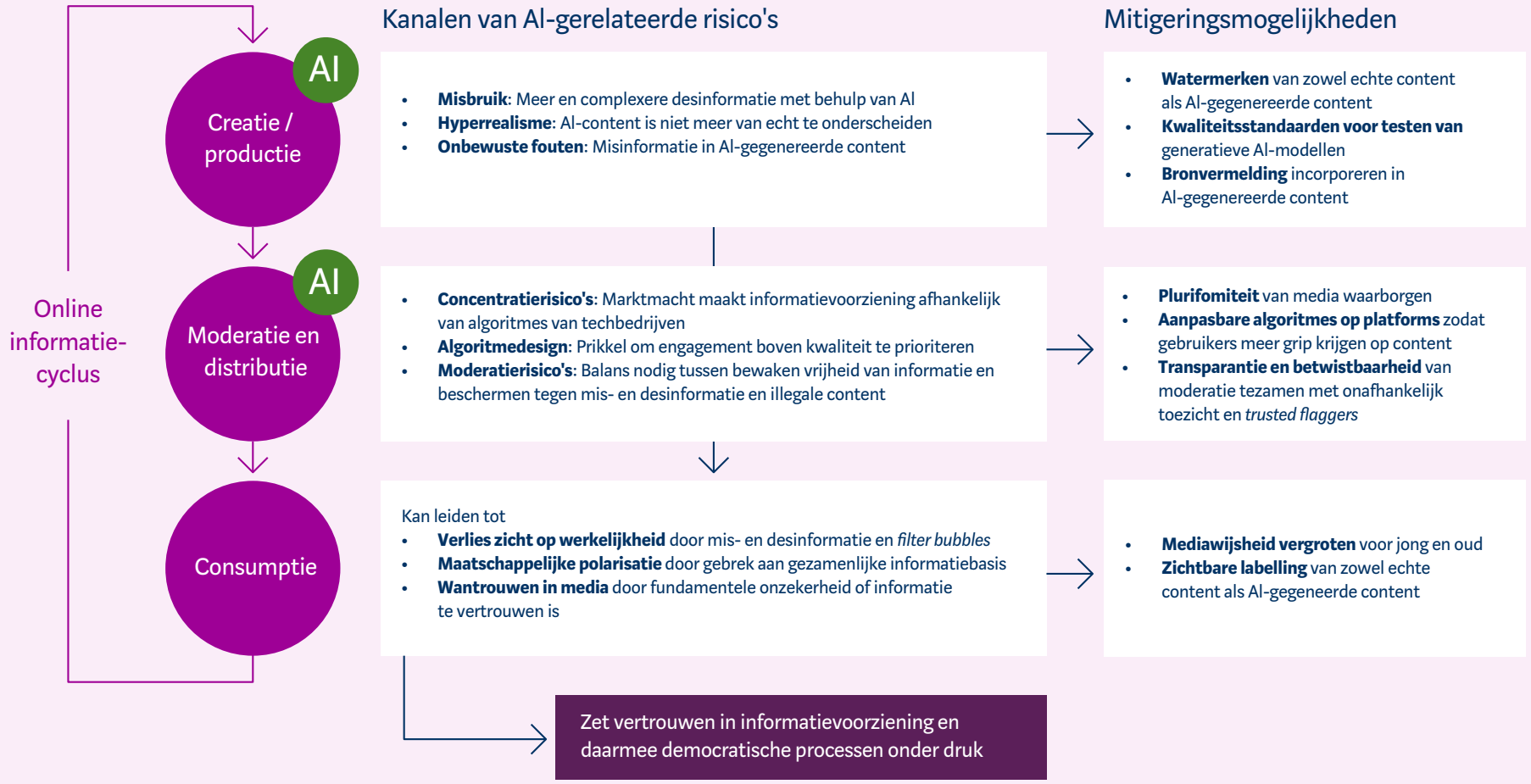
Zeer grote online platforms moeten over zulke maatregelen wel transparant en duidelijk zijn. Ook stellen de richtsnoeren voor om de herkenbaarheid van officiële accounts te vergroten. Dit kan helpen om de verspreiding van desinformatie en misinformatie op zeer grote online platforms tegen te gaan. Een ander voorbeeld is het zodanig ontwerpen van aanbevelingssystemen dat gebruikers zinvolle controle krijgen over de informatie die zij tot zich nemen.

De richtsnoeren bevatten ook specifieke voorstellen voor generatieve AI. De EC beveelt zeer grote online platforms aan om door AI gegenereerde content te voorzien van een watermerk. Op die manier is door AI gegenereerde content eenvoudiger te herkennen. Ook moet door AI-systemen gegenereerde informatie zo veel mogelijk op betrouwbare bronnen worden gebaseerd om onjuiste output te voorkomen. Daarnaast noemen de richtsnoeren herhaaldelijk het belang van het verhogen van de mediawijsheid. Het verbeteren van de algoritmische geletterdheid kan burgers helpen om alert te zijn op de mogelijke verspreiding van misinformatie of desinformatie bij verkiezingen.

De EC, nationale toezichthouders, civil society organisaties en de zeer grote online platforms hebben een stresstest uitgevoerd om te kijken of grote platforms klaar zijn voor verkiezingstijden. De EC liep in de stresstest verschillende scenario's door en bekeek of de interne procedures en werkwijzen van grote online platforms en zoekmachines effectief systeemrisico's bij verkiezingen tegengaan. De scenario's betreffen bijvoorbeeld de verspreiding van politieke deepfakes om kiezers te misleiden.[85] De EC maakt de uitkomsten van de stresstest niet openbaar.

De EC ziet op basis van de DSA actief toe op de beheersing van systeemrisico's bij verkiezingen. De EC is recentelijk een onderzoeksprocedure gestart tegen Meta. De EC vermoedt dat Meta de verplichtingen om systeemrisico's te beheersen niet naleeft.[86] Zo zijn er zorgen over de verspreiding van desinformatie en de zichtbaarheid van politieke boodschappen op de feeds van gebruikers van Meta-platforms.

AI beïnvloedt de online informatiecyclus in alle stadia en via verschillende kanalen



3. Uitdagingen in democratische controle op AI-systemen

SNEL NAAR DIT ONDERDEEL

De vormgeging van het proces voor democratische sturing en controle van AI-systemen is bepalend voor de wijze waarop volksvertegenwoordigers - van Tweede Kamer tot gemeenteraad - grip kunnen hebben op AI-systemen die worden ingezet door de overheid. Deze sturing en controle moet mogelijk zijn tijdens elke fase van ontwikkeling, inzet en evaluatie van een AI-systeem. Dit hoofdstuk verkent dit onderwerp aan de hand van de situatie in het lokaal bestuur. De inzichten en aanbevelingen zijn daarbij relevant voor alle bestuurslagen.

Overheden gebruiken een divers palet aan AI-systemen om allerlei processen te automatiseren. Dat blijkt onder andere uit een verkennend onderzoek van TNO.[87]
De laatste jaren heeft automatisering met AI-systemen bij overheden ook tot incidenten en grondrechtenschendingen geleid. De lokale overheid voert veel taken uit met een grote impact op burgers en een groot deel van de AI-systemen in de publieke sector wordt door Nederlandse gemeentes ingezet.[88]

Enquêteresultaten tonen dat gemeentelijke organisaties nog maar beperkt overzicht hebben over hun AI-systemen, dat raadsleden twijfelen over de adequaatheid van hun AI-kennis en dat slechts enkele lokale rekenkamers sporadisch onderzoek doen naar AI-systemen. Gemeentes hebben behoefte aan duidelijke kaders en regels om de democratische beheersingscyclus voor AI-systemen vorm te geven. Concreet gaat het om verduidelijking van vragen als: (i) hoe verantwoordt de uitvoerende macht AI-gebruik aan de volksvertegenwoordiging? (ii) hoe kan een externe controleur, zoals een rekenkamer, AI-systemen effectief controleren? en (iii) welke vragen zouden volksvertegenwoordigers op welk moment het beste kunnen stellen over AI-systemen? Daarbij kan ook gedacht worden aan een overkoepelend AI-coördinatiecentrum en/of expertisecentra ter ondersteuning van de democratische beheersingscyclus van AI-systemen bij overheden.

3.1 Casus: Controle op AI-systemen in de lokale democratie

Gemeentes gebruiken, net als andere overheden, AI-systemen voor diverse doeleinden – dat heeft in het verleden grondrechtenschendingen met zich meegebracht. Gemeentes gebruiken AI-systemen om efficiënter hun taken uit te voeren. Bijvoorbeeld fraude voorkomen bij inschrijvingen in de Basisregistratie Personen,[89] communiceren met mensen die geen Nederlands spreken[90] of proactief helpen bij schulden.[91] En in veel gemeentes rijden inmiddels scanauto's rond die met AI-systemen controleren op parkeerovertredingen.[92] Op dit moment is bijna 75% van de geregistreerde algoritmes in het nationale Algoritmeregister afkomstig van gemeentelijke organisaties (zie grafiek 5.4 in hoofdstuk 5). Recentelijk leidde AI-gebruik door gemeentes een aantal keren tot risico's voor of inbreuk op fundamentele waarden en grondrechten. De AP heeft daar eerder op gewezen, bijvoorbeeld in het jaarverslag 2023.[93]

Dit hoofdstuk is mede gebaseerd op een enquête onder gemeentelijke organisaties, raadsleden, lokale rekenkamers en lokale ombudsorganisaties. Samen waarborgen deze organisaties idealiter een verantwoorde en democratisch verankerde inzet van AI-systemen, met zicht op risico's en incidenten. Zodat gemeentes AI-technologie inzetten op een manier die bijdraagt aan fundamentele waarden en de bescherming van grondrechten. In totaal is de survey beantwoord door 85 gemeentelijke organisaties, 35 rekenkamers en 27 raadsleden. Het aantal lokale ombudsorganisaties dat in staat bleek om op de enquête te reageren, was minimaal. De enquête is door de AP opgezet vanuit de nieuwe coördinerende algoritmetaak en is uitgevoerd in maart en april 2024.

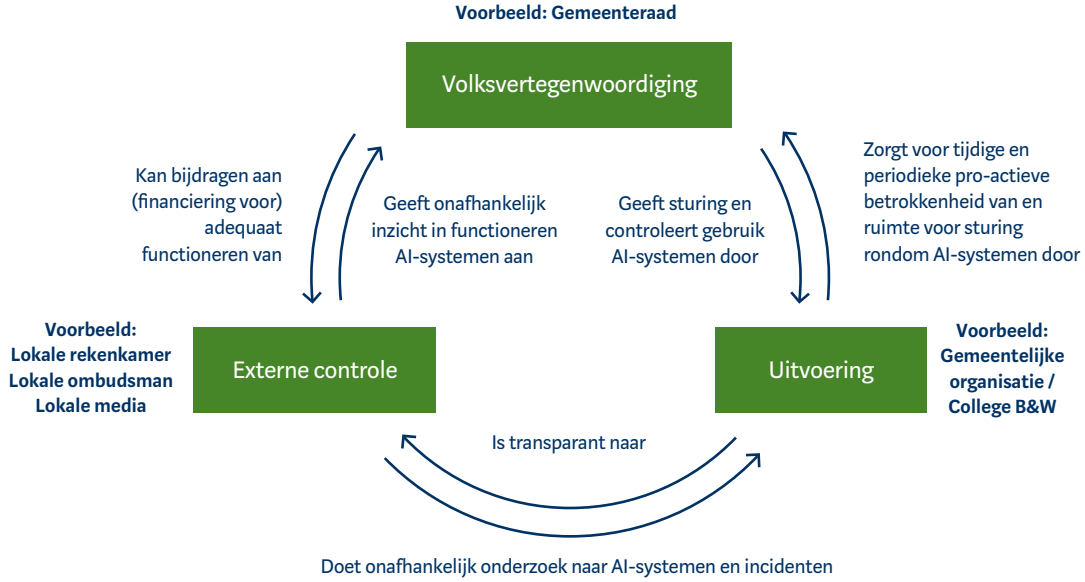
3.2 De democratische cyclus voor AI-systemen

Het gebruik van AI-systemen in de publieke sector moet onderdeel zijn van een democratische cyclus van (aan) sturing en verantwoording. Dit geldt op landelijk, regionaal en lokaal niveau. Volksvertegenwoordigers behoren de uitvoering van overheidstaken te sturen én te controleren. Dit omvat dus ook het gebruik van AI-systemen in die uitvoering. Daarbij kunnen ze bouwen op het werk van andere instituties, zoals de media of toezichthouders. Volksvertegenwoordigers geven kaders mee aan een uitvoerende overheid. Hierin kan ook worden aangeven of en zo ja, hoe de overheid AI-systemen kan gebruiken. De overheid ontwikkelt binnen die kaders een AI-beleid en voert dat uit. De volksvertegenwoordigers beoordelen vervolgens het AI-gebruik van de overheid en stellen de beleidskaders bij op basis van hun oordeel. Zo ontstaat idealiter een democratische cyclus van (aan)sturing en verantwoording van het gebruik van AI-systemen door overheden (zie figuur 3.1).

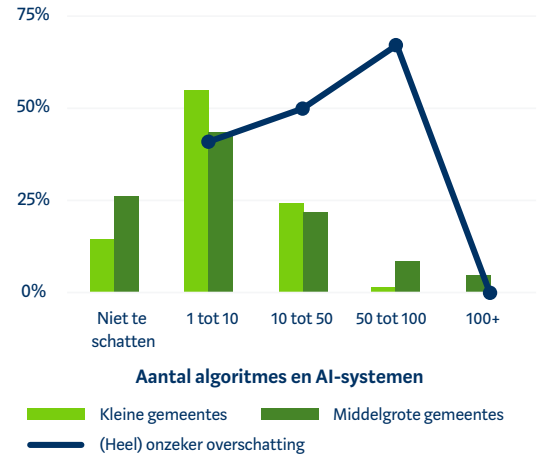
Gemeentelijke organisaties krijgen kaders voor hun beleid van de gemeenteraad en leggen via het college van burgemeester en wethouders ook verantwoording aan de raad af. De gemeenteraad heeft een kaderstellende én een controlerende taak, naast de taak om de inwoners te vertegenwoordigen.[94] Als het college verantwoording aflegt over hoe het AI-systemen gebruikt, kan de gemeenteraad daar vanuit de controlerende taak van de raad over oordelen.

Lokale rekenkamers, lokale ombudspersonen en lokale media controleren het college ook en kunnen de kaderstelling door de raad voeden. Rekenkamers onderzoeken bijvoorbeeld onafhankelijk de doeltreffendheid, doelmatigheid en rechtmatigheid van door een gemeente gevoerd beleid.[95] Als zij over de rechtmatigheid van gemeentelijk gebruik van AI-systemen rapporteren, kan de gemeenteraad daarmee het college beoordelen. Lokale ombudspersonen helpen burgers met klachten tegen het lokale bestuur. Het doel is burgers te beschermen tegen handelen door de gemeente dat hen schaadt. Daarvoor heeft de ombudspersoon uitgebreide onderzoeksbevoegdheden.[96] Uitkomsten van onderzoeken door de ombudspersonen kunnen eveneens worden gebruikt om het college te beoordelen en bij te sturen. Lokale media kunnen het college openlijk kritisch bevragen en kunnen maatschappelijke kwesties agenderen.[97] Ook daar kan de gemeenteraad gebruik van maken bij de beoordeling en bijsturing van het college.

FIGUUR 3.1: CONCEPTUELE DEMOCRATISCHE BEHEERSINGSCYCLUS VOOR AI-SYSTEMEN



FIGUUR 3.2: VEEL GEMEENTES SCHATTEN IN DAT ZIJ SLECHTS EEN BEPERKT AANTAL ALGORITMES EN AI-SYSTEMEN GEBRUIKEN, MAAR ZIJN HIER WEL ONZEKER OVER



BRON: EIGEN ONDERZOEK OP BASIS VAN SURVEY ONDER GEMEENTELIJKE ORGANISATIES (N=85)

3.3 Uitdagingen bij gebruik AI-systemen door gemeenten

Veel gemeentes veronderstellen maar een beperkt aantal algoritmes en AI-systemen te gebruiken, maar zijn wel onzeker over deze inschatting. Uit de enquête onder gemeentelijke organisaties blijkt dat meer dan de helft van de kleinere gemeentes denkt hoogstens 10 algoritmes en AI-systemen te gebruiken (zie figuur 3.2). De inschattingen zijn echter divers: meerdere kleine gemeentes denken meer dan 100 algoritmes en AI-systemen te gebruiken. De twijfel is groot, zeker bij kleinere en middelgrote gemeentes die 50 tot 100 algoritmes gebruiken. Bijna driekwart geeft aan dat de schatting onzeker of heel onzeker is. Die onzekerheid geeft aan dat gemeentes nog stappen moeten zetten om overzicht te krijgen. Dat is een voorwaarde voor risicobeheersing.

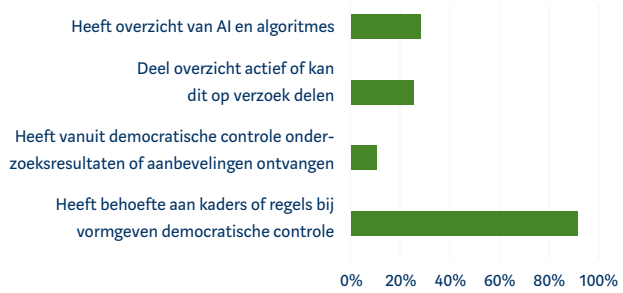
De huidige onzekerheid is begrijpelijk, want gemeentelijke organisaties kampen vanwege hun schaalgrootte met uitdagingen op het gebied van kennis, expertise en middelen. De gemeentelijke uitdagingen maken enerzijds de keuze om AI-systemen te gebruiken aantrekkelijk, maar bemoeilijken anderzijds het verantwoord inzetten van diezelfde AI-systemen. Ambtenaren hebben bijvoorbeeld kennis nodig om een ingekocht AI-systeem op waarde te kunnen schatten en de impact ervan te kunnen overzien. Als voorbeeld wordt in de enquête door een gemeente aangegeven dat het "in het verleden niet altijd bekend was – als wij een applicatie in gebruik namen – of hierbij [wel of niet] sprake was van AI of algoritmegebruik." Om hier een stap in te zetten, moeten gemeentes werken aan voldoende AI-geletterdheid (zoals voorgeschreven door de AI-verordening per 1 februari 2025, zie hoofdstuk 5). Beperkte middelen en een uitdagende arbeidsmarkt maken dat echter moeilijk. Expertise inkopen of inhuren is duur, net als het trainen van de ambtenaren die een toepassing dagelijks moeten inzetten en beheersen.

Gemeentes zijn zich niet altijd bewust (geweest) van de impact en het politieke karakter van keuzes in het gebruik van AI-systemen. Dat is een probleem, vooral omdat gemeentes AI-systemen gebruiken in processen die mensen in kwetsbare posities treffen, bijvoorbeeld in het sociaal domein.[98] Gemeentes zagen de toepassing van AI-systemen tot voor kort primair als een technische kwestie en een efficiëntiemaatregel. Bovendien is het voor (kleine) gemeentelijke organisaties gegeven hun omvang moeilijk om aan de juiste informatie over een AI-systeem te komen. Daardoor blijven de implicaties van de inzet van een AI-systeem voor grondrechten en fundamentele waarden buiten zicht.[99] Zoals verwoord door een respondent:

> "Er is een sterke behoefte aan een onafhankelijke certificering voor algoritmes en AI. De meeste organisaties zijn afhankelijk van leveranciers ('wij van wc-eend') voor het inwinnen van informatie over een algoritme of AI en niet in staat om de achterliggende trainingsdataset en het algoritme of de AI technisch voldoende te toetsen op onderwerpen als bias."

Gemeentes die grondrechten bij hun gebruik van AI-systemen wel in zicht hebben, focussen voornamelijk op het recht op gegevensbescherming.[100]

FIGUUR 3.3: GEMEENTELIJKE ORGANISATIES HEBBEN SLECHTS BEPERKT INTERACTIE MET GEMEENTERAAD, REKENKAMERS EN OMBUDSPERSONEN OVER AI EN ALGORITMES



Toelichting: *28% van de gemeentes weet zeker een overzicht te hebben van algoritmes en AI-systemen in de gemeente. 26% deelt dit overzicht minstens 1x per jaar, deelt het op verzoek of is in staat het overzicht te delen. 11% heeft in de afgelopen drie jaar aanbevelingen of onderzoeksresultaten ontvangen over algoritmes en AI vanuit controlerende organen. 92% van de gemeentes zegt behoefte te hebben aan kaders of regels voor hoe zij de democratische controle op algoritmes en AI (verder) vorm kunnen geven. Bron: enquêteresultaten (n=85)*

Gemeentelijke organisaties delen informatie over AI-systemen nog maar mondjesmaat met de gemeenteraad, rekenkamers en ombudspersonen. Uit de enquête blijkt dat iets meer dan 20% van de gemeentes een overzicht heeft van AI en algoritmes. Eveneens iets meer dan 20% deelt dit overzicht ook actief, of op verzoek, met de andere partijen in de democratische cyclus. Een kleiner deel (ongeveer 10%) heeft vervolgens ook vanuit de gemeenteraad, of controleurs zoals de rekenkamer, resultaten van onderzoek naar of adviezen over AI-systemen ontvangen (zie figuur 3.3). In de toelichting geeft een gemeentelijke organisatie bijvoorbeeld aan betrokkenheid van gemeenteraden of burgers ook niet als vanzelfsprekendheid te zien: "Waarom [worden] er vragen gesteld of burgers en/of [gemeente]raad betrokken worden? Is dit straks een wettelijke verplichting?". Een mogelijke best practice voor gemeentelijke organisaties is de

overweging om informatie over AI-systemen in te bedden in de reguliere informatievoorziening. Zo geeft een respondent aan dat betrokken algoritmes worden benoemd in besluitvormende stukken naar de gemeenteraad:

> "In het raadsvoorstel nemen we een alinea op over het gebruikte algoritme om tot dit voorstel te komen (beleidsvorming), of om dit voorstel uit te voeren (toezicht en handhaving)."

Dit sluit aan bij de observatie dat het voor gemeentes nog altijd een uitdaging is het eigen gebruik van AI-systemen overzichtelijk te organiseren. Verantwoord gebruik is pas mogelijk als er een overzicht is welke AI-systemen de eigen organisatie gebruikt en wie daarvoor verantwoordelijk is. Uit onderzoek in opdracht van het College voor de Rechten van de Mens bleken de verantwoordelijkheden voor AI-systemen in sommige gemeentes echter zo onoverzichtelijk te zijn, dat de onderzoekers het een 'bestuurlijke spaghetti' noemen:[101] een lastig te ontwarren kluwen. Sommige gemeentelijke organisaties laten zich ook nog afschrikken door een brede interpretatie van het concept 'algoritmes' en gesuggereerde juridische belemmeringen. Zo antwoord een respondent dat "elke geautomatiseerde toepassing een stelsel van algoritmes [bevat]. Het is niet te doen deze in kaart te brengen, nog afgezien van het feit dat op de broncode auteursrecht geldt." Het moge duidelijk zijn dat het bij het in kaart brengen van AI-systemen moet gaan om systemen die via hun output impact hebben, bijvoorbeeld door voorspellingen te doen, inhoud te genereren, aanbevelingen te doen of beslissingen te nemen. Vandaar het belang van de brede, maar gerichte definitie van AI-systemen (zie de inleiding van deze RAN).

De behoefte onder gemeentelijke organisaties aan kaders of regels voor het vormgeven van de democratische controle op AI-systemen is groot. In de enquête onderschrijft meer dan 90% van de gemeentelijke organisaties behoefte te hebben aan dergelijke kaders (zie figuur 3.3). In toelichtende vragen geven gemeentelijke organisaties aan dat zij vooral behoefte hebben aan concrete kaders met weinig interpretatieruimte en juist veel zekerheden. Een respondent geeft bijvoorbeeld aan dat het "nadeel tot nu toe is dat alle kaders op dit vlak ontzettend vaag zijn qua definities of de plank misslaan op dit vlak". Een andere respondent geeft aan dat "het afwegingskader alsook de inschattingen die gemaakt moeten worden om algoritmes in te delen vrij complex [zijn] en bovenal subjectief". Een behoefte uitgesproken door nog een andere respondent is het delen van best practices waar gemeentes mee aan de slag kunnen.

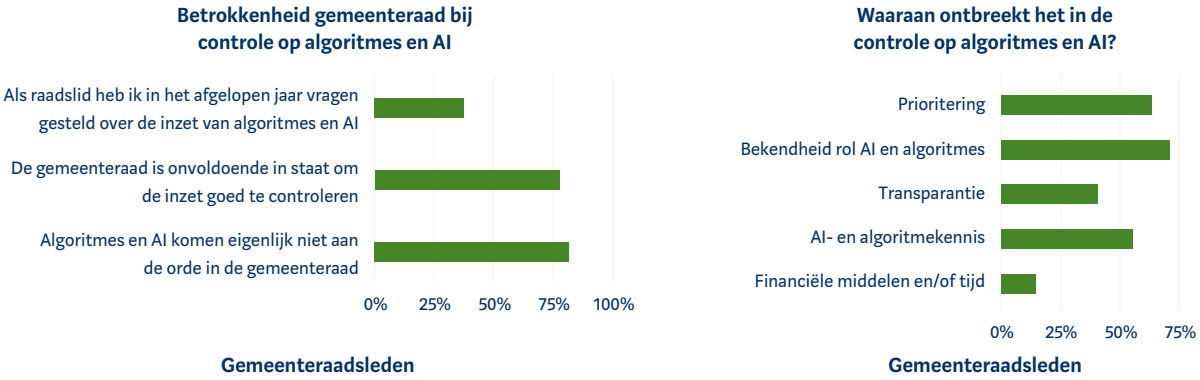
3.4 Vergelijkbare uitdagingen voor sturende en controlerende partijen

> "In gemeenteraden komen algoritmes en AI eigenlijk niet aan de orde."

Met die stelling is driekwart van de raadsleden die de enquête hebben beantwoord het eens. Eveneens driekwart van de raadsleden vindt daarom dat de raad onvoldoende in staat is om het gebruik van algoritmes en AI te controleren (zie figuur 3.4).

Dat komt door gebrek aan prioritering, aan bekendheid met de werking van AI-systemen en aan AI-geletterdheid. Meer dan de helft van de raadsleden die de enquête hebben beantwoord, ziet dit als knelpunten. Hier spelen volgens de raadsleden verschillende aandachtspunten.

FIGUUR 3.4: VOLGENS RAADSLEDEN KOMEN ALGORITMES EN AI BEPERKT AAN DE ORDE IN DE GEMEENTERAAD – HET ONTBREEKT VOORAL AAN PRIORITERING EN BEKENDHEID



BRON: EIGEN ONDERZOEK OP BASIS VAN SURVEY ONDER GEMEENTERAADSLEDEN (N=27)

Zo geeft een raadslid aan "[dat] als het niet ter sprake komt, het ook niet te controleren is". Een ander raadslid geeft aan dat "er ontzettend veel gebeurt en niemand lijkt het overzicht te hebben. Dat betekent dat we onze controlerende taak niet kunnen uitvoeren." Het gebrek aan diepgravende aandacht is een zorgpunt, omdat het ook in de weg kan staan van waardevolle toepassingen. In de woorden van een raadslid:

> "Raadsleden weten niets van AI. Ze vinden het eng en [dat] zorgt voor veel weerstand."

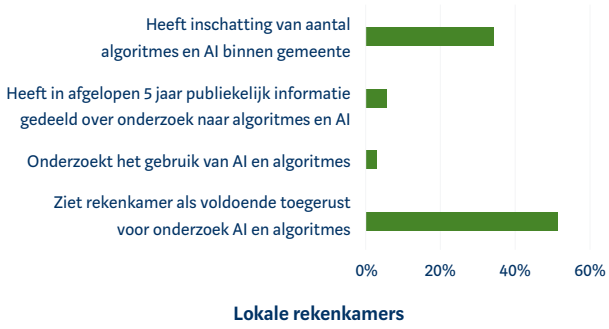
Gemeenteraden zijn zich ook niet altijd bewust van de impact en het politieke karakter van keuzes bij het gebruik van AI-systemen. Daardoor benutten raden vaak niet de kans om kaders te stellen voor en democratische controle uit te oefenen op AI-systemen bij de gemeente. Dat stelt het Rathenau Instituut op basis van een eigen onderzoek uit 2020. Het Rathenau Instituut constateerde hierbij dat raadsleden te weinig bewust leken te zijn van de impact van digitale technologie. Raden bespreken digitalisering volgens het Rathenau Instituut zelden als een onderwerp waarover politieke en ethische keuzes gemaakt kunnen worden.[102] De Raad voor het Openbaar Bestuur (ROB) concludeerde ook al dat gebruik van AI-systemen te weinig als een morele en politieke kwestie wordt benaderd. De ROB merkte ook op dat het bewustzijn over AI-systemen wel toeneemt.[103] De enquête uitgevoerd voor dit hoofdstuk geeft op punten nog altijd een vergelijkbaar beeld. Zo geeft een raadslid aan dat "uitvoering van het beleid wordt weggezet bij het ambtelijk apparaat. In die zin is het een competentie van de directeur. Die is enkel aan te spreken op het goed functioneren van het ambtelijk apparaat en niet direct op de instrumenten die daarvoor worden ingezet."

Kennis en expertise op het gebied van AI-technologie is een uitdaging voor volksvertegenwoordigers. Dat maakt het moeilijk om te beseffen welke vragen zij bij gebruik van AI-systemen door de gemeente moeten stellen om het college te controleren. Maar juist ook voor de kaderstellende taak hebben volksvertegenwoordigers kennis nodig. Het Rathenau Instituut stelde in 2020 vast dat raadsleden zichzelf vaak onvoldoende bekwaam achten om over digitaliseringskwesties te oordelen.[104] De enquête bevestigt dit beeld. Zo antwoord een raadslid dat er "te weinig kennis [is] om goede vragen te stellen." Praktische oplossingen daarvoor zijn denkbaar. Een ander raadslid vraagt bijvoorbeeld of "er ergens een voorbeeldlijst van vragen beschikbaar [is] die we kunnen gebruiken om gericht vragen te stellen?".

Beperkte financiering van lokale democratische instituties maakt effectieve controle en kaderstelling moeilijker, bijvoorbeeld voor raadsleden. Voldoende financiering van de gemeenteraad is belangrijk vanwege het parttime karakter van het raadslidmaatschap en de beperkte omvang van gemeenteraden, waardoor specialisatie onder raadsleden beperkt mogelijk is. Daarom zijn raadsleden gebaat bij hulp van bijvoorbeeld fractieondersteuning. Het gemiddelde jaarbudget hiervoor is echter zeer beperkt.[105]

Rekenkamers zijn competent maar kunnen maar beperkt onderzoek doen naar algoritmes en AI. Minder dan een op de tien lokale rekenkamers heeft in de afgelopen vijf jaar onderzoek gedaan naar het gebruik van AI en algoritmes en publiek informatie gedeeld over dit onderzoek. Meer dan de helft van de lokale rekenkamers acht zich op dit onderwerp wel competent (en ziet het onderwerp daarmee als onderdeel van het mandaat) – zie figuur 3.5. Prominent zichtbaar zijn de activiteiten van de rekenkamers van de grootste gemeenten, zoals Amsterdam, Den Haag en Rotterdam. De Rekenkamer Metropool Amsterdam publiceerde in oktober 2023 een onderzoek naar de toepassing van algoritmes in Amsterdam.[106] De Rekenkamer Rotterdam publiceerde in maart 2024 een vervolgonderzoek naar het gebruik van algoritmes.[107] In maart 2024 heeft de Rekenkamer Den Haag bekendgemaakt te starten met een verkenning naar het gebruik van algoritmes in Den Haag om tot een volwaardige onderzoeksopzet te komen.[108] De werkwijze van deze lokale rekenkamers in grote gemeentes kan, via best practices, inspiratie bieden voor lokale rekenkamers in kleinere gemeentes.

FIGUUR 3.5: WEINIG AANDACHT VAN LOKALE REKENKAMERS VOOR ALGORITMES EN AI



Toelichting: *Waar een lokale rekenkamer verantwoordelijk is voor meerdere gemeentes, gaat het antwoord op de eerste vraag over de grootste gemeente.*

BRON: EIGEN ONDERZOEK OP BASIS VAN SURVEY ONDER LOKALE REKENKAMERS (N=35)

Meer dan 90% van de lokale rekenkamers weet nog niet of zij algoritmes en AI in de komende jaren wél gaan meenemen. Dit geven lokale rekenkamers aan in respons op de enquête. Deels komt dit omdat veel rekenkamers – gegeven de nieuwe Wet ter versterking van decentrale rekenkamers – in de huidige vorm nog maar kort bestaan. Tegelijkertijd geven de meeste lokale rekenkamers aan wel open te staan voor het onderwerp. In antwoord op de enquête geven rekenkamers aan het een "interessant onderwerp" te vinden, waarbij zij "willen bespreken of dit onderwerp zich leent voor een onderzoek" en waarbij zij er ook rekening mee houden "dat dit onderzoek regelmatig herhaald moet worden".

Het gebrek aan middelen beperkt de mogelijkheden voor rekenkamers, ombudspersonen en lokale media om hun controlerende taak uit te oefenen. Veel lokale rekenkamers hebben budget voor een of twee onderzoeken per jaar. Een respondent op de enquête maakt het concreet: "lokale rekenkamers, zeker van kleine gemeenten [hebben] een zeer beperkte tijd en capaciteit, onderzoeken als deze [over algoritmes en AI] bevinden zich buiten onze mogelijkheden. De inzet van bestuursleden en totale ondersteuning omvat maximaal een of twee dagen per week. Het budget voor alle onderzoeken is 30.000 euro per jaar." Ook de basisfinanciering van lokale omroepen schiet tekort volgens zowel het Stimuleringsfonds voor de Journalistiek[110] als de Raden voor Openbaar Bestuur en cultuur. Ook lokale private media hebben weinig geld en komen daardoor niet toe aan hun rol als controleur van de gemeente.

3.5 Inrichting decentrale overheid en rol van kaders

Een complexe lokale organisatie rondom AI-systemen bemoeilijkt democratische controle. Voor veel uitvoeringskwesties kiezen colleges tegenwoordig voor samenwerkingen met andere gemeenten.[110] Of voor het uitbesteden van taken aan marktpartijen. Ook de inzet van AI-systemen wordt vaak (deels) uitbesteed.[111] Een ondoorzichtige verantwoordelijkheidsverdeling maakt het niet alleen moeilijk voor gemeenten om hun eigen AI-systemen te beheersen. Het maakt het voor controlerende partijen, zoals de gemeenteraad, ook moeilijker om gemeentelijk gebruik van AI-systemen op waarde te schatten en bij te sturen. Dat doet af aan de democratische legitimiteit van gemeentelijk gebruik van AI-systemen.

De inrichting van het lokale bestuur zorgt ook voor hindernissen bij controle en correctie van gemeentelijk gebruik van AI-systemen. De financiering van raad, rekenkamer, lokale omroep en ombudspersoon stelt de raad voor lastige dilemma's bij het vaststellen van het gemeentebudget.[112] Veel gemeenteraden hebben een krap budget.[113] Kiest de raad voor een uitbreiding van het budget van de raadsfracties, de lokale omroep, de ombudspersoon of de rekenkamer? Dan gaat dat in de huidige situatie ten koste van de budgetten van diensten die de inwoners direct ten goede komen. Via nationale wet- en regelgeving kan wel sturing worden gegeven aan hoe het systeem van controle en correctie op lokaal niveau werkt. Een voorbeeld is de Wet versterking decentrale rekenkamers die op 1 januari 2023 in werking is getreden.

Regelgeving biedt houvast bij de inzet van AI-systemen in het publieke domein. De aankomende AI-verordening brengt in aanvulling op bestaande regelgeving, zoals de AVG, verplichtingen mee voor veel AI-systemen in de publieke sector. Inhoudelijk gaat het vooral om een verankering van beheersingsmaatregelen die toch al aan te raden of onvermijdelijk waren. Bijvoorbeeld dat het verplicht wordt een risicomanagementsysteem te hebben, risico's te monitoren en AI-systemen te registreren. Het is verstandig om zo snel mogelijk gebruik te maken van het houvast die de AI-verordening biedt. Zie de bijlage van deze RAN voor een indicatie van de wijze waarop dit beheersingsraamwerk vorm krijgt.

3.6 Aanbevelingen

Landelijke politici en beleidsmakers kunnen lokaal bestuur helpen door decentrale overheden te ondersteunen met voldoende middelen, kennis en flexibele uitvoerende capaciteit voor beheersing en controle van AI. Naar analogie gelden deze aanbevelingen voor de democratische controle op landelijk niveau, ook in relatie tot de inzet van AI-systemen bij uitvoeringsorganisaties en andere bestuursorganen. Automatiseren van handelingen via AI-systemen lijkt vaak een aantrekkelijke optie, gelet op personeelstekorten, de behoefte aan efficiëntie en kostenbesparingsoperaties. Maar de AI-paradox is dat daarvoor eerst flink geïnvesteerd moet worden in een infrastructuur voor beheersing van AI, personeel dat voldoende opgeleid is en genoeg personeel voor menselijke controle van AI-systemen. Gerichte eisen, financiering en ondersteuning vanuit nationaal niveau kunnen hieraan bijdragen.

Daarbij is mogelijk een belangrijke rol weggelegd voor een overkoepelend AI-coördinatiecentrum en/of expertisecentra, waarop lokale overheden een beroep kunnen doen bij de beheersing van en controle op AI-systemen. Over de mogelijkheden voor de verdeling van kennis en expertise zijn al adviezen verschenen. De WRR adviseerde in 2021 bijvoorbeeld een beleidsinfrastructuur voor AI op te zetten, te beginnen met een nationaal AI-coördinatiecentrum. Dat zou onder andere het leervermogen van overheden bij AI-systemen vergroten.[114] Het Britse Alan Turing Institute zou daarvoor inspiratie kunnen bieden (zie box 3.1). In de afgelopen jaren zijn veel (nieuwe) AI-systemen, bijvoorbeeld voor generatieve AI, alleen maar complexer geworden. Dit vergroot de meerwaarde van nationale en regionale structuren om de inzet van AI te ondersteunen. Denk aan expertisecentra die gemeentelijke organisaties en lokale rekenkamers structureel kunnen helpen om AI-systemen te beheersen en controleren. Gezien de technische complexiteit van AI-systemen moet het niet zo zijn dat organisaties het wiel telkens opnieuw moeten uitvinden.

Gemeentelijke organisaties kunnen hierbij de raad en andere controlerende instituties op weg helpen door heldere AI-governance en door de raad proactief te betrekken. Een overzicht over het eigen AI-gebruik is de basis. Dit stelt de raad, rekenkamer, ombudspersoon en lokale media beter in staat om het AI-gebruik binnen de gemeente kritisch te bekijken en eventueel bij te sturen. Want door zo'n overzicht wordt het bijvoorbeeld makkelijker om de raad te informeren en krijgen raadsleden beter gestructureerde informatie. Het is daarbij de verantwoordelijkheid van het college om de democratische cyclus verder te bevorderen door de raad proactief en vroeg te betrekken bij vraagstukken over AI binnen de gemeente. Daarbij zou het college keuzes over AI-systemen kunnen presenteren als keuzes waarbij democratische zeggenschap gewenst is, omdat deze keuzes impact kunnen hebben op fundamentele waarden en grondrechten. Zie daartoe ook de bijlage bij deze RAN, inclusief het accent op expliciete doelbepaling en een afgewogen besluit tot (het verkennen van) de inzet van een AI-systeem.

Het ministerie van Binnenlandse Zaken en Koninkrijksrelaties is een algoritmekader aan het ontwikkelen. Dat is op dit moment het belangrijkste instrument waaraan wordt gewerkt om overheden te ondersteunen. De staatssecretaris voor Digitale Zaken en Koninkrijksrelaties heeft eind juni 2024 aangegeven dat het algoritmekader een overzicht moet bieden van de belangrijkste eisen aan het gebruik van algoritmes en AI-systemen. Daarbij wordt het kader op een open source-manier ontwikkeld, in brede werkgroepen waaraan interbestuurlijke partijen deelnemen. De intentie is daarbij het kader te laten gelden voor de gehele overheid. De staatssecretaris geeft aan dat de controlerende organen, zoals auditdiensten en rekenkamers, toezien op de normstelling.[115] De AP merkt daarbij op dat het kader zo concreet mogelijk moet zijn om overheidsorganisaties te helpen. Daarnaast is het belangrijk dat het kader niet alleen gaat over (i) de eisen aan de (kwaliteits)beheersing van individuele AI-systemen, maar ook over (ii) eisen aan de AI-governance van overheidsorganisaties en (iii) hoe de democratische beheersingscyclus wordt vormgegeven. Bijvoorbeeld door basisinformatie te bieden die volksvertegenwoordigers en externe organen kunnen gebruiken bij hun sturende en controlerende taken. De AP wijst er daarbij op dat het cruciaal is om het algoritmekader zo dicht mogelijk te laten aansluiten op de (bindende) eisen uit Europese regelgeving, zoals de AI-verordening en de AVG. In de tweede helft van dit jaar geeft de AP vanuit de coördinerende algoritmetaak een nadere analyse van het algoritmekader in ontwikkeling.

Controleverplichtingen versterken specifiek voor AI-systemen, kan het lerend vermogen vergroten. Dit kan bijvoorbeeld door publieke AI-systemen met impact – van simpele algoritmes tot complexe modellen – verplicht te laten controleren d.m.v. audits. De uitkomsten kunnen een goede basis vormen voor het uitvoeren van de controletaak binnen gemeentes. Een professionele partij moet deze audits dan volgens heldere criteria uitvoeren. Vooraf opgestelde kaders kunnen meegenomen worden in het auditeren van publieke AI-systemen, om zo de gemeenteraad en andere partijen in het democratische controlebestel optimaal van informatie te voorzien. Het invoeren van een wettelijke auditplicht die de controlerende partijen versterkt en optimaal van kennis en informatie voorziet kan bijdragen aan het lerend vermogen van gemeentes.

Box 3.1

Ondersteunen van decentrale overheden via AI-kennisinstituten – ervaring in het VK

Het Alan Turing Institute (ATI) is het Britse nationale instituut voor data science en is hét adres voor publieke organisaties in het VK om aan te kloppen met datasciencevragen. Het instituut richt zich op het vergaren van kennis en het benutten daarvan in projecten op het gebied van data science en AI. Zodoende adviseert het ATI onder andere de publieke sector op het gebied van data science. Publieke organisaties hebben door het ATI één instituut om aan te kloppen voor kennis en bijstand bij complexe AI-vraagstukken. Het ATI wordt voor een belangrijk deel gefinancierd door overheidsfondsen voor onderzoek en innovatie.

Het Londense district Camden klopte bij het ATI aan om samen met inwoners een visie op datagebruik te ontwikkelen. Dat deed het district in 2021. Camden wilde een visie ontwikkelen om ethischer en doordachter met persoonsgegevens om te gaan en deze verantwoord te gebruiken in algoritmes. Tijdens dit proces was Camden voortdurend met inwoners in gesprek. Hun ideeën en bijdragen, verzameld met een enquête en inputsessies, vormden de basis voor de uiteindelijke visie.

De externe expertise van het ATI droeg bij aan het niveau van de maatschappelijke discussie en het uiteindelijke beleid. Het ATI hielp om inwoners of andere leken goed te informeren en een onderhoudend debat op gang te brengen. Op die manier kregen overheden en inwoners de kans om van elkaar te leren. En konden zij toewerken naar een vorm van overeenstemming en een breed gedragen aanpak. De multidisciplinaire benadering van het ATI droeg bovendien bij aan het stellen van zoveel mogelijk relevante vragen.

Zie voor meer informatie de website van het Alan Turing Institute. Waaronder ook informatie over het Camden Council project.

4. Profilerende en selecterende AI-systemen: risico's en de aselecte steekproef

SNEL NAAR DIT ONDERDEEL

Veel organisaties gebruiken algoritmes voor profilering of soortgelijke processen waarbij onderscheid tussen mensen wordt gemaakt. Dit hoofdstuk verkent dit onderwerp aan de hand van voorbeelden op het gebied van fraudedetectie. Belangrijk is deze algoritmes altijd als AI-systeem en daarmee onderdeel van een breder proces te zien. Het risico op discriminatie is een terugkerend thema in deze context. Op verschillende plekken in het proces rondom een profilerend AI-systeem kan discriminatie ontstaan. Bijvoorbeeld door niet-representatieve data en overmatig vertrouwen op algoritmische uitkomsten.
Een aselecte steekproef kan worden ingezet als beheersingsinstrument. Voordelen van deze techniek zijn dat algoritmes beter gemonitord kunnen worden en dat deze techniek waarborgt dat er een menselijke beslissing is in het proces. Het aanvullen van fraude-algoritmeprocessen met een aselecte steekproef in is in veel gevallen dan ook aan te bevelen.

4.1 Risicoprofilering en selectie

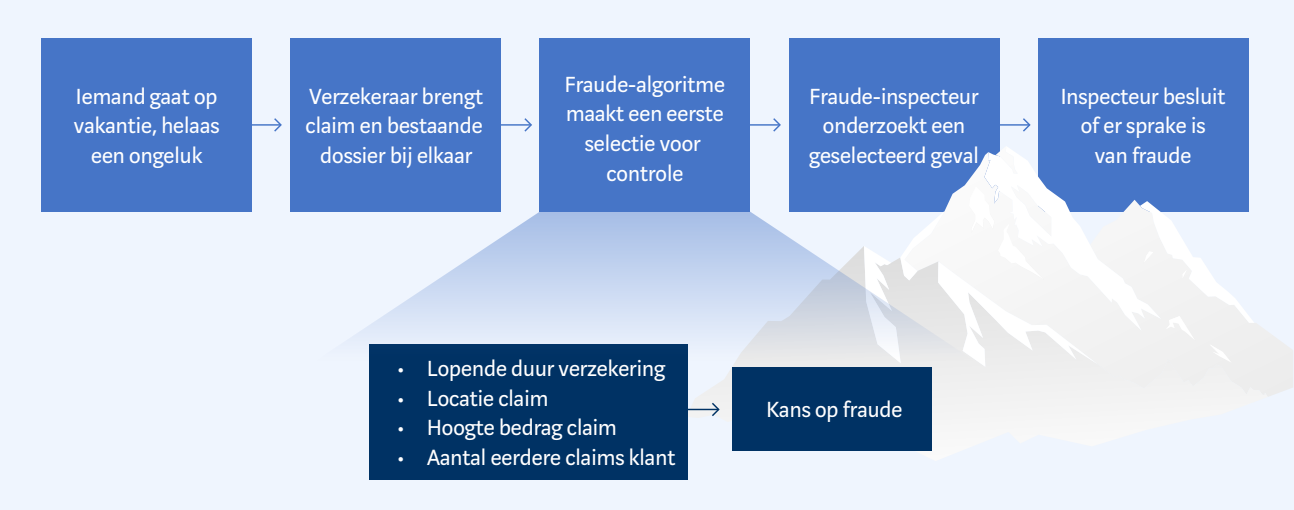
Veel organisaties zetten algoritmes in om te profileren en te selecteren. Op basis van gevallen uit het verleden maken organisaties een inschatting. Hiermee kunnen zij actie ondernemen, zoals een gericht onderzoek naar bepaalde personen. De inschatting dient dan als selectiemiddel om personen te selecteren die een inspecteur gaat onderzoeken.

Een voorbeeld van profilerende en selecterende systemen zijn fraudealgoritmes, waar vrijwel iedereen mee in aanraking komt. Verzekeraars zoeken met algoritmes naar verzekeringsfraude,[116] banken gebruiken fraudemodellen om transacties te controleren[117] en online platforms zoeken naar fraude binnen nieuwe gebruikersaccounts.[118] Vaak weet een burger, klant of gebruiker niet dat er een controle op fraude plaatsvindt. Zolang er geen verdenking is van fraude, blijft het fraudealgoritme een onzichtbare processtap.

“Het ontbreken van waarborgen in de uitvoeringspraktijk van risicogericht toezicht en de schending van wet- en regelgeving, hebben als gevolg dat sommige mensen meer kans hebben dan andere mensen om in beeld te komen bij de overheid en te worden gecontroleerd in het kader van fraudebestrijding.”
Parlementaire enquêtecommissie: Blind voor mens en recht, 7.1

Fouten bij fraudealgoritmes kunnen grote gevolgen hebben voor personen. Naast rechtmatigheidsvraag-stukken– mag een dergelijk algoritme in bepaalde situatie ingezet worden en mogen bepaalde indicatoren gebruikt worden – is het een essentieel aandachtspunt dat fouten in deze algoritmes grote gevolgen hebben. Dit is te zien in de toeslagenaffaire. Uit onderzoek van de Autoriteit Persoonsgegevens (AP) bleek dat in fraudeopsporing bij de kinderopvangtoeslag onrechtmatig gebruik is gemaakt van gegevens over dubbele nationaliteit en dat het een discri-minatoire verwerking van persoonsgegevens betrof.[119] De parlementaire enquêtecommissie bevestigde de schending van grondrechten en besloot dat hier sprake was van een schending van de eerbiediging van de persoonlijke levens-sfeer en op gelijke behandeling.[120] Een tweede voorbeeld is het Systeem Risico Indicatie (SyRI). SyRI werd ingezet om socialezekerheidsfraude te detecteren. De rechter oordeelde dat de methode in strijd is met het recht op respect voor privé- en familieleven. De rechtbank woog hierbij mee dat het systeem onvoldoende inzichtelijk en controleerbaar was, terwijl de inzet van SyRI (onbedoeld) wel discriminerende effecten met zich kon meebrengen.[121] Een derde en recenter

FIGUUR 4.1: EEN FRAUDEALGORITME IS ALTIJD ONDERDEEL VAN EEN BREDER PROCES. IN DEZE WEERGAVE IS HET FRAUDE-ALGORITME EEN VAN DE STAPPEN IN DE VERWERKING BIJ EEN SCHADECLAIM VOOR EEN REISVERZEKERING



voorbeeld is de controle op fraude met de uitwonende beurs voor studenten, waarbij DUO een selectiealgoritme gebruikte. In een onderzoeksrapport werd geconcludeerd dat er sprake was van discriminatie, waar het kabinet en DUO vervolgens hun excuses voor aanboden.[122] Dit komt in het geval van deze casus bovenop de discriminatie die samenhangt met het gebruik van indicatoren als opleidingsniveau als onderscheidende factor voor frauderisico.[123]

4.2 Discriminatie en overmatig vertrouwen

Risico's bij de inzet van fraudealgoritmes ontstaan vaak in de processtappen rondom het algoritme. De uitkomsten van een algoritme hangen af van de processtappen ervoor en de uiteindelijke impact hangt af van hoe er met de uitkomsten wordt omgegaan. Om dit te illustreren is een fictief voorbeeld van een fraudecontroleproces van een reisverzekeraar schematisch weergegeven.

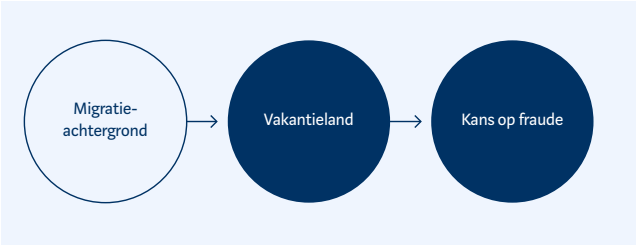
Het algoritme is hier afhankelijk van de ingediende claim en bestaande databronnen. Vervolgens dient het algoritme hier als een voorselectie in het fraudecontroleproces: een inspecteur onderzoekt vervolgens de gevallen met een hoge risico-indicatie (zie figuur 4.1).

Discriminatie is een belangrijk risico bij de inzet van fraudealgoritmes. Het logische doel van een algoritme is om onderscheid te maken, daarbij moet wel altijd bekeken worden of het maken van onderscheid juridisch wel te rechtvaardigen is. De selectie moet fraudeurs bevatten en juist géén mensen die niet frauderen. De term 'discriminatie' refereert hier aan artikel 21 van het Handvest van de Grondrechten van de Europese Unie, waarin staat: "Elke discriminatie, met name op grond van geslacht, ras, kleur, etnische of sociale afkomst, genetische kenmerken, taal, godsdienst of overtuigingen, politieke of andere denkbeelden, het behoren tot een nationale minderheid, vermogen, geboorte, een handicap, leeftijd of seksuele geaardheid, is verboden."[124] Een duidelijk voorbeeld van discriminatie is het benadelen van een kandidaat in een sollicitatieprocedure op basis van geslacht of nationaliteit.[125] Discriminatie kan zowel direct als indirect plaatsvinden. Bij indirecte discriminatie wordt er geen rechtstreeks onderscheid gemaakt op basis van een verboden grond. Bijvoorbeeld een postcode als selectiecriterium. Er zijn postcodegebieden waar veel mensen met een migratieachtergrond wonen. Als postcodegebieden met een hoog risico in een algoritme hiermee samenvallen, kan er sprake zijn van indirecte discriminatie op basis van migratieachtergrond.[126] Naast discriminatie zijn er ook andere grondrechtenrisico's bij de inzet van AI-systemen. Vaak zijn deze verbonden met het toepassingsgebied van het systeem. Zie ook box 4.1.

Niet-representatieve data zorgen voor oneerlijke uitkomsten. Een statistisch model is afhankelijk van de data waarmee het getraind wordt. Als een groep mensen weinig voorkomt in de data, zal het algoritme meer foute voorspellingen doen voor die groep. Dit leidt tot discriminerende uitkomsten wanneer een toepassing nadelig uitpakt.[127]

Voorbeeld: Stel dat de reisverzekeraar uit het eerdergenoemde schema geen goed systeem heeft voor claims die in een andere taal binnenkomen. Deze claims gaan eerst door het fraudealgoritme, maar de data komen zonder dat iemand dit doorheeft in een andere vorm in de database terecht. Het algoritme kan dan niet goed leren om met

FIGUUR 4.2: WANNEER EEN VERBODEN GROND (ONGEOBSERVEERD) EEN RELATIE HEEFT MET EEN INDICATOR IN HET ALGORITME, BESTAAT ER EEN VERHOOGD RISICO OP DISCRIMINATIE



claims in een andere taal om te gaan en zal hier meer fouten in maken. Dit heeft mogelijk discriminerende gevolgen.

Wanneer een beschermde groep wel voorkomt in de data maar niet op een representatieve manier, kan dit ook leiden tot discriminatie.

Voorbeeld: Stel dat het proces bij de verzekeraar enige tijd onbewust vooringenomen verliep. Hierdoor zijn mensen in een beschermde groep vaker aangeduid als fraudeur. Het algoritme zal dit patroon overnemen en deze groep in de toekomst een te hoog risico toewijzen.

Een negatieve feedbackloop verergert discriminatie. Het algoritme leert van eerdere fraudegevallen om nieuwe gevallen te vinden. Wanneer deze nieuwe gevallen na verloop van tijd gebruikt worden om weer van te leren, ontstaat er een feedbackloop. Deze feedbackloop kan discriminatie verergeren door selectiebias.

Voorbeeld: Stel dat volgens de data claims uit een bepaald land vaker frauduleus zijn. Claims uit dat land krijgen dan een hogere risico-indicatie en zullen vaker worden onderzocht. Intensiever zoeken naar fraude, is op zichzelf al een manier om daar ook meer fraude te vinden. Wanneer het algoritme opnieuw getraind wordt, zal dit land dan ook een nog sterkere focus krijgen. Dit herhaalt zich: er is een zelfversterkende feedbackloop ontstaan.

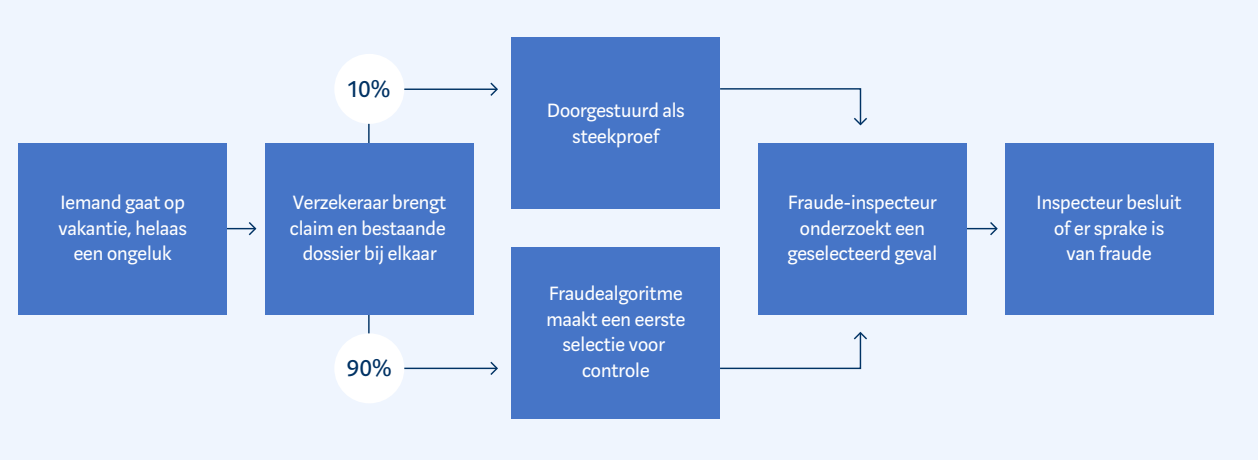
Voldoende en correcte data hebben is geen garantie voor non-discriminatie. Kenmerken voor het beoogde onderscheid zijn namelijk vaak door proxies gerelateerd aan de verboden gronden voor onderscheid.

Voorbeeld: Stel dat migratieachtergrond invloed heeft op het land waar mensen op vakantie gaan. En daarmee indirect ook de inschatting van het algoritme van de kans op fraude. Migratieachtergrond wordt niet geobserveerd (zie figuur 4.2). Ervan uitgaande dat migratieachtergrond een verboden grond is om onderscheid op te maken, zal dit algoritme mogelijk ongewenst discrimineren. Zelfs wanneer de data die gebruikt zijn voor de training van het algoritme een perfecte afspiegeling zijn van de werkelijkheid.

Overmatig vertrouwen op algoritmes is een belangrijk risico. Er is interactie tussen de menselijke beoordelaar en de uitkomst van een algoritme. In plaats van de uitkomst te controleren, hebben mensen de neiging om de uitkomst snel voor waarheid aan te nemen: "de computer zal het wel weten." Dit fenomeen wordt ook wel automation bias genoemd. Daardoor lijkt het alsof er een waarborg is met menselijke tussenkomst, terwijl deze in werkelijkheid beperkt effectief is.

Overmatig vertrouwen op algoritmes geeft ruimte aan discriminatie. Zoals besproken, kan er op verschillende manieren discriminatie ontstaan in de voorspellingen van fraudealgoritmes. De menselijke tussenkomst dient onder andere als waarborg tegen discriminatie. Als menselijke beoordelaars overmatig vertrouwen op algoritmes, kunnen discriminerende voorspellingen worden overgenomen. Het is essentieel dat de menselijke beoordelaar kritisch blijft kijken naar de uitkomst van een algoritme.

FIGUUR 4.3: EEN VERZEKERAAR KAN EEN STEEKPROEF INZETTEN DOOR EEN PERCENTAGE VAN DE SCHADECLAIMS WILLEKEURIG TE SELECTEREN, EN DEZE ONAFHANKELIJK VAN HET ALGORITME TE STUREN VOOR ONDERZOEK.



4.3 Aselecte steekproef

Een mogelijke maatregel om de genoemde risico's te verminderen is de aselecte steekproef. Wanneer een aselecte steekproef wordt toegepast, wordt een deel van de gevallen willekeurig geselecteerd om onderzocht te worden op fraude. In de afbeelding is een aselecte steekproef van 10% weergegeven. (zie figuur 4.3).

Wat het optimale percentage gevallen is dat de steekproef moet selecteren, verschilt per fraudealgoritme. Een belangrijke overweging hierbij is dat de steekproef pas kan dienen als referentie als er voldoende gevallen onderdeel van uitmaken.

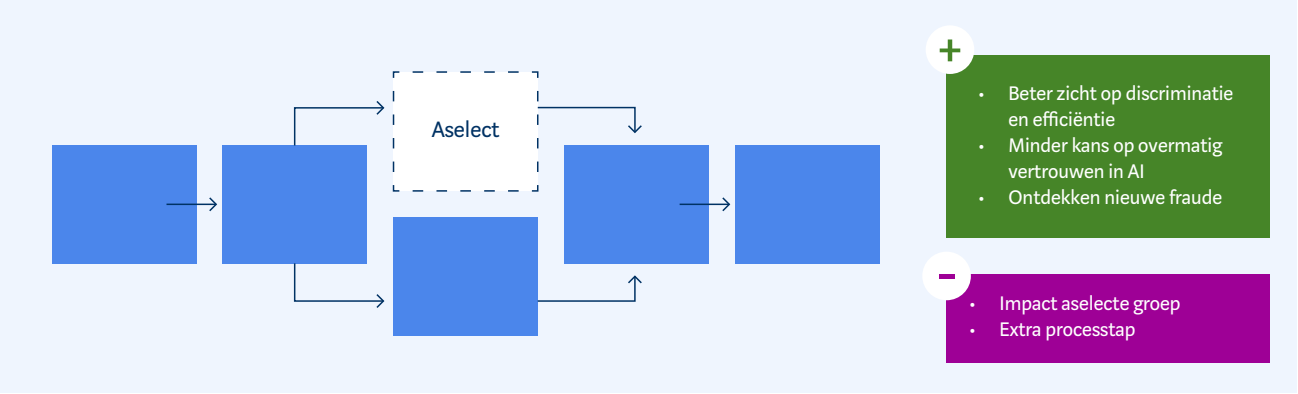
Met de aselecte steekproef als referentie kan een deel van de discriminatierisico's worden gemonitord. Door de steekproef als referentie te gebruiken, kan bijvoorbeeld worden gevolgd of een groep niet onevenredig naar voren komt door het algoritme. Het belang van willekeurige selectie voor het verkrijgen van representatieve data is eerder bijvoorbeeld beschreven door de EU Agency for Fundamental Rights (FRA).[128] Ter illustratie wederom het voorbeeld van de reisverzekering.

Voorbeeld: stel dat het fraude-algoritme van de reisverzekeraar voor 95% gevallen selecteert met nationaliteit Y. Er wordt een steekproef ingezet, en van de fraude die binnen de steekproef wordt gevonden blijkt slechts 15% van de mensen nationaliteit Y te hebben. Het lijkt er hier op dat het algoritme onevenredig veel nadruk legt op het controleren van mensen met nationaliteit Y.

Door de aselecte steekproef kan het risico op overmatig vertrouwen worden verminderd. Wanneer een steekproef een deel van de selectie voor fraudeonderzoek verzorgt, zijn niet alle onderzoeken meer op basis van een hoogrisicosignaal. Hiermee is het proces zo in te richten dat de inspecteur niet weet of een te onderzoeken geval aselect geselecteerd is. Daardoor kan de inspecteur niet meer blind uitgaan van de algoritme-uitkomsten, maar wordt de inspecteur aangemoedigd kritisch te zijn.

Op welke groep de aselecte steekproef van toepassing is, verschilt per context. In sommige toepassingen zal een fraudeindicatie geen selectie betekenen, maar juist een uitsluiting. Het blokkeren van een online bestelling is hier een voorbeeld van. Bij zo'n geval hoort een ander procesoverzicht. Een steekproef kan hier worden ingezet door aselect hoogrisicogevallen door te laten. Er zijn ook toepassingen waarbij de impact op de willekeurig geselecteerde personen significant is. Denk aan huisbezoeken als onderdeel van een fraudeonderzoek. Hierbij staat een belangenafweging centraal: een steekproef inzetten betekent niet dat je zomaar bij iemand mag binnenstappen.

FIGUUR 4.4: DE BESLISSING OM EEN ASELECTE STEEKPROEF IN TE ZETTEN HANGT AF VAN DE GEVOLGEN VOOR HET HELE PROCES



Naast risicovermindering draagt een aselecte steekproef bij aan het meten van efficiëntie en verkennen van nieuwe soorten fraude. De referentie die een steekproef verzorgt, kan worden gezien als basis om de prestaties van het algoritme mee te vergelijken. Daarnaast zal het element van willekeur ervoor zorgen dat onbekende en nieuwe vormen van fraude in de loop van de tijd ook voorkomen in de data. Een nieuwe versie van het algoritme kan hier vervolgens van leren.

4.4 Het overwegen waard

De beslissing om een aselecte steekproef in te zetten, hangt af van de gevolgen voor het hele proces. De techniek raakt aan verschillende onderdelen in het proces en kan dus niet worden afgewogen tegen een op zichzelf staand risico. Om de beslissing inzichtelijk te maken, kan een systeem zonder steekproef worden vergeleken met een systeem met steekproef (zie figuur 4.4).

De aselecte steekproef draagt in veel gevallen bij aan een verantwoordere inzet van profilerende en selecterende AI-systemen. Bij de inzet van dit soort AI-systemen is de aselecte steekproef dan ook het overwegen waard.

De inzet van een aselecte steekproef raakt aan de AI-verordening en de AVG. Een systeem rondom een fraudealgoritme moet (met of zonder steekproef) aan de wetgeving voldoen. Een profilerend of selecterend AI-systeem verwerkt persoonsgegevens en dus is de AVG van toepassing. Daarnaast treedt de komende jaren de AI-verordening in werking. Binnen de AI-verordening is het relevant of een AI-systeem als hoogrisicosysteem classificeert.

Voor bijvoorbeeld een fraudealgoritme voor essentiële overheidsuitkeringen en -diensten zal dit zo zijn. Aanbieders van hoogrisicosystemen zijn verplicht de redelijkerwijs te voorziene risico's vast te stellen en te beperken. Potentiële discriminatie is zo'n risico. De aselecte steekproef kan hierbij worden ingezet als beheersingsinstrument.

Bepaalde methoden om discriminatie tegen te gaan zijn afhankelijk van de verwerking van bijzondere persoonsgegevens. Om bijvoorbeeld te meten of een groep met een bepaalde afkomst anders behandeld wordt, zal informatie over afkomst verwerkt moeten worden. Deze gevoelige gegevens, 'bijzondere persoonsgegevens' genoemd, krijgen extra bescherming in de AVG. De verwerking van bijzondere persoonsgegevens is verboden, tenzij er een uitzondering is. De AI-verordening kan voor AI-systemen met een hoog risico onder strikte voorwaarden voorzien in een uitzondering voor het verwerken van bijzondere persoonsgegevens om discriminatie op te sporen of tegen te gaan.

Box 4.1

Grondrechtenrisico's bij de inzet van AI-systemen

Het gebruik van AI-systemen kan direct en indirect leiden tot schending van grondrechten. Dit risico speelt zowel bij zeer simpele algoritmes zoals beslisbomen als complexe systemen, bijvoorbeeld op basis van neurale netwerken.

Een bekend risico bij AI-systemen is te zien bij het recht op non-discriminatie (artikel 21 van het Handvest), maar er zijn ook risico's voor andere grondrechten. Non-discriminatie vraagstukken springen vaak in het oog omdat AI-systemen in veel gevallen immers juist worden ingezet om onderscheid te maken. Daarbij moet nadrukkelijk bekeken worden of het maken van onderscheid juridisch wel te rechtvaardigen is. Maar er is bijvoorbeeld ook het grondrecht op rechtvaardige arbeidsomstandigheden (artikel 31 van het Handvest). Dit kan onder druk komen te staan door algoritmisch management. Bijvoorbeeld wanneer dit afbreuk doet aan gezonde, veilige of waardige arbeidsomstandigheden. Een ander voorbeeld is de bescherming van persoonsgegevens (artikel 8 van het Handvest), een grondrecht dat waarborgt dat gegevens eerlijk worden verwerkt met toestemming van de betrokkene of op basis van een wettelijke grondslag. AI-systemen werken op basis van grote hoeveelheden data, waaronder vaak ook persoonsgegevens. Deze persoonsgegevens moeten rechtmatig en in lijn met dit grondrecht worden verwerkt, ook wanneer het een AI-systeem betreft. De AP ziet als onafhankelijke autoriteit toe op de naleving van deze regels.

Twee andere relevante grondrechten zijn vrijheid van informatie en het recht op behoorlijk bestuur. Beide grondrechten komen terug in deze rapportage en zijn relevant in de context van AI-systemen. Het grondrecht dat raakt aan de vrijheid en de pluriformiteit van de media is van belang bij de inzet en het gebruik van AI in de online informatievoorziening (zie hoofdstuk 2). Het recht op behoorlijk bestuur waarborgt dat zaken onpartijdig, billijk en binnen een redelijke termijn behandeld moeten worden door overheidsinstellingen en organen. De inzet van simpele algoritmes en AI-systemen in het publieke domein heeft de afgelopen jaren tot risico's en incidenten geleid die moeilijk samen gaan met dit recht.

Grondrechtenrisico's en schendingen bij de inzet van AI-systemen zijn nog te vaak buiten beeld. Ook wanneer specifieke wetgeving niet alle aspecten van nieuwe technologie adresseert, of wanneer er nog geen specifieke wetgeving is, zullen grondrechten altijd moeten worden eerbiedigt en beschermd. Om hier nader op in te gaan komt de AP later dit jaar met een factsheet over grondrechtenrisico's bij de inzet van AI-systemen.

5. Beleid en regelgeving

SNEL NAAR DIT ONDERDEEL

Wereldwijd is er steeds meer aandacht voor de regulering van AI en algoritmes. De inwerkingtreding van de Europese AI-verordening per 1 augustus 2024 is een mijlpaal. Daarbij treden sommige bepalingen al per 1 februari 2025 in werking. Bijvoorbeeld bepalingen voor verboden AI-toepassingen en voor AI-geletterdheid binnen organisaties. Punt van zorg is de lange overgangstermijn (tot augustus 2030) voordat bestaande hoog-risico AI-systemen binnen de overheid aan alle eisen moeten voldoen. Een ander aandachtspunt is dat de productstandaarden onder hoge tijdsdruk worden afgerond. Deze zijn allesbepalend voor de daadwerkelijke effectiviteit en uitvoerbaarheid van de eisen uit de AI-verordening. Ondertussen zijn toezichthouders in Nederland zich aan het voorbereiden op nieuwe toezichtstaken onder de AI-verordening.

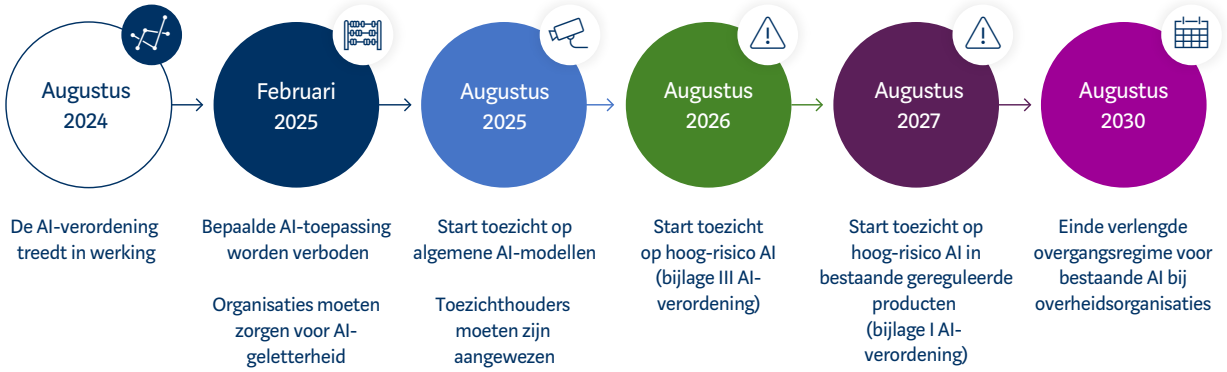
Het hoofdlijnenakkoord uit mei 2024 biedt goede aanknopingspunten om op nationaal niveau verder te werken aan adequate beheersing van de risico's van AI-systemen. Hierbij is het belangrijk dat het vullen van het algoritmeregister een prioriteit blijft en een verplichting wordt, met aandacht voor voldoende reikwijdte. Het werk aan algoritmekaders biedt organisaties een governance-instrument voor AI-systemen. Dit is belangrijk. Tegelijkertijd moet gewaakt worden voor kaders die te vrijblijvend zijn of die (onbewust) ruimte geven aan onvoldoende precieze of meetbare standaarden, die soms achterlopen op of strijdig zijn met wetenschappelijke inzichten. In de optiek van de AP is het aantreden van een nieuw kabinet een moment om de nationale AI-strategie tegen het licht te houden.

5.1 AI-verordening treedt in werking

Op 1 augustus 2024 treedt de AI-verordening in werking en start het overgangsregime voordat de AI-verordening volledig van toepassing zal zijn. Voor bedrijven, overheden en andere organisaties die met AI-systemen werken, is dit hét moment om te starten met voorbereidingen om systemen en organisaties compliant te maken (zie box 5.1). Nu de vereisten vaststaan, zullen de verschillende onderdelen van de verordening stapsgewijs – maar snel – in werking treden. Het eerste grote onderdeel waarvoor dit geldt, zijn de bepalingen waarmee bepaalde AI-toepassingen worden verboden op de Europese markt. Dit onderdeel treedt in februari 2025 in werking. Op hoofdlijnen zullen tussen februari 2025 en augustus 2027 ook de vereisten uit andere onderdelen van kracht worden (zie figuur 5.1).

Voor een werkbare naleving van deze vereisten moet nog veel geregeld en verduidelijkt worden, op nationaal én Europees niveau. Veel begrippen en procedures uit de AI-verordening vragen om nadere uitleg of invulling. Ook moet duidelijk worden hoe ontwikkelaars met Europese standaarden kunnen voldoen aan de eisen aan hoog-risico AI-systemen. Verder moet het toezicht op AI-systemen op nationaal niveau snel en met voldoende middelen worden ingericht.

FIGUUR 5.1: DE AI-VERORDENING TREEDT DE KOMENDE JAREN GETRAPT IN WERKING



Het is daarnaast positief dat de eisen aan AI-geletterdheid binnen organisaties al snel van toepassing zijn. Organisaties die AI-systemen ontwikkelen of gebruiken, moeten zorgen dat er voldoende kennis van AI is bij de medewerkers die met de AI-systemen werken. Het kennisniveau moet aansluiten bij de kennis, ervaring en opleiding van het personeel, maar ook bij de context waarin de AI-systemen worden gebruikt en hoe de systemen (groepen) personen kunnen raken. Dit is een algemene bepaling, die geldt voor alle AI-systemen en die al in februari 2025 van toepassing zal zijn. Deze bepaling is ook een belangrijke stap naar de verplichting voor gebruikers van AI-systemen om adequaat menselijk toezicht te houden.

Minder positief zijn de lange termijnen die overheidsorganisaties krijgen om aan de AI-verordening te voldoen. De meeste hoogrisicosystemen die vóór februari 2026 in gebruik zijn of worden genomen, hoeven pas aan de AI-verordening te voldoen als er na februari 2026 een aanzienlijke wijziging in het ontwerp van het systeem is. AI-systemen die daarna op de markt worden gebracht, zullen meteen moeten voldoen aan de vereisten. Voor AI-systemen die bedoeld zijn voor overheidsorganisaties geldt dat zij uiterlijk februari 2030 aan de eisen moeten voldoen. De AP maakt zich zorgen over hoe lang deze periode is voor AI-systemen die bedoeld zijn voor overheidsorganisaties. Ook geeft het overgangsregime een perverse prikkel aan bedrijven en de overheid om vóór februari 2026 aanzienlijke wijzigingen door te voeren en na die datum juist niet. De praktijk heeft juist bij bestaande AI-systemen – en onderliggende algoritmes – uitgewezen dat grondrechten en fundamentele waarden in het geding kunnen zijn. Tegelijkertijd benadrukt de AP dat een hoger ambitieniveau mogelijk is en dat dit in nationale wetgeving verankerd kan worden. Een uitvoerbaar transitieplan voor bestaande hoog-risico AI-systemen bestaat eruit (i) deze eerst in kaart te brengen, (ii) overheidsorganisaties een plan te laten maken voor hoe zij zelfontwikkelde hoogrisicosystemen compliant kunnen maken en (iii) voor te schrijven wanneer – eerder dan februari 2030 – deze systemen aan de AI-verordening moeten voldoen. Aanvullend zouden overheden ook voor die tijd alleen systemen moeten inkopen die al zo veel mogelijk voldoen aan de regels in de AI-verordening.

5.2 Verboden AI

Sommige AI-toepassingen worden vanaf februari 2025 verboden, maar de precieze reikwijdte van deze verbodsbepalingen vraagt om verdere verduidelijking en uitleg. De verboden zijn een belangrijk onderdeel van de AI-verordening, omdat er straks systemen zijn waarvan het in de handel brengen of in gebruik nemen volledig aan banden moet worden gelegd. Dit gaat gelden voor systemen die een onaanvaardbaar risico met zich meebrengen, bijvoorbeeld omdat die systemen de vrije keuze van mensen te veel beperken, mensen uitbuiten of mensen manipuleren. Alleen wanneer er snel een praktische en concretere interpretatie komt van de in de verbodsbepalingen opgenomen AI-systemen, wordt duidelijk wanneer Europese toezichthouders wel of niet moeten ingrijpen. Zo wordt zekerheid geboden naar de markt en de samenleving. De AP verwelkomt daarom de opheldering die de Europese Commissie gaat geven met richtsnoeren. Een eerste richtsnoer gaat nadere invulling geven aan de definitie van een AI-systeem. Een tweede richtsnoer gaat meer duidelijkheid geven over de invulling van de verboden AI-toepassingen.

De AP doet nog dit jaar een call for input om de concrete invulling van de verboden AI-toepassingen te verkennen. Belanghebbenden krijgen de mogelijkheid inbreng te geven op twee in de AI-verordening opgenomen verboden producttoepassingen. Het doel hiervan is kennis en praktische vragen bij organisaties op te halen, om uiteindelijk adequaat toezicht en rechtszekerheid mogelijk te maken. Dit biedt ook een basis voor verdere beleidsvorming met andere (Europese) toezichthouders, bijvoorbeeld in de context van de te ontwikkelen richtsnoeren.

5.3 AI-standaarden

Europese productstandaarden zijn essentieel voor het naleven van de AI-verordening. Dergelijke standaarden bieden AI-ontwikkelaars houvast bij de vereisten uit de verordening. In het standaardisatieproces is de tijdsdruk echter hoog. Ook worden de resultaten in de vorm van productstandaarden vooralsnog niet vrij toegankelijk. Normalisatieorganisaties CEN en CENELEC ontwikkelen standaarden om nadere invulling te geven aan de eisen uit de AI-verordening. Wanneer organisaties werken volgens deze standaarden, wordt verondersteld dat hun hoogrisicosystemen voldoen aan de in de AI-verordening gestelde eisen. In de praktijk zullen de normen dus een grote rol spelen in het aantonen van compliance en de beoordeling van conformiteit.

De AP maakt zich echter zorgen over de snelheid waarmee de standaarden moeten worden opgeleverd. De normalisatieorganisaties hebben vanaf het standaardisatieverzoek van de Europese Commissie[129] slechts drie jaar de tijd voor de ontwikkeling. Doorgaans neemt het opstellen van technische productstandaarden echter veel tijd in beslag. Het opleveren van standaarden voor de AI-verordening is bovendien nog complexer, gezien de brede focus op zowel gezondheid en veiligheid als op grondrechten. Aanbieders en gebruikers van AI-systemen moet er hierdoor rekening mee houden dat de standaarden mogelijk gelijktijdig met – of pas na – de inwerkingtreding van de bepalingen over hoogrisicotoepassingen beschikbaar zijn. Beleidsmakers moeten werk maken van (het voorkomen van) een scenario waarin organisaties moeten voldoen aan de producteisen uit de verordening voordat zij de standaarden kunnen gebruiken.

Het blijft een aandachtspunt dat deze productstandaarden in beginsel enkel na betaling toegankelijk gaan worden. Vooral omdat het gaat om standaarden die moeten bijdragen aan de bescherming van grondrechten en fundamentele waarden. Doordat de normen niet algemeen toegankelijk zijn, kunnen deze minder snel doorwerken in de algemene AI-geletterdheid die organisatie- en maatschappijbreed noodzakelijk is. Het werpt ook een extra drempel op voor het algemeen publiek om controle uit te oefenen op een belangrijke uitwerking van de AI-verordening. Tegelijkertijd moet erkend worden dat het staande praktijk is dat productstandaarden – en het onderliggende standaardisatieproces – op deze manier gefinancierd worden. Als beleidsmakers ervoor kiezen om de productstandaarden publiekelijk beschikbaar te stellen, moet hier dus een passende oplossing voor gevonden worden.

Box 5.1

Starten met de voorbereiding op de AI-verordening

Bedrijven, overheden en andere organisaties die AI gebruiken of ontwikkelen doen er goed aan zich voor te bereiden op de nieuwe regels. Aanbieders en gebruikers van AI-systemen raden we aan direct te starten met een implementatieplan en het opzetten van het interne risicomanagement. Een eerste stap daarbij is het in kaart brengen van de systemen die zij ontwikkelen of gebruiken en of die vallen binnen de definitie van 'AI-systeem' in de verordening. Vervolgens moeten zij meestal een inschatting maken of deze systemen in een van de volgende risicogroepen vallen:

1. **Verboden AI.** Deze systemen moeten uit de handel genomen worden en het gebruik ervan moet worden gestopt. De bepalingen over verboden AI zijn vanaf februari 2025 al van kracht. Ook is de kans groot dat met deze systemen nu al de wet wordt overtreden, zoals wetgeving op het gebied van gelijke behandeling, privacy of arbeidswetgeving.
2. **Hoog-risico-AI.** Deze systemen moeten voldoen aan de eisen aan onder andere risicobeheer, de kwaliteit van de gebruikte data, technische documentatie en registratie, transparantie en menselijk toezicht. Voor overheden of uitvoerders van publieke taken gelden soms aanvullende vereisten, zoals het uitvoeren van een 'grondrechteneffectbeoordeling'.
3. **AI met een beperkt risico.** Voor systemen die bedoeld zijn om met personen in contact te komen of die inhoud genereren, zoals deepfakes, gelden transparantieverplichtingen. Als deze systemen worden aangeboden of gebruikt, moeten mensen daarover worden geïnformeerd.

Gebruikers van AI-systemen kunnen alvast inschatten of hun AI-systemen voldoen of zullen voldoen aan de AI-verordening. Zij kunnen bij aanbieders informeren in hoeverre de gebruikte AI al voldoet aan de eisen. Bij de aanschaf van een AI-systeem is het zaak om de inkoopvoorwaarden te controleren en bijvoorbeeld ook goed te letten op wat er gebeurt met data die het AI-systeem verwerkt en hoe de rechten op die data zijn geregeld. Er wordt ook gewerkt aan model-contractbepalingen voor de aanbesteding van AI in de publieke sector vanuit de Europese *public buyers community*.[130]

Rollen en verantwoordelijkheden in de AI-verordening kunnen overlappen en verschuiven. Zowel aanbieders als gebruikers moeten aan bepaalde verplichtingen voldoen. Organisaties die zelf AI-systemen ontwikkelen en gebruiken, vervullen in wezen beide rollen en moeten zich dan ook aan alle verplichtingen houden. De rollen kunnen ook verschuiven. Wordt een ingekocht systeem aangepast of voor een ander doel gebruikt? Dan kan een organisatie een aanbieder worden van een systeem, die zich dan bijvoorbeeld aan de regels voor hoogrisicosystemen moet houden. Gebruikers van AI-systemen raden we daarom aan goed in de gaten te blijven houden of (en zo ja, hoe) zij zelf bijdragen aan de ontwikkeling van de AI-systemen die zij gebruiken.

Ontwikkelaars van AI-systemen kunnen terecht in de Nederlandse testomgeving voor de regelgeving, ook wel regulatory sandbox genoemd. De AP, de Rijksinspectie Digitale Infrastructuur (RDI) en het ministerie van EZK werken aan de voorbereiding van de sandbox. Aanbieders kunnen hier in de loop van 2026 terecht om compliancevragen over de AI-verordening op te lossen. Tot die tijd testen de toezichthouders in pilots de werking van de sandbox en sandboxtrajecten.

5.4 Toezicht op de AI-verordening in Nederland

De in Nederland lopende voorbereidingen op het toezicht op de AI-verordening verdienen blijvende aandacht. In mei heeft de AP samen met de RDI een advies[131] uitgebracht aan de ministeries van EZK en BZK over hoe het toezicht op de AI-verordening goed geregeld kan worden.[132] Dit advies is het resultaat van een samenwerking van ruim tien verschillende toezichthouders, colleges en inspecties. De AP heeft een coördinerende rol gespeeld in de totstandkoming van dit advies. De toezichthouders hebben in het advies uitgewerkt hoe het toezicht op de AI-verordening zou moeten worden ingericht en wie er als zogenoemde markttoezichtautoriteiten moeten worden aangewezen.

De AI-verordening biedt de kans om het toezicht op AI te versterken. Dit vraagt echter voldoende capaciteit en de juiste randvoorwaarden. Allereerst moet snel duidelijk worden welke instanties de verschillende delen van het toezicht gaan uitvoeren. Ten tweede moet er op tijd voldoende budget en personeel beschikbaar zijn voor alle betrokken toezichthouders. Zodat zij op tijd kunnen beginnen aan hun taken, zoals voorlichting en handhaving. Ten derde is AI een systeemtechnologie die niet alleen door de AI-verordening wordt gereguleerd. Het is daarom belangrijk dat het toezicht op de AI-verordening en het bestaande toezicht op bestaande wetgeving elkaar versterken en aanvullen. Bijvoorbeeld in de toezichtsrelatie op het gebied van consumentenbescherming, data, onderwijs en arbeid. Het inrichten van toezicht op AI is een opgave waarbij de RDI namens Nederland en Europa ook samenwerkt met UNESCO (zie box 5.2).

5.5 Internationale ontwikkelingen in AI-regelgeving en -beleid

De noodzaak om AI te reguleren wordt ook buiten de EU sterk gevoeld. Wereldwijd wordt er daarom iets gedaan, maar wel met verschillende invalshoeken. Verschillende landen en op internationaal niveau wordt volop gewerkt aan wetgeving, normen en uitgangspunten voor de regulering van AI. Daarbij bestaat het risico dat gebrek aan coördinatie fragmentatie met zich meebrengt. Wereldwijd wordt er verschillend gereageerd op de ontwikkelingen rondom AI. Dit is afhankelijk van de specifieke uitdagingen waar landen mee kampen en van onderliggende waarden die prioriteit krijgen. Zo ligt de focus in het mondiale zuiden vooral op de mogelijkheden die AI creëert voor de nationale economie. Terwijl in bijvoorbeeld Europa en de VS de risico's van verder ontwikkelde AI centraal staan.[133] Dit sluit ook aan bij de verschillen in risicoperceptie tussen burgers uit deze regio's, zoals besproken in hoofdstuk 1. In de EU hebben de rechten van het individu een expliciete positie in het beleid voor en de regulering van het internet. De VS deelt op dit moment de Europese ambitie om te waarborgen dat AI betrouwbaar en veilig is en bijdraagt aan de bescherming van mensenrechten. De visies verschillen in de afweging tussen het stimuleren van innovatie enerzijds en het waarborgen van maatschappelijke waarden anderzijds.[134] Een andere invalshoek voor AI-regulering is zichtbaar in China, waar het landsbelang nadrukkelijk voorop staat.[135]

Internationale ontwikkelingen bieden steeds meer een basis voor samenwerking bij de ontwikkeling van AI-systemen. In China staat voor het eerst een meer alomvattende wet op AI op de wetgevingsagenda. De regulering van AI in China is tot zover meer fragmentarisch geweest. Hierbij ziet regulering niet op AI als geheel, is het doel ervan eerder om grip te krijgen op specifieke systemen of producten, zoals aanbevelingssystemen.[136] Daarnaast ligt bijvoorbeeld in Brazilië een nieuwe wet[137] voor die met name de ontwikkeling van AI probeert te reguleren op een manier die vergelijkbaar is met de benadering die de EU heeft gekozen in de AI-verordening.[138] Het is zichtbaar dat begrippen als betrouwbaarheid, veiligheid en mensenrechten meer aandacht krijgen in nationale reguleringsinitiatieven wereldwijd. Tegelijkertijd moet zorgvuldige aandacht blijven uitgaan naar de verschillen in uitdagingen, perspectieven en regelgevingsinitiatieven. Fragmentatie in nationale benaderingen kan namelijk leiden tot spanningen wanneer AI grensoverschrijdend wordt gebruikt.

Inclusieve internationale normen en standaarden kunnen bijdragen aan de benodigde harmonisatie. Internationale normen en standaarden zijn van belang voor een veilige ontwikkeling en inzet van AI op globaal niveau, zowel door private partijen als door nationale overheden. Inclusieve internationale normen kunnen de kloof in beleid en regulering verminderen en bijdragen aan geharmoniseerde implementaties. Internationale organisaties als de OECD, UNESCO en de G20 spelen een belangrijke rol in het harmoniseren van normen en regulering.[139]

De OECD publiceerde begin 2024 een update van de AI-principes van de OECD. Het gebruik van deze principes door de OECD-lidstaten (momenteel 38 landen) legt een basis voor een wereldwijde interoperabiliteit tussen landen. De principes dienen als leidraad voor een betrouwbare ontwikkeling van AI en bieden aanbevelingen voor beleid en strategie van de overheid voor AI. In 2019 hebben de OECD-lidstaten deze principes ondertekend. De OECD heeft

de principes dit jaar bijgewerkt vanwege de nieuwe ontwikkelingen rondom AI-systemen, met name de opkomst van foundation models en generatieve AI. De bijgewerkte versie adresseert uitdagingen voor privacy, intellectuele eigendomsrechten, veiligheid en informatie-integriteit in de context van mis- en desinformatie. Ook wordt het belang onderstreept van verantwoord ondernemen en een goede samenwerking in beleid en governance.[140]

Binnen de Verenigde Naties is in maart 2024 een eerste mondiale resolutie[141] over AI aangenomen. Deze niet-bindende VN-resolutie is ingediend door 122 landen, waaronder China. De resolutie roept staten op om mensenrechten te waarborgen en met regelgeving en governance te zorgen voor een betrouwbare ontwikkeling en inzet van AI. Ook roept de resolutie op de digitale kloof te dichten, zodat landen waarin de ontwikkeling van AI minder ver gevorderd is ook kunnen profiteren van de mogelijkheden die AI teweegbrengt.[142]

Een mondiaal instituut voor AI-governance kan een belangrijke stap zijn naar sterke gemeenschappelijke standaarden. Hoewel formeel niet bindend, weerspiegelt de VN-resolutie een internationale consensus over de standaarden die gevolgd moeten worden bij de ontwikkeling en het gebruik van AI. Uit eerdere, vergelijkbare initiatieven is gebleken dat dergelijke afspraken sturing geven aan het opstellen van nationaal beleid en nationale regelgeving.[143] Desondanks moet niet zonder meer aangenomen worden dat dergelijke standaarden de versnippering zullen tegengaan in nationale strategieën en reguleringsinitiatieven. Naast nationale governance moet daarom gewerkt worden aan een internationaal governance-raamwerk.

Om dit te verkennen is binnen de VN een speciaal adviesorgaan opgericht. De High-Level Advisory Body on Artificial Intelligence heeft onlangs het tussenrapport 'Governing AI for Humanity' opgeleverd.[144] Daarin wordt opgeroepen om internationale governance te versterken. Het adviesorgaan benadrukt een inclusieve globale aanpak om tot harmonisatie te komen. Dit wordt mede bereikt door het opstellen van een globaal AI-governance-raamwerk, waarin de volgende functies aan bod komen: (i) signaleren en monitoren van AI-ontwikkelingen, (ii) werken aan consensus over internationale standaarden en (iii) monitoren van systemische kwetsbaarheden voor de mondiale stabiliteit. Een globaal instituut, waarbij verschillende stakeholders betrokken zijn, kan door deze activiteiten zicht houden op globale ontwikkelingen. En kan deze inzichten vervolgens gebruiken om inclusieve internationale normen en standaarden op te stellen. Het toezicht op AI mag daarbij geen ondergeschoven kindje worden. In reactie op dit tussenrapport schreef en publiceerde de AP als coördinerend algoritmetoezichthouder daarom een *discussion paper*[145] waarin ook het belang wordt aangekaart van nationaal en internationaal toezicht bij governance van AI.[146] Toezichthoudende autoriteiten zijn immers in staat om ontwikkelingen en risico's in een vroeg stadium te monitoren. Zij kunnen gezamenlijk bijdragen aan het opstellen van richtlijnen en standaarden voor een verantwoorde inzet van AI.

5.6 Nationale ontwikkelingen in AI-regelgeving en -beleid

Het Hoofdlijnenakkoord uit mei 2024 bevat belangrijke bepalingen over algoritmes en AI, die het huidige beleid kunnen versterken. In het akkoord is afgesproken dat er een wetenschappelijke standaard komt voor het gebruik van modellen en algoritmes. Aan deze standaard worden de eisen gesteld dat deze openbaar en navolgbaar zijn, met een duidelijke instructie waarvoor deze modellen en algoritmes wel en niet bedoeld zijn en waarvoor ze gebruikt mogen worden.

De AP verwelkomt deze eisen aan het gebruik van modellen en algoritmes. De eisen moeten echter in samenhang met de bepalingen uit de AI-verordening bekeken worden. De AI-verordening bevat namelijk vergelijkbare eisen. De verordening schrijft voor dat hoogrisicosystemen moeten voldoen aan kwaliteitseisen, rekening houdende met het doel van het gebruik en de algemeen erkende stand van de techniek.

Een van die eisen uit de AI-verordening is dat AI-systemen nauwkeurig moeten zijn bij het doel van het gebruik ervan. Dit voorkomt willekeur door algoritmes en AI-systemen, wat één van de doelstellingen is waarvoor de AP zich inzet vanuit de coördinerende algoritmetaak. De AI-verordening zal ertoe leiden dat benchmarks en meetmethoden worden ontwikkeld voor het beoordelen van AI-systemen op nauwkeurigheid en robuustheid. Bijvoorbeeld in samenwerking met metrologie- en benchmarkingsautoriteiten (zie artikel 15 van de AI-verordening).

Ook stelt de AI-verordening dat hoogrisico AI-systemen zo moeten worden ontworpen dat de werking ervan transparant genoeg is. Dit stelt organisaties die zo'n systeem gebruiken in staat de output van het systeem te interpreteren en op de juiste manier te gebruiken. De gebruiksinstructie voor een hoog-risico AI-systeem moet ook ingaan op het doel en de mate van nauwkeurigheid (zie artikel 13 van de AI-verordening).

Het hoofdlijnenakkoord ziet de voordelen van het gebruik van AI door de overheid, maar acteert ook op de risico's. Er is de ambitie om kennis van digitalisering binnen de overheid te versterken. Tegelijkertijd is in het hoofdlijnenakkoord afgesproken dat aan het gebruik van AI door de overheid speciale voorwaarden verbonden zijn om de veiligheid, privacy en rechtsbescherming te waarborgen. De AP ziet dit als goede ambities. Het belangrijkste instrument om het gebruik van AI door de overheid in goede banen te leiden, is een zo proactief mogelijke voorbereiding op, en implementatie door (overheids)organisaties van, de vereisten uit de AI-verordening. Daarbij geldt het belang van (i) de brede definitie van AI-systeem (zie Kernboodschappen in deze RAN) en (ii) een ambitieniveau voor de overheid dat hoger ligt dan de ondergrens in de verlengde overgangsperiode voor bestaande AI-systemen (zie het onderdeel 'AI-verordening treedt in werking' in dit hoofdstuk). Daarnaast bieden de kennisvereisten voor AI-geletterdheid (per 1 februari 2025) een aangrijpingspunt om het niveau van kennis over AI (en daarmee digitalisering in brede zin) te verhogen.

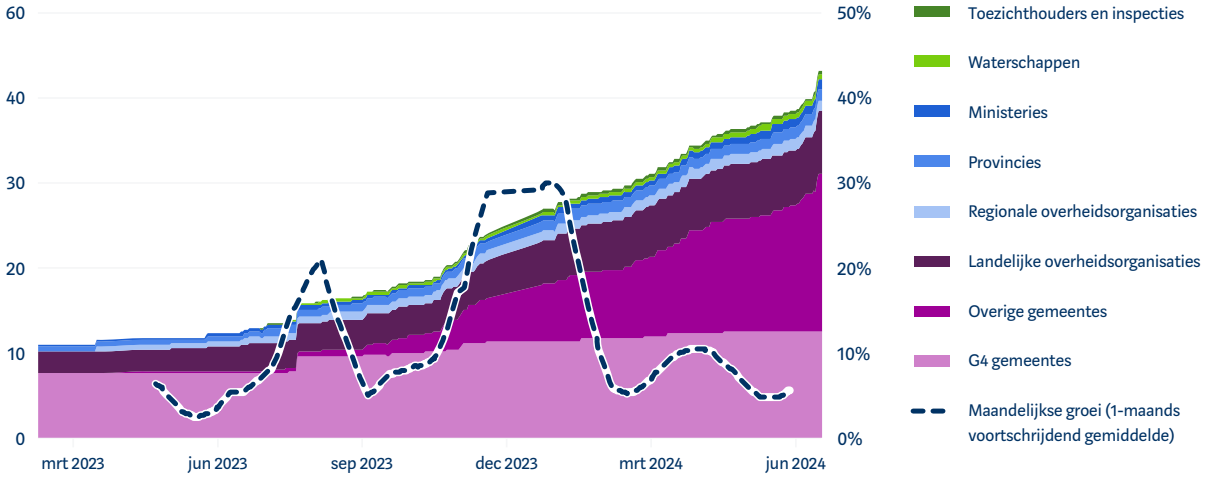
Op basis van het hoofdlijnenakkoord moet het nieuwe kabinet de maatschappij ook weerbaarder maken tegen desinformatie en deepfakes. Dit sluit aan bij de observaties die de AP vanuit het coördinerend algoritmetoezicht doet over de impact van AI op informatievoorziening in de democratie (zie hoofdstuk 2 van deze RAN) en de opkomst van deepfakes, die door generatieve AI makkelijker en op grote schaal te produceren zijn (zie hoofdstuk 1 en 2 van deze RAN). Ook de afspraken om de aanpak van digitale dreigingen te versterken, zijn in lijn met het risicobeeld dat volgt uit deze RAN. Hierbij is de observatie dat kwaadwillenden (generatieve) AI op ontwrichtende wijze kunnen inzetten (zie hoofdstuk 1 van deze RAN).

Op korte termijn is het belangrijk dat algoritmeregistratie een prioriteit blijft. Registratie is essentieel om transparant te zijn naar de burger en inzicht te hebben in eigen algoritmegebruik. Registratie is ook een goede basis voor risicobeheersing. Het afgelopen half jaar is het nationale Algoritmeregister voor publieke organisaties verder gevuld, tot meer dan 400 algoritmes (eind juni 2024). Daarbij valt op dat er eind 2023 een piek was in de groei van het aantal registraties. Het nationale Algoritmeregister is een goed middel om aan de oproep tot registratie te voldoen. Genoeg tempo houden in de groei van dit register is van belang om tot volledige registratie te komen. De groei in de afgelopen maanden is voornamelijk afkomstige van registraties door gemeentelijke organisaties. Het valt op dat toezichthouders en inspecties, waterschappen, ministeries, provincies en regionale overheidsorganisaties vooralsnog maar weinig algoritmes in het register hebben geplaatst (zie figuur 5.2).

De regering heeft met de Tweede Kamer afgesproken uiterlijk eind 2025 alle hoog-risico AI-systemen binnen de Rijksoverheid te registreren. De AP ziet graag een bredere reikwijdte en duidelijkheid over de consequenties als organisaties niet aan deze deadline voldoen. De AP blijft er voorstander van dat registratie door overheidsorganisaties snel verplicht wordt. Het Algoritmeregister zou hierbij op het niveau van registrerende overheidsorganisaties ook inzicht kunnen bieden in de organisatorische beheersing van AI en algoritmes en transitiepaden ter ondersteuning van algoritmeregistratie en AI-beheersing. De AP benadrukt daarnaast dat de scope van het Algoritmeregister breed genoeg moet zijn. De afweging is niet altijd simpel of een bepaald AI-systeem (of algoritme) al dan niet een toepassing met een hoog risico is. Daarom is het belangrijk dat ook politieke organen en derden deze afweging kunnen beoordelen. Algoritmeregistratie biedt hiervoor de basis. Registratieverplichtingen moeten daarom niet beperkt worden tot algoritmes waarvan het (op voorhand) al zeker is dat deze een hoog risico kennen.

Een aanvullend aandachtspunt hierbij is de mate waarin algoritmes en AI-systemen in de niet-commerciële dienstverleningssector in beeld zijn en geregistreerd worden. Organisaties in de gezondheidszorg, het onderwijs, de volkshuisvesting en het openbaar vervoer dragen bij aan essentiële dienstverlening. De reikwijdte van het huidige Algoritmeregister strekt zich echter niet uit tot deze organisaties die op afstand van de overheid staan. Tegelijkertijd zijn AI-systemen steeds vaker een onderdeel van de wijze waarop deze organisaties hun diensten verzorgen. Door gebrek aan registers is het vanuit het coördinerend algoritmetoezicht op dit moment moeilijk om zicht te krijgen op hoe deze organisaties algoritmes en AI precies inzetten en hoe het staat met de risicobeheersing.

FIGUUR 5.2: HET AANTAL GEREGISTEERDE SYSTEMEN IN HET ALGORITMEREGISTER VAN DE OVERHEID GROEIT DE AFGELOPEN PERIODE MET NAME DANKZIJ GEMEENTES, DE GROEISNELHEID NEEMT AF NA EEN PIEK ROND JAAREINDE 2023



Toelichting: *Ontwikkeling geregistreerde algoritmes op algoritmes.overheid.nl van 13 maart 2023 tot en met 30 juni 2024.*

Eerder dit jaar publiceerde de regering een overheidsbrede visie op generatieve AI. Positief aan de visie is dat deze risico's belicht en erkent die aanwezig zijn bij de inzet van generatieve AI. Risico's voor privacy en auteursrechten worden specifiek benoemd, net als de mogelijke marktmacht van bigtechbedrijven. Het uitbrengen van deze visie versterkt de maatschappelijke dialoog en kan een impuls geven aan toekomstig beleid. Aandachtspunten die de AP ziet, zijn dat de technologie volop in ontwikkeling is en dat de duiding hiervan beperkt is. Daarnaast zijn er indirecte gevolgen die maar beperkt worden meegenomen, zoals algoritmevervorming. Ook is meer aandacht nodig voor maatschappijbrede educatie. De wijze waarop systemen voor generatieve AI voldoen aan vereisten voor gegevensbescherming blijft voor de AP een groot zorgpunt, zoals eerder belicht.[147]

De regering werkt ook aan een nieuw algoritmekader. De opzet van het algoritmekader lijkt ondersteunend. Het kader heeft volgens de regering als doel een praktisch overzicht te bieden van "de belangrijkste bestaande normen [...] en maatregelen die daarbij kunnen helpen." Daarnaast bevat het algoritmekader "Richtsnoeren en maatregelen die niet verplicht zijn, maar als handreiking dienen voor het borgen van fundamentele waarden. De eisen uit de AI-verordening zullen hierin nog worden opgenomen". Aandachtspunten bij een dergelijk ondersteunend kader zijn eventuele vrijblijvendheid en hoe organisaties invulling geven aan open normen (zie ook het onderdeel 'Kaders, normen en instrumenten voor de inzet van AI' in dit hoofdstuk). Daarom is het belangrijk dat er ook eisen komen aan de (afdwingbare) opbouw van AI-kennis, AI-governance en AI-strategieën binnen overheidsorganisaties.

Dit vraagt ook nadrukkelijk om investeringen in personeel, IT en educatie, met bijbehorende financiële middelen en bestuurlijke verantwoordelijkheid binnen organisaties. Zo wordt het mogelijk om stapsgewijs de beheersing van de risico's van algoritmes en AI te verbeteren.

In de optiek van de AP is het aantreden van een nieuw kabinet ook een moment om de nationale AI-strategie tegen het licht te houden. Dit is eerder aan de orde gekomen in hoofdstuk 1 van deze RAN. De huidige nationale AI-strategie is het Strategic Action Plan for AI (SAPAI) uit oktober 2019. Door de stormachtige ontwikkeling van AI-technologie in de afgelopen vier jaar, is het passend om de strategie te evalueren en opnieuw vast te stellen. Dit is nodig om aan te sluiten bij nieuwe uitdagingen en de verdere maatschappelijke transitie die de komende jaren moet worden doorlopen.

Hierbij kan ook een nationale AI-adviesraad een rol spelen, zoals beleidsmakers op dit moment verkennen. De AP ziet daarbij ruimte voor een multistakeholderaanpak, waarbij de adviesraad kennis samenbrengt uit wetenschap, toezicht, beleid, praktijk (sectoren) en burgerperspectief.

5.7 Kaders, normen en instrumenten voor de inzet van AI

Het risico bestaat dat er een wildgroei komt aan deelkaders (inclusief toetsingsnormen en implementatie-instrumenten) bij gebrek aan voldoende precieze, volledige en meetbare standaarden voor AI-systemen. Dit brengt twee risico's met zich mee. Ten eerste kan het brede palet aan kaders – gegeven het partiële karakter – schijnzekerheid bieden. Bijvoorbeeld als (i) een implementatie-instrument veel interpretatievrijheid geeft, (ii) criteria moeilijk (objectief) meetbaar of te wegen zijn en (iii) er geen deskundigheidsvereisten zijn voor de gebruikers van het kader. Ten tweede kunnen de kaders juist selectief gebruikt worden om de inzet en prestaties van een AI-systeem te legitimeren. Bijvoorbeeld wanneer een bepaald kader zich richt op een bepaald type criterium en/of compliance vanuit een bepaalde invalshoek (dit doet dan overigens niets af aan nut en noodzaak van dit kader voor het beoogde doel). Het is dan niet de bedoeling dat de (positieve) uitkomst breder wordt getrokken naar andere domeinen of invalshoeken vanwaaruit een systeem óók beoordeeld moet worden.

In het ergste geval kunnen organisaties de situatie met kaders rooskleuriger laten lijken dan deze is, wat kan leiden tot een vorm van *AI ethics washing*. Bijvoorbeeld wanneer een organisatie zich op papier committeert aan een ethische (risico)beheersing van AI en hiervoor (procedurele) bewijzen aandraagt, maar dit in de praktijk onvoldoende opvolgt met daadwerkelijke maatregelen om de risico's langdurig te beheersen.

Een recent rapport van een internationale denktank wijst er daarbij op dat in veel kaders en instrumenten serieuze tekortkomingen zitten. Het World Privacy Forum heeft in december 2023 een rapport gepubliceerd over 'AI-ondersteunende governance-instrumenten', de overkoepelende term voor handreikingen, toetsingskaders, raamwerken en soortgelijke instrumenten. Uit een onderzoek naar bijna 20 van deze instrumenten blijkt dat bijna 40 procent verwijst naar meetmethoden die volgens wetenschappelijke literatuur ongeschikt, ongepast of niet-relevant zijn bij het meten van AI-systemen. Een voorbeeld is het voorschrijven van de 80%-regel voor het beoordelen van de bias (vooringenomenheid) van een AI-systeem, terwijl dit een maatstaf is die voor veel toepassingen ongeschikt is. Verder valt op dat er grote verschillen zijn in de vorm van dit soort instrumenten. Soms is een 'kader' beperkt tot praktische guidance, eventueel aangevuld met een vragenlijst voor een selfassessment. In andere gevallen omvat het kader ook een technisch raamwerk, inclusief software, scores en schalen om de uitkomst te beoordelen, inclusief drempelwaarden om vast te stellen of – conform dat kader – een AI-systeem voldoet

Box 5.2

UNESCO versterkt AI-toezicht in Nederland en de Europese Unie

Door: Rijksinspectie voor Digitale Infrastructuur

Nederlandse en Europese toezichthouders staan voor een gezamenlijk opgave: effectief toezicht houden op AI. Die opgave brengt de nodige uitdagingen met zich mee. AI overstijgt zowel fysieke als digitale grenzen en vraagt daarom om een gecoördineerde aanpak. Verder moeten toezichthouders rekening houden met zowel bestaande als nieuwe wetgeving, zoals de AI-verordening. Daarbij is er nog weinig duidelijkheid in de vorm van bijvoorbeeld guidance of best practices. En misschien nog wel de belangrijkste uitdaging: niet alle toezichthouders hebben momenteel genoeg ervaring met en kennis over toezicht op AI.

De Rijksinspectie Digitale Infrastructuur (RDI) werkt met UNESCO aan de ontwikkeling van de capaciteiten van de Europese toezichthouders om toezicht te houden op AI. In het licht van de genoemde uitdagingen heeft de RDI, als voorzitter van de Europese en Nederlandse werkgroepen van toezichthouders op AI, namens andere Europese toezichthouders de Europese Commissie om ondersteuning gevraagd. De Europese Commissie schakelde daarop UNESCO in om de toezichthouders bij te staan in de uitdagingen van effectief toezicht op AI. De samenwerking is een aanvulling op bestaande activiteiten, zoals die van de Nederlandse en Europese werkgroep van toezichthouders op AI.

De opdracht aan UNESCO luidt: "Ondersteuning bieden aan de RDI en leden van de Nederlandse en Europese werkgroepen van toezichthouders op AI, om hun toezichtcapaciteiten te versterken in overeenstemming met de AI-verordening en andere relevante wetgeving."

De samenwerking tussen de RDI en UNESCO heeft een aantal concrete doelstellingen: Allereerst een nulmeting op basis van een uitgebreid rapport over de huidige praktijken van AI-toezicht in Europa en daarbuiten. AI-systemen beperken zich niet tot domeinen en landsgrenzen binnen of buiten de EU. Een brede scope op ontwikkelingen is noodzakelijk. Ten tweede de ontwikkeling en discussie van casestudies naar AI-toezicht met de leden van de Europese en Nederlandse werkgroepen waarvan de RDI voorzitter is. Ten derde het opstellen en verspreiden van een reeks best practices voor het omgaan met specifieke AI-toezichtskwesties. Ten vierde het verkennen van benaderingen en opties, en deze presenteren aan relevante stakeholders binnen AI-toezicht. Een vijfde doelstelling is het trainen van toezichthouders, onder andere op basis van de ontwikkelde best practices.

De samenwerking levert naast verbeterde capaciteit een uniformer toezicht op. UNESCO en de RDI richten zich primair op inspecteurs die moeten toezien op AI-systemen. Doordat verschillende toezichthouders echter dezelfde lessen leren, krijgen ze ook dezelfde toezichtsaanpak aangeleerd. Dat resulteert in een uniformer toezicht door de verschillende nationale en Europese autoriteiten heen.

De samenwerking leidt op korte termijn tot tastbare resultaten. Een eerste rapport wordt medio 2024 opgeleverd. De overige doelstellingen volgen gefaseerd. Het project wordt eind 2025 afgerond.

Deze box is geschreven door de Rijksinspectie voor Digitale Infrastructuur, toezichthouder op de beschikbaarheid, continuïteit en betrouwbaarheid van de digitale infrastructuur in Nederland.

Bijlage: wat maakt het beheersen van AI-risico's zo complex?

Van impact assessments, ethische normen en implementatiekaders tot evaluatiekaders, fairness metrics en transparantie-verplichtingen: allemaal waarborgen die bijdragen aan de bescherming van grondrechten en fundamentele waarden bij de inzet van AI-systemen. Maar hoe hangen deze concepten samen binnen de gehele levenscyclus van AI-systemen?

Er is geen silver bullet voor het beheersen van de risico's van AI-systemen. Een verantwoorde inzet van AI binnen organisaties, of aanbieding van AI-systemen aan particuliere eindgebruikers, vraagt om wisselwerking en samenhang tussen risicobeheersingsmaatregelen in de ontwikkel-, inzet- en evaluatiefase van een AI-systeem. Maar verantwoorde inzet of aanbieding raakt evenzeer aan de overkoepelende cultuur, ethiek, kennis en governance die daarvoor op het niveau van de organisaties (die AI-systemen ontwikkelen en/of inzetten) nodig zijn. Dit is precies de reden waarom het voor organisaties ingewikkeld is om overzicht te krijgen bij de vele kaders die zij aangereikt krijgen en bijbehorende invalshoeken en accenten (zie ook hoofdstuk 5).

De schets in deze bijlage geeft op hoofdlijnen een overzicht van de samenhang in de risicobeheersing van een simpel AI-systeem. De schets geeft een niet-limitatief overzicht van bouwstenen voor de risicobeheersing van een simpel AI-systeem dat binnen een en dezelfde organisatie wordt ontwikkeld en ingezet (zie infographic). Bijvoorbeeld door een overheidsorganisatie. Dit is in zekere zin de eenvoudigste situatie die zich kan voordoen. Het beheersingsraamwerk wordt complexer zodra meerdere organisaties zijn betrokken en meerdere AI-modellen (algoritmes) met elkaar samenhangen in een proces dat zich baseert op een AI-systeem. In algemene zin kan gesteld worden dat het benodigde risicobeheersingsraamwerk een vorm van maatwerk moet zijn voor iedere individueel AI-systeem, afhankelijk van de doelstelling, autonomie en context waarbinnen het AI-systeem wordt ingezet.

Het fundament van de beheersing van een AI-systeem wordt geboden door (i) gedrag en cultuur en (ii) governance van de organisatie die betrokken is bij het AI-systeem. Vaak wordt bijvoorbeeld gesproken over het belang van ethisch besef, diversiteit en zorgvuldigheid door personen betrokken bij de ontwikkeling en inzet van AI-systemen. Dit zijn organisatiebrede vraagstukken. Hetzelfde geldt voor de AI-governance van een organisatie. Een goede governance geeft duidelijkheid wie (eind)verantwoordelijk is voor de inzet van AI binnen de organisatie en schept kaders voor hoe beheersing en kennis van AI vorm krijgen binnen de organisatie. Een voorbeeld van een overkoepelende organisatiebrede eis is te vinden in artikel 4 van de AI-verordening. Dit artikel schrijft voor dat organisaties die AI-systemen inzetten, moeten zorgen voor een toereikend niveau van AI-geletterdheid bij hun personeel.

Op organisatieniveau geldt vervolgens als eerste stap het doel bepalen van (de eventuele verkenning van) de inzet van een AI-systeem en een afgewogen besluit hiertoe nemen. Het scherp bepalen van het (expliciete of impliciete) doel van het AI-systeem biedt de basis voor het beoordelingskader: waarop moet het AI-systeem afgerekend worden? Het maakt ook een afweging mogelijk: welke voordelen biedt het AI-systeem? Met welke zekerheid kunnen deze voordelen behaald worden? En hoe wegen deze voordelen op tegen de nadelen en risico's en de (on) zekerheid of deze nadelen en risico's zich ook daadwerkelijk voordoen – inclusief bijbehorende beheersingskosten? Proportionaliteit speelt hierin ook een rol. Cruciaal is ook dat vooraf wordt vastgelegd dat er de intentie is om een AI-systeem in te zetten of te ontwikkelen. Bij publieke organisaties speelt hier ook democratische legitimiteit. Uit casuïstiek kan worden opgemaakt dat dit nog vaak tekortschiet – zie hoofdstuk 1 en 3, maar ook de eerste editie van de RAN (zomer 2023).

De continue risicobeheersing bij een AI-systeem binnen een organisatie vraagt vervolgens om een doorlopende cyclus van (i) (door)ontwikkeling en implementatie, (ii) inzet en (iii) evaluatie. Binnen elk onderdeel van de cyclus zijn er verschillende bouwstenen, die ieder op hun eigen manier bijdragen aan de beheersing van AI-risico's tijdens de gehele levenscyclus. Zo draagt het vooraf vaststellen van de te hanteren fairness- maatstaven gedurende de (door) ontwikkelings- en implementatiefase eraan bij dat na afloop, tijdens de evaluatiefase, getoetst kan worden of een AI-systeem aan de maatstaven voldoet.

Evengoed maakt het zekerstellen van registratie en transparantie van een AI-systeem het mogelijk dat tijdens de inzetfase informatie over het AI-systeem beschikbaar is voor het publiek. Dit draagt bij aan het kunnen melden (en verwerken) van incidenten met AI-systemen.

Tijdens de (door)ontwikkelings- en implementatiefase vindt uitgebreide toetsing plaats en worden de voorwaarden ingericht die daadwerkelijke inzet van het model-AI-systeem mogelijk moeten maken. Het schema geeft een weergave van enkele bouwstenen waaraan gedacht kan worden. Die vinden deels ook hun weerslag in vereisten uit de AI-verordening, maar ook andere wet- en regelgeving. Alvorens het systeem ingezet kan worden, moet worden voldaan aan vereisten op het gebied van kwaliteitsbeheersing (zie o.a. artikel 17 AI-verordening) en risicobeheersing – in feite, de cyclus van risico-identificatie, -beoordeling, -evaluatie en -beheersingsmaatregelen (zie o.a. artikel 9 AI-verordening). Bij de implementatie van een AI-systeem is een eerste bouwsteen een grondrechteneffectbeoordeling, die verplicht kan zijn voor gebruiksverantwoordelijken (zie o.a. artikel 27 AI-verordening). Deze beoordeling bestaat uit het identificeren, wegen en ondervangen van grondrechtenrisico's. Zo een beoordeling kan ook (gedeeltelijk) gepubliceerd worden, bijvoorbeeld in een register. Hieraan gerelateerd is de gegevensbescherming-seffectbeoordeling (DPIA) waarmee vooraf privacyrisico's in kaart worden gebracht en organisaties maatregelen kunnen nemen om die te verkleinen.

Ook moet er aandacht zijn voor testen, documentatie en registratie bij de ontwikkeling van AI. Onderdeel van de (door)ontwikkeling is het doorlopen van een testfase aan de hand van vooraf vastgestelde beoordelingsmaatstaven en betrouwbaarheidsdrempels die passend zijn voor het beoogde doel (zie o.a. artikel 9 AI-verordening). Het vaststellen van fairness-maatstaven is hieraan verbonden (zie o.a. artikel 10 AI-verordening), maar tegelijkertijd moet het AI-systeem nauwkeurig zijn in relatie tot de doelbepaling en niet tot willekeur leiden (zie o.a. artikel 15 AI-verordening). Registratie van een AI-systeem dat op de markt wordt gebracht, draagt bij aan transparantie en herleidbaarheid van een systeem en is voor AI-systemen met een hoog-risicoonderdeel van de conformiteitsprocedure die moet worden gevolgd bij het op de markt brengen van dergelijke systemen (zie o.a. artikel 49 AI-verordening). Evenzeer moet technische gebruiksdocumentatie worden opgesteld (zie o.a. artikel 11 en artikel 13 AI-verordening) die ook bij de implementatie van een AI-systeem het gebruik in goede banen moet leiden. Verder staat of valt de bruikbaarheid van een AI-systeem bij de datakwaliteit, waarvoor ook de datagovernance op orde moet zijn (zie o.a. artikel 10 AI-verordening).

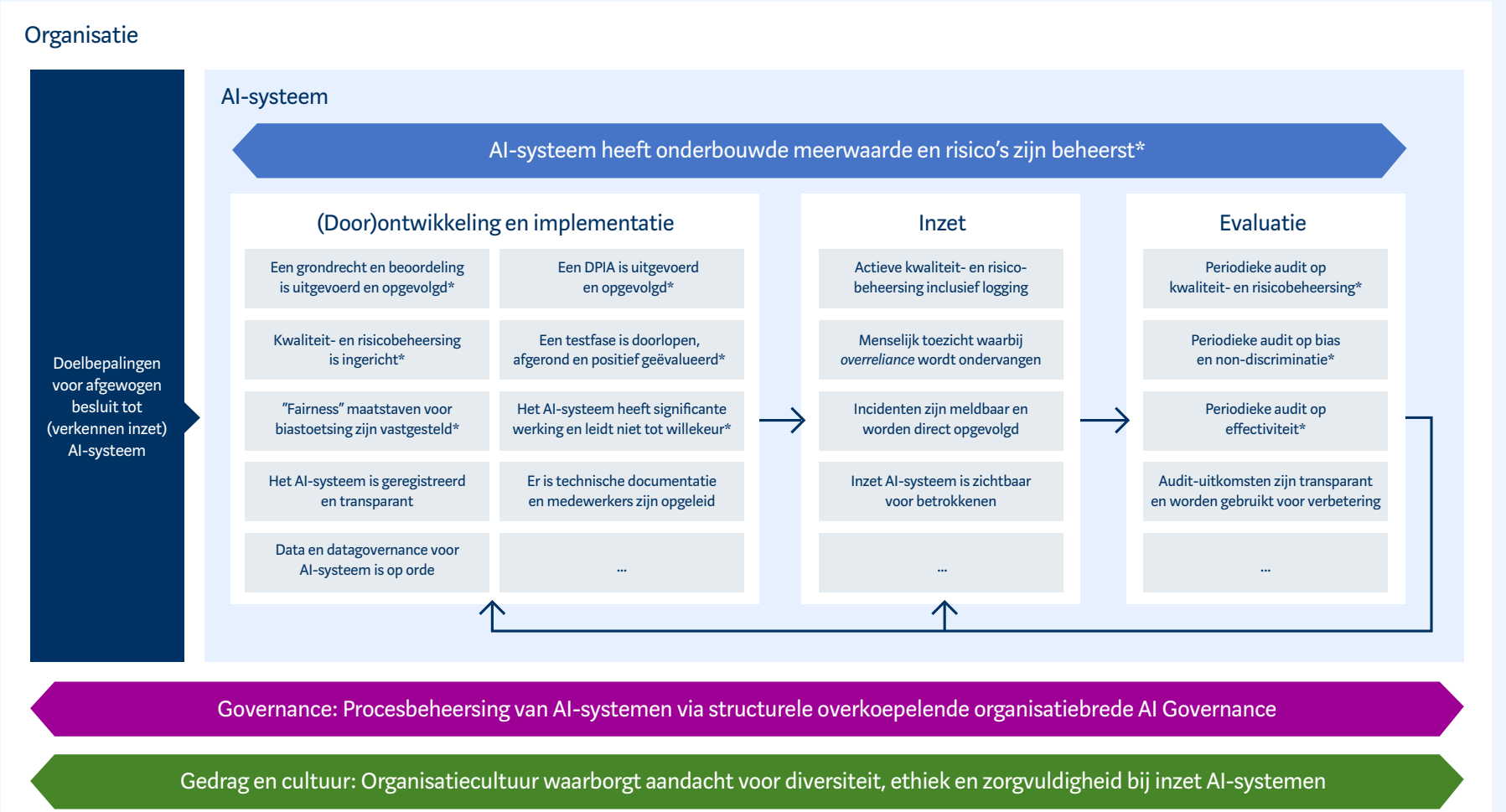
Tijdens de inzetfase moeten gebruikers en derden in staat zijn om gevaren van en incidenten met een AI-systeem snel te identificeren en te corrigeren. Een eerste bouwsteen daartoe is logging van de gebeurtenissen tijdens de inzet van het AI-systeem, zodat de oorzaak van incidenten achterhaald kan worden (vgl. artikel 12 AI-verordening dat ontwikkelaars verplicht tot inbouw van loggingmogelijkheden).

Overmatig vertrouwen in systemen kan beperkingen geven voor geautomatiseerde besluitvorming. Mocht een incident zich voordoen, dan moet een gebruiker of derde dit kunnen melden, zodat herstel mogelijk is (vgl. artikel 72 AI-verordening). Transparantie- en uitlegbaarheidsvereisten helpen de betrokkenheid van het AI-systeem zichtbaar te maken voor gebruikers of derden (zie o.a. artikel 50 en 86 AI-verordening, maar ook de AVG op het gebied van geautomatiseerde besluitvorming).

De periodieke evaluatie van het AI-systeem waarborgt dat het systeem doelmatig en kwalitatief in orde is en dat grondrechtenrisico's worden verkleind. Het evalueren is onderdeel van de beheersingscyclus die wordt voorgeschreven in het risicobeheersingsraamwerk voor AI-ontwikkelaars (zie o.a. artikel 9 AI-verordening). Dit is ook onderdeel van het systeem voor monitoring na ingebruikname van het AI-systeem. Hiertoe werkt de ontwikkelaar van het AI-systeem op basis van een plan voor monitoring. Cruciaal is de vastlegging van de evaluatie op een dusdanige manier dat noodzakelijke corrigerende of preventieve maatregelen worden vastgelegd. Deze maatregelen zijn vervolgens weer onderdeel van de doorontwikkeling van het AI-systeem.

Het in lijn brengen van AI met grondrechten en publieke waarden vereist acties tijdens de gehele levenscyclus van een AI

Schematische weergave van bouwstenen die bijdragen aan een beheersing van een simpel AI-systeem ontwikkeld en ingezet binnen één organisatie.



* Specialistische en/of externe ondersteuning ondersteunt de legitimiteit

Toelichting rapportage

Deze rapportage gaat over systemen en toepassingen van algoritmes en artificiële intelligentie (AI) die impact kunnen hebben op (groepen) personen.

Dit is de derde editie van de RAN, die halfjaarlijks verschijnt. De inhoud is gebaseerd op de kennis die is verkregen via het toezichtnetwerk van de AP. Zoals bureau-analyse en gesprekken met meer dan honderd relevante nationale en internationale organisaties. Maar de ontwikkelingen gaan snel en het zicht is op veel fronten nog onvolledig. Met dit in het achterhoofd probeert de AP toch een zo goed mogelijk beeld te vormen van actuele risico's en ontwikkelingen in beheersingsmaatregelen. En hieraan op een constructieve manier beleidsaanbevelingen te koppelen. Fouten of omissies in deze RAN zijn echter mogelijk.

AI-systemen automatiseren, in de kern, handelingen en beslissingen die mensen voorheen deden. Of die voorheen niet op deze manier mogelijk waren. Eenvoudig gezegd spreken we dan over algoritmes en AI. Dit strekt van relatief simpele toepassingen, waarin een enkel algoritme functioneert op basis van statische beslisregels, tot zeer complexe toepassingen van machine learning of neurale netwerken. De risicoanalyse in deze rapportage maakt geen onderscheid op basis van de technische werking van algoritmes en AI. Hiermee wordt aangesloten bij de beleidsconsensus over de betekenis van de term 'AI-systeem' (zie box 'AI-systeem als brede definitie').

De Rapportage AI- & Algoritmerisico's Nederland (RAN) beschrijft (trends en ontwikkelingen in) risico's. Dit zijn risico's bij de inzet van algoritmes en AI die individuele personen, groepen personen of de samenleving als geheel kunnen raken. En daarmee uiteindelijk ook de samenleving kunnen ontwrichten. De AP stelt de RAN op om belanghebbenden – private en publieke organisaties, politiek, beleidsmakers en het publiek – tijdig bewust te maken van deze risico's, zodat zij actie kunnen ondernemen. Bij de beschrijving van trends en ontwikkelingen in de risico's gelden twee kanttekeningen. Ten eerste brengt de inzet van algoritmes en AI niet alleen risico's mee, maar kan deze ook positieve bijdragen leveren, ook om fundamentele waarden en grondrechten juist te versterken. In het toezicht ligt de nadruk op (het wegnemen van) risico's. Ten tweede ligt de nadruk in deze periodieke rapportage op trends en ontwikkelingen. Dit betekent dat accenten worden gelegd in de analyse, in aanvulling op structurele risico's.

De RAN bevat geen voorspellingen. De AP wil met de huidige kennis en beschikbare informatie een compact en begrijpelijk beeld geven van de huidige risico's van de inzet van algoritmes en AI en de uitdagingen bij de beheersing van deze risico's. Waar mogelijk doet de AP voorstellen voor beleid dat risico's kan tegengaan.

Dit moet daarmee nog niet worden gezien als concrete guidance. De analyses en aanbevelingen in de RAN bieden organisaties en beleidsmakers inzichten om bij de inzet van algoritmes de kans op ongewenste effecten te verkleinen. Ook is de RAN te gebruiken om algoritmes en AI beter te begrijpen en de dialoog te versterken over kansen en risico's van algoritmes in de samenleving.

De RAN blijft pionierswerk en kan fouten bevatten. Nederland loopt mondiaal gezien voorop in het werken aan een zorgvuldige beheersing van algoritmes en AI, zodat de inzet hiervan ten dienste staat van mensen en de samenleving. De inrichting van het coördinerende AI- en algoritmetoezicht bij de AP en de periodieke systeemanalyses in deze RAN zijn daar voorbeelden van. Deze nieuwe taak is vorig jaar van start gegaan en is in opbouw. De eerste editie van de RAN (zomer 2023) ging in op de werkzaamheden van de DCA.

Kom met ons in contact. Uw reacties op de RAN en suggesties zijn welkom. U kunt die mailen aan dca@autoriteitpersoonsgegevens.nl

[1] OECD. (Maart 2024). Explanatory Memorandum on the Updated OECD Definition of an AI system. OECD Artificial Intelligence Papers, No. 8. https://www.oecd-ilibrary.org/docserver/623da898-en.pdf.

[2] Financial Times. (6 juni 2024). 'Most exciting moment' since birth of WiFi: chipmakers hail arrival of AI PCs. https://www.ft.com/content/6a546ad6-ae03-4c2d-92f5-c8efd-d4bba3b

[3] BCG. (21 september 2023). How People Can Create—and Destroy—Value with Generative AI [nieuwsbericht]. https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai

[4] Google. (30 mei 2024). AI Overviews: About last week. [nieuwsbericht]. https://blog.google/products/search/ai-overviews-update-may-2024/

[5] Microsoft. (7 juni 2024). Update on the Recall preview feature for Copilot+ PCs [nieuwsbericht]. https://blogs.windows.com/windowsexperience/2024/06/07/update-on-the-recall-preview-feature-for-copilot-pcs/

[6] Open AI. (19 mei 2024). How the voices for ChatGPT were chosen [nieuwsbericht]. https://openai.com/index/how-the-voices-for-chatgpt-were-chosen/

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2 augustus 2023). Attention is all you need. https://arxiv.org/abs/1706.03762

[8] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton. J., Kelton, F., Miller, J., Simens, M., Askell, A., Welinder P., Christiano, P., Leike, J., Lowe, R. OpenAI. (4 maart 2022). Training language models to follow instructions with human feedback https://arxiv.org/pdf/2203.02155

[9] AI Safety Institute (Verenigd Koninkrijk). (april 2024). Fourth Progress Report. https://www.aisi.gov.uk/work/fourth-progress-report.

[10] AI Seoul Summit. (21 mei 2024). Seoul Declaration for safe, innovative and inclusive AI by participants attending the leaders' session of the AI Seoul Summit. https://www.president.go.kr/download/664ca1113f0e7.

[11] AI Safety Institute (Verenigd Koninkrijk). (20 mei 2024). Advanced AI evaluations at AISI: May update. https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

[12] National AI Advisory Committee. (mei 2024). Finding & Recommendations: AI Safety. https://ai.gov/wp-content/uploads/2024/06/FINDINGS-RECOMMENDATIONS_AI-Safety.pdf.

[13] U.S. AI Safety Institute. (21 mei 2024). The US AI Safety Institute: Vision, Mission, and Strategic Goals. https://www.nist.gov/system/files/documents/2024/05/21/AISI-vision-21May2024.pdf.

[14] Autoriteit Persoonsgegevens en Rijksinspectie Digitale Infrastructuur. (16 mei 2024). Tweede (tussen) advies toezichtstructuur AI-verordening. https://www.autoriteitpersoonsgegevens.nl/system/files?-file=202406/20240516%20AI%20Act%20tweede%20tussen-advies.pdf.

[15] OECD. (2023). Using AI to support people with disability in the labour market. https://read.oecd.org/10.1787/008b32b7-en.

[16] UWV. (5 december 2023). Onderzoek toont aan: inclusieve technologie werkt. https://www.uwv.nl/nl/kennis-en-cijfers/uwv-als-kennisorganisatie/onderzoek-toont-aan-inclusieve-technologie-werkt.

[17] EU Funding & Tenders Portal. (2023). Innovative and Inclusive Democratic Spaces for Deliberation and Participation (iDEM). https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/org-details/999999999/project/101132431/program/43108390/details.

[18] MinBZK. (januari 2024). Overheidsbrede visie Generatieve AI. p. 41 Overheidsbrede visie Generatieve AI

[19] TNO. (4 april 2024). Generatieve AI in de Nederlandse zorg. p. 7-8. TNO2024 R10662 Generatieve AI in de Nederlandse zorg getekend | Publicatie | Gegevensuitwisseling in de zorg

[20] DNB, AFM. (2024). De impact van AI op de financiele sector en het toezicht. p. 23-25 (*IA PDF AI Rapport (dnb.nl)

[21] DNB, AFM. (2024). De impact van AI op de financiele sector en het toezicht. p. 26 *IA PDF AI Rapport (dnb.nl)

[22] Axis. (2024). Enhancing Safety and Security through Remote Surveillance. Enhancing Safety and Security through Remote Surveillance | Axis Communications

[23] Mobiliteit.nl. (7 juni 2024). NS test slimme camera die potentieel geweld herkent en vastlegt. https://www.mobiliteit.nl/ov/2024/06/07/ns-test-in-amsterdam-met-slimme-camera-om-potentieel-geweld-te-vast-te-leggen/?gdpr=deny&gdpr=deny

[24] Adams, R., Adeleke, F., Florido, A., de Magalhães Santos, L. G., Grossman, N., Junck, L., & Stone, K. (2024). Global Index on Responsible AI 2024 (1st Edition). South Africa: Global Center on AI Governance. The Global Index on Responsible AI (global-index.ai)

[25] Algemene Rekenkamer. (15 mei 2024). Resultaten verantwoordingsonderzoek 2023 Ministerie van Infrastructuur en Waterstaat Resultaten verantwoordingsonderzoek 2023 Ministerie van Infrastructuur en Waterstaat | Rapport | Algemene Rekenkamer

[26] Tweede Kamer der Staten- Generaal. (26 februari 2024). Rapport parlementaire enquêtecommissie Fraudebestrijding en Dienstverlening: 'Staatsmachten waren blind voor mens en recht' https://www.tweedekamer.nl/nieuws/

persberichten/rapport-parlementaire-enquetecommissie-fraudebestrijding-en-dienstverlening

[27] De Alliantie. Het woonfraude algoritme. https://www.de-alliantie.nl/over-de-alliantie/wat-we-doen/innovatie/innovaties/toekomstgerichte-organisatie/woonfraude/

[28] Berg, J. van den, Zwaan, I. de. (19 maart 2024). Onrust op basisscholen over uitkomsten nieuwe doorstroomtoets: 'Dit klopt gewoon niet'. De Volkskrant. https://www.volkskrant.nl/binnenland/onrust-op-basisscholen-over-uitkomsten-nieuwe-doorstroomtoets-dit-klopt-gewoon-niet~b2bebdb4/.

[29] Ministerie van Onderwijs, Cultuur en Wetenschap (3 juni 2014). Toetsbesluit PO. Staatsblad 2014, 209. https://zoek.officielebekendmakingen.nl/stb-2014-209.html.

[30] PO-Raad. (9 april 2024). PO-Raad duikt dieper in de cijfers rondom doorstroomtoets en organiseert masterclass toetsing. https://www.poraad.nl/po-raad-duikt-dieper-in-de-cijfers-rondom-doorstroomtoets-en-organiseert-masterclass-toetsing.

[31] Minister voor Primair en Voortgezet Onderwijs. (17 april 2024). Primair Onderwijs; Brief regering; Eerste afname doorstroomtoets. https://zoek.officielebekendmakingen.nl/kst-31293-729.html.

[32] Algoritmeregister Gemeente Amsterdam (2024). Onderzoekswaardigheid: Slimme check levensonderhoud https://algoritmeregister.amsterdam.nl/ai-system/onderzoekswaardigheid-slimme-check-levensonderhoud/1086/

[33] Het Parool. (14 februari 2024). Opinie: 'Onderzoek vooringenomenheid van zowel algoritme als ambtenaar' https://www.parool.nl/columns-opinie/opinie-onderzoek-voor-ingenomenheid-van-zowel-algoritme-als-ambtenaar~bd69aa5e/

[34] Zie voor de volledige documentatie die is beschikbaar gesteld door de gemeente Amsterdam het "Overzicht Verwerkte Data en Features" onder "Non-discriminatie" bij de beschrijving van het algoritme "Onderzoekswaardigheid: Slimme check levensonderhoud" (https://algoritmeregister.amsterdam.nl).

[35] Autoriteit Persoonsgegevens. (18 december 2023). Rapportage AI- & algoritmerisico's Nederland (RAN) - najaar 2023 (https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-ai-algoritmerisicos-nederland-ran-najaar-2023

[36] Rijksoverheid. (8 oktober 2019). Strategisch Actieplan voor Artificiële Intelligentie [beleidsnota]. Strategisch Actieplan voor Artificiële Intelligentie | Beleidsnota | Rijksoverheid.nl

[37] TNO Vector. (4 juni 2024). De economische waarde van strategische autonomie. De economische waarde van strategische autonomie - TNO Vector

[38] ACM FAccT conference. (2024). List accepted papers. https://facctconference.org/2024/acceptedpapers

[39] Autoriteit Persoonsgegevens Autoriteit Persoonsgegevens en Rijksinspectie Digitale Infrastructuur. (16 mei 2024). Tweede (tussen)advies toezichtstructuur AI-verordening. https://www.autoriteitpersoonsgegevens.nl/system/files?file=2024-06/20240516%20AI%20Act%20tweede%20tussenadvies.pdf

[40] Autoriteit Persoonsgegevens. (18 december 2023). Rapportage AI- & algoritmerisico's Nederland (RAN) - najaar 2023 (https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-ai-algoritmerisicos-nederland-ran-najaar-2023

[41] Commissariaat voor de Media. (juni 2024). Digital News Report Nederland 2024. 2031086-CvdM-DigitalNewsReport-2024_def.pdf

[42] Wired. (23 januari 2024). The Biden Deepfake Robocall Is Only the Beginning [nieuwsartikel]. https://www.wired.com/story/biden-robocall-deepfake-danger/

[43] The New York Times. (juni 2024). A Small Army Combating a Flood of Deepfakes in India's Election. [nieuwsartikel]. https://www.nytimes.com/2024/06/01/world/asia/india-election-deepfakes.html

[44] NRC. (30 mei 2024). "Alla yes on Rafah'gaat viral, toch is er ook kritiek: 'Dit is een AI-foto waardoor jouw ogen juist niet naar Rafah hoeven te kijken'. https://www.nrc.nl/nieuws/2024/05/30/is-het-protestbeeld-all-eyes-on-gaza-leunstoelactivisme-zo-kweek-je-bewustwording-bij-jongeren-a4200501?t=1717404010

[45] Waag Futurelab, Nederlandse AI Coalitie. (30 april 2024). Een maatschappelijke onderzoeksagenda voor AI. Eindrapport-Een-maatschappelijke-onderzoeksagenda-voor-AI.pdf (waag.org)

[46] Commissariaat voor de Media. (juni 2024). Digital News Report Nederland 2024. 2031086-CvdM-DigitalNewsReport-2024_def.pdf

[47] Autoriteit Persoonsgegevens. (18 december 2023). Rapportage AI- & algoritmerisico's Nederland (RAN) - najaar 2023 (https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-ai-algoritmerisicos-nederland-ran-najaar-2023

[48] Rijksoverheid. (23 december 2022). Kamerbrief over rijksbrede startegie effectieve aanpak van desinformatie. [kamerstuk]. https://open.overheid.nl/repository/ronl-d3369562e78345a02126dcd644ae9e6edc1a5b12/1/pdf/kamerbrief-over-rijksbrede-strategie-effectieve-aanpak-van-desinformatie.pdf

[49] NCTV. (25 juni 2024). Hoofdrapport Trendanalyse Nationale Veiligheid 2024. https://www.nctv.nl/binaries/nctv/documenten/rapporten/2024/06/25/hoofdrapport-trendanalyse-nationale-veiligheid-2024/Hoofdrapport+Trendanalyse+Nationale+Veiligheid+2024.pdf

[50] Rathenau. (13 oktober 2022). Digitale dreigingen voor de

democratie. p. 21. https://www.rathenau.nl/sites/default/files/2020-10/RAPPORT_Digitale_dreigingen_voor_de_democratie_Rathenau_Instituut.pdf

[51] NOS (3 mei 2024). Chatbots adviseerden: verspreid desinformatie en zaai angst over EU-verkiezingen. [Nieuwsartikel]. Chatbots adviseerden: verspreid desinformatie en zaai angst over EU-verkiezingen (nos.nl)

[52] NRC. (31 mei 2024). Hoe Google zijn zoekmachine op de schop gooit, en daarmee het hele internet. [nieuwsartikel]. Hoe Google zijn zoekmachine op de schop gooit, en daarmee het hele internet - NRC

[53] De Volkskrant. (24 mei 2024). Nieuwe AI-zoekdienst Google: doe lijm op je pizza en stop chloorgas in je wasmachine. [nieuwsartikel]. Nieuwe AI-zoekdienst Google: doe lijm op je pizza en stop chloorgas in je wasmachine | de Volkskrant

[54] De Groene Amsterdammer. (5 juni 2024). Verantwoording bij het onderzoek naar AI-boeken op Bol. Verantwoording bij het onderzoek naar AI-boeken op Bol – De Groene Amsterdammer

[55] The Verge. (6 februari 2024). Meta says you better disclose your AI fakes or it might just pull them. [nieuwsartikel]. Meta to label AI-generated images on Facebook, Instagram, Threads - The Verge

[56] Content Autheniticity Initiative. (2024). Restoring trust and transparency in the age of AI. Content Authenticity Initiative

[57] BBC. (4 maart 2024). Haiti violence: Haiti gangs demand PM resign after mass jailbreak. [nieuwsartikel]. Haiti violence: Haiti gangs demand PM resign after mass jailbreak (bbc.com)

[58] Arnold, M., Goldschmitt, M., Rigotti, T.(21 juni 2023). Dealing with information overload: a comprehensive review. Dealing with information overload: a comprehensive review - PMC (nih.gov)

[59] Forbes advisor. (4 juni 2024). Top website statistics for 2024. Top Website Statistics For 2024 – Forbes Advisor

[60] The Brussels Times. (15 februari 2024). TikTok overtakes Google as most popular search engine among Gen Z [nieuwsartikel]. TikTok overtakes Google as most popular search engine among Gen Z (brusselstimes.com)

[61] European Parlement. (8 november 2023). REPORT on addictive design of online services and consumer protection in the EU single market. REPORT on addictive design of online services and consumer protection in the EU single market | A9-0340/2023 | European Parliament (europa.eu)

[62] Handvest van de grondrechten van de Europese Unie. Artikel 11 Artikel 11 - De vrijheid van meningsuiting en van informatie | European Union Agency for Fundamental Rights (europa.eu)

[63] Funk, A., Shahbaz, A., Vesteinsson, K. (2023) Freedom House. Freedom on the net 2023. The repressive power of artificial intelligence. p.16 https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[64] The Economic Times India. (22 februari 2024). X disagrees with govt blocking orders for certain posts and accounts. [nieuwsartikel]. https://economictimes.indiatimes.com/tech/technology/x-complies-with-govt-blocking-orders-for-certain-posts-accounts/articleshow/107904325.cms

[65] DW. (17 april 2024). X blocks posts in India after election commission order. [nieuwsartikel]. https://www.dw.com/en/x-blocks-posts-in-india-after-election-commission-order/a-68846469

[66] Funk, A., Shahbaz, A., Vesteinsson, K. (2023) Freedom House. Freedom on the net 2023. The repressive power of artificial intelligence. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[67] ACM. Betrouwbare flaggers. Betrouwbare flaggers | ACM.nl

[68] Commissariaat voor de media. (juni 2024). Tussen Bits en Principes: Hoe AI de kernwaarden van mediabeleid uitdaagt. AI-Verkenning-CvdM-Tussen-Bits-en-Principes.pdf

[69] AIVD. (2023) Jaarverslag 2022. https://www.aivd.nl/onderwerpen/jaarverslagen/jaarverslag-2022

[70] Defend Democracy. (19 april 2024). Automated anarchy: the rising tide of bad bots on the internet. https://defenddemocracy.eu/automated-anarchy-the-rising-tide-of-bad-bots-on-the-internet/

[71] The Guardian. (19 juli 2023). Disinformation reimagined: how AI could erode democracy in the 2024 US elections. [nieuwsartikel]. Disinformation reimagined: how AI could erode democracy in the 2024 US elections | US elections 2024 | The Guardian

[72] Rathenau Instituut. (13 oktober 2020). Digitale dreigingen voor de democratie, p. 55 Digitale dreigingen voor de democratie | Rathenau Instituut

[73] Funk, A., Shahbaz, A., Vesteinsson, K. (2023) Freedom House. Freedom on the net 2023. The repressive power of artificial intelligence. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[74] UCL, University of Kent. (2 februari 2024). Safter scrolling: How algorithms popularise and gamify online hate and misogyny for young people. Safer-scrolling.pdf (ascl.org.uk)

[75] Funk, A., Shahbaz, A., Vesteinsson, K. (2023) Freedom House. Freedom on the net 2023. The repressive power of artificial intelligence. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[76] NRC. (19 maart 2024). Ook bekende Nederlanders

slachtoffer deepfake porno: schending seksuele privacy. [nieuwsartikel]. Ook bekende Nederlanders slachtoffer deepfake porno: schending seksuele privacy - NRC

77 Commissariaat voor de media. (juni 2024). Digital News Report Nederland 2024 2031086-CvdM-DigitalNewsReport-2024_def.pdf

78 Commissariaat voor de media. (juni 2024). Digital News Report Nederland 2024 2031086-CvdM-DigitalNewsReport-2024_def.pdf

79 Wagner, M. (2024). Affective polarization in Europe. European Political Science Review, 1–15. doi:10.1017/S1755773923000383 Affective polarization in Europe | European Political Science Review | Cambridge Core

80 Mediawijheid.nl. Wat is mediawijsheid Wat is mediawijsheid? - Mediawijsheid.nl

81 MinBZK. (7 juni 2024). Voortgangsbrief Rijksbrede strategie voor de effectieve aanpak van desinformatie en aankondiging nieuwe acties. Brief - Voortgangsbrief Rijksbrede strategie voor de effectieve aanpak van desinformatie en aankondiging nieuwe acties (overheid.nl)

82 Rijksoverheid. Digitale geletterdheid op school. Digitale geletterdheid op school | Digitalisering in het onderwijs | Rijksoverheid.nl

83 SLO. (6 maart 2024). Conceptkerndoelen leergebied digitale geletterdheid + toelichtingsdocument. Conceptkerndoelen leergebied digitale geletterdheid + toelichtingsdocument - SLO

84 Publicatieblad van de Europese Unie. (26 april 2024). Mededeling van de Commissie — Richtsnoeren van de Commissie voor aanbieders van zeer grote online-platforms en zeer grote onlinezoekmachines inzake de beperking van systeemrisico's voor verkiezingsprocessen overeenkomstig artikel 35, lid 3, van Verordening (EU) 2022/2065 (europa.eu)

85 Europese Commissie. (24 april 2024). Commission stress tests platforms' election readiness under the Digital Services Act. [nieuwsartikel]. Commission stress tests platforms' election readiness under the Digital Services Act | Shaping Europe's digital future (europa.eu)

86 Europese Commissie. (30 april 2024). Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act. [Press release]. Commission opens formal proceedings under DSA (europa.eu)

87 TNO. (Juni 2024). Quickscan AI in de Publieke Dienstverlening (TNO 2024 R11005). https://publications.tno.nl/publication/34642601/SASNc3ZW/TNO-2024-R11005.pdf.

88 Zie TNO. (Juni 2024).

89 Algoritmeregister. 2024. Geautomatiseerde documentcheck en gezichtsvergelijker (Gemeente 's Hertogenbosch). https://algoritmes.overheid.nl/nl/algoritme/38325188.

90 Algoritmeregister. (2024). Chatbot Guus (AI Versie) (Gemeente Goes). https://algoritmes.overheid.nl/nl/algoritme/15943226.

91 Algoritmeregister. (2024). Vroeg eropaf (Gemeente Amsterdam). https://algoritmes.overheid.nl/nl/algoritme/66453169.

92 Zo becijferde RTL nieuws nog in januari 2020 dat op dat moment 25% van de Nederlanders toen al in een gemeente woonde die scanauto's gebruikt. RTL nieuws. 16 januari 2020. Scanauto's leveren gemeentes miljoenen op (en het worden er steeds meer. Scanauto's leveren gemeentes miljoenen op (en het worden er steeds meer) | RTL Nieuws | RTL.nl

93 Autoriteit Persoonsgegevens. (2024). Jaarverslag 2023. https://www.autoriteitpersoonsgegevens.nl/documenten/ap-jaarverslag-2023.

94 Nederlandse Vereniging voor Raadsleden. Taken Gemeenteraad. Taken gemeenteraad | Nederlandse Vereniging voor Raadsleden

95 Gemeentewet, art. 182(1). wetten.nl - Regeling - Gemeentewet - BWBR0005416 (overheid.nl)

96 Algemene Wet Bestuursrecht 9(22-27).

97 Raad voor het Openbaar Bestuur en Raad voor Cultuur. 2020. Lokale media: niet te missen. p. 22. Adviesrapport Lokale media: niet te missen | Publicatie | Raad voor het Openbaar Bestuur (raadopenbaarbestuur.nl)

98 Zie TNO. (Juni 2024).

99 Hooghiemstra & Partners. (Juni 2021). Hoe gemeenten besluiten over algoritmen & mensenrechten. p. 7 en 16. https://publicaties.mensenrechten.nl/publicatie/60dd2c7b98d7821c6468363e.

100 Hooghiemstra & Partners. (Juni 2021). Hoe gemeenten besluiten over algoritmen & mensenrechten. https://publicaties.mensenrechten.nl/publicatie/60dd2c7b98d7821c6468363e.

101 Zie Hooghiemstra & Partners. (Juni 2021). p. 19.

102 Rathenau Instituut. (September 2020). Raad weten met digitalisering. p. 4. https://www.rathenau.nl/nl/kennis-voor-transities/raad-weten-met-digitalisering.

103 Raad voor het openbaar bestuur. (2021). Sturen of gestuurd worden? Over de legitimiteit van sturen met data. https://www.raadopenbaarbestuur.nl/documenten/publicaties/2021/05/25/advies-sturen-of-gestuurd-worden.

104 Rathenau. (september 2020). Raad weten met digitalisering. p. 20

105 Raad voor het Openbaar Bestuur. (November 2020). Goede ondersteuning sterke democratie: over de ondersteuning van decentrale volksvertegenwoordiging. p. 20. Adviesrapport Goede ondersteuning, sterke democratie | Publicatie | Raad voor het Openbaar Bestuur (raadopenbaarbestuur.nl).

[106] Rekenkamer Metropool Amsterdam. (Oktober 2023). Algoritmen: hoe Amsterdam algoritmen beter kan toepassen. Onderzoeksrapport-Algoritmen-DEF.pdf (amsterdam.nl)

[107] Rekenkamer Rotterdam. (2024). Kleur Bekennen: vervolgonderzoek naar algoritmes. kleur bekennen (rotterdam.nl).

[108] Rekenkamer Den Haag. (Maart 2024). Jaarverslag 2023 Rekenkamer Den Haag. RIS318254_Jaarverslag_2023_Rekenkamer_Den_Haag.pdf (rekenkamerdenhaag.nl)

[109] Stimuleringsfonds voor de Journalistiek. (Juni 2022). Van Stoom naar Stroom: De weg naar professionalisering. p. 29. Van Stoom naar Stroom: de weg naar professionalisering - SVDJ

[110] Hooghiemstra & Partners. Juni 2021. Hoe gemeenten besluiten over algoritmen & mensenrechten. p. 20. https://publicaties.mensenrechten.nl/publicatie/60dd2c7b98d7821c6468363e.

[111] Hooghiemstra & Partners. (Juni 2021). Hoe gemeenten besluiten over algoritmen & mensenrechten. p. 16. https://publicaties.mensenrechten.nl/publicatie/60dd2c7b98d7821c6468363e.

[112] Raad voor het Openbaar Bestuur. (November 2020). Goede ondersteuning sterke democratie: over de ondersteuning van decentrale volksvertegenwoordiging. p. 35. Adviesrapport Goede ondersteuning, sterke democratie | Publicatie | Raad voor het Openbaar Bestuur (raadopenbaarbestuur.nl).

[113] BDO. (Januari 2024). Benchmark Nederlandse gemeenten 2024. P. 22-24. Download nu de BDO-Benchmark Nederlandse gemeenten 2024

[114] Wetenschapelijke raad voor het regeringsbeleid. (2021). Opgave AI. De nieuwe systeemtechnologie. p. 437-442 Opgave AI. De nieuwe systeemtechnologie | Rapport | WRR

[115] MinBZK. (17 juni 2024). Verzamelbrief Digitalisering juni 2024. Brief – Verzamelbrief Digitalisering juni 2024 (overheid.nl)

[116] Schrijver, G., Sarmah, D. K., & El-Hajj, M. (2024). Automobile Insurance Fraud Detection Using Data Mining: A Systematic Literature Review. Intelligent Systems With Applications, 200340. https://doi.org/10.1016/j.iswa.2024.200340

[117] DNB AFM. (2024). De impact van AI op de financiële sector en het toezicht. https://www.dnb.nl/nieuws-voor-de-sector/toezicht-2024/afm-en-dnb-publiceren-rapport-over-de-impact-van-ai-in-de-financiele-sector-en-het-toezicht-daarop/

[118] Meta. (3 augustus 2020). How Does Facebook Measure Fake Accounts? https://about.fb.com/news/2019/05/fake-accounts/

[119] Dienst Toeslagen, Ministerie van Financiën. (18 maart 2024). Gebruik van dubbele nationaliteit. https://herstel.toeslagen.nl/gebruik-van-dubbele-nationaliteit/

[120] Tweede Kamer Der Staten-Generaal. (26 februari 2024). Rapport parlementaire enquêtecommissie Fraudebestrijding en Dienstverlening: 'Staatsmachten waren blind voor mens en recht'. https://www.tweedekamer.nl/nieuws/persberichten/rapport-parlementaire-enquetecommissie-fraudebestrijding-en-dienstverlening

[121] Rechtbank Den Haag. (5 februari 2020). 6.87-6.94, ECLI:NL:RBDHA:2020:865. https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:865

[122] DUO. (2024). Excuses voor indirecte discriminatie bij controles op de uitwonendenbeurs. https://duo.nl/organisatie/pers/excuses-voor-indirecte-discriminatie-bij-controles-op-de-uitwonendenbeurs.jsp

[123] Rensen, Frank. 2 juli 2024. 'Bij zo'n beetje elke tegel die we lichten, ontdekken we discriminerende algoritmes bij de overheid', zegt de Autoriteit Persoonsgegevens. De Volkskrant. https://www.volkskrant.nl/tech/bij-zo-n-beetje-elke-tegel-die-we-lichten-ontdekken-we-discriminerende-algoritmen-bij-de-overheid-zegt-de-autoriteit-persoonsgegevens~b360ed6e

[124] Dit wijkt af van een technische interpretatie van discriminatie waarbij het regelmatig als synoniem voor onderscheid gebruikt wordt

[125] College Voor de Rechten van de Mens. (2022). Vraag 4, Vraag en antwoord over werving- en selectie-algoritmes voor werkgevers. https://www.mensenrechten.nl/themas/digitalisering/werving-en-selectie/qa-over-hr-algoritmes-voor-werkgevers

[126] College Voor de Rechten van de Mens. (2021). Discriminatie door risicoprofielen - Een mensenrechtelijk toetsingskader. https://publicaties.mensenrechten.nl/publicatie/61a734e65d726f72c45f9dce

[127] In een casus bij de VU is er een aannemelijk vermoeden van discriminatie aangetoond bij AI-technologie die niet goed werkte voor een student met een donkere huidskleur. Volgens het eindoordeel is er hier niet gediscrimineerd. Zie: College Voor de Rechten van de Mens. (17 oktober 2023). Student niet gediscrimineerd door tentamensoftware Proctorio, maar VU had de klacht zorgvuldiger moeten behandelen. https://www.mensenrechten.nl/actueel/nieuws/2023/10/17/student-niet-gediscrimineerd-door-tentamensoftware-proctorio-maar-vu-had-de-klacht-zorgvuldiger-moeten-behandelen

[128] FRA. (2021). Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf

[129] Europese Commissie. (22 mei 2023). C(2023)3215 – Standardisation request M/593 COMMISSION IMPLEMENTING

DECISION of 22.5.2023 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence https://ec.europa.eu/growth/tools-databases/enorm/mandate/593_en

[130] Europese Commissie. (September 2023). EU model contractual AI clauses to pilot in procurements of AI. https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/eu-model-contractual-ai-clauses-pilot-procurements-ai.

[131] Autoriteit Persoonsgegevens en Rijksinspectie Digitale Infrastructuur. (16 mei 2024). Tweede (tussen) advies toezichtstructuur AI-verordening. https://www.autoriteitpersoonsgegevens.nl/system/files?-file=202406/20240516%20AI%20Act%20tweede%20tussenadvies.pdf.

[132] Autoriteit Persoonsgegevens en Rijksinspectie Digitale Infrastructuur. (16 mei 2024). Tweede (tussen) advies toezichtstructuur AI-verordening. https://www.autoriteitpersoonsgegevens.nl/system/files?-file=202406/20240516%20AI%20Act%20tweede%20tussenadvies.pdf.

[133] Klein, E. & Patrick, S. (21 maart 2024). Envisioning a Global Regime Complex to Govern Artificial Intelligence. https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en

[134] Klein, E. & Patrick, S. (21 maart 2024). Envisioning a Global Regime Complex to Govern Artificial Intelligence. https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en

[135] Roberts, H., Cowls, J., Hine, E., Morley, J., Wang, V., Taddeo, M., & Floridi, L. (2022). Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes. The Information Society, 39(2), 79–97. https://doi.org/10.1080/01972243.2022.2124565

[136] Ryan-Mosley, T., Heikkilä, M., Yang, Z. (5 januari 2024). What's next for AI regulation in 2024? https://www.technologyreview.com/2024/01/05/1086203/whats-next-ai-regulation-2024/

[137] Braziliaans voorstel voor regels AI. (2024). https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

[138] Schertel Mendes, L. & Kira, B. (21 december 2023). The road to regulation of artificial intelligence: the Brazilian experience. https://policyreview.info/articles/news/road-regulation-artificial-intelligence-brazilian-experience/1737

[139] Klein, E. & Patrick, S. (21 maart 2024). Envisioning a Global Regime Complex to Govern Artificial Intelligence. https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en

[140] OECD. Update to the OECD I principles. (2024). https://oecd.ai/en/ai-principles

[141] United Nationas News. (21 maart 2024). General Assembly adopts landmark resolution on artificial intelligence. https://news.un.org/en/story/2024/03/1147831

[142] United Nations. (21 maart 2024). General Assembly Adopts Landmark Resolution on Steering Artificial Intelligence towards Global Good, Faster Realization of Sustainable Development General Assembly Adopts Landmark Resolution on Steering Artificial Intelligence towards Global Good, Faster Realization of Sustainable Development | Meetings Coverage and Press Releases (un.org)

[143] Zo liggen de OECD Privacyrichtlijnen, die grenzen stellen aan het verzamelen en gebruiken van persoonsgegevens, ten grondslag aan veel privacywetten. Forty-two countries adopt new OECD Principles on Artificial Intelligence - OECD

[144] United Nations Ai advisory board. (December 2023). Interim Report: Governing AI for humanity. https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf

[145] Autoriteit Persoonsgegevens. (April 2024). Supervisory Perspective on Global AI Governance (Discussion Paper). https://www.autoriteitpersoonsgegevens.nl/en/documents/supervisory-perspective-on-global-ai-governance-discussion-paper

[147] Autoriteit Persoonsgegevens. (April 2024). Supervisory Perspective on Global AI Governance (Discussion Paper). https://www.autoriteitpersoonsgegevens.nl/en/documents/supervisory-perspective-on-global-ai-governance-discussion-paper.

[147] Autoriteit Persoonsgegevens. (7 december 2023). Blogpost: zorgen om generatieve AI. https://www.autoriteitpersoonsgegevens.nl/actueel/blogpost-zorgen-om-generatieve-ai.

AUTORITEIT
PERSOONSGEGEVENS