Rapportage AI- & Algoritmerisico's Nederland

Rapportage najaar 2023

Autoriteit Persoonsgegevens | Directie Coördinatie Algoritmes (DCA)

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

AUTORITEIT
PERSOONSGEGEVENS

Inhoudsopgave

Toelichting rapportage

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

Deze AI-systemen automatiseren, in de kern, handelingen en beslissingen die mensen voorheen deden. Of die niet op deze manier mogelijk waren. Eenvoudig gezegd: het gaat over algoritmes en AI. Dit strekt van relatief simpele toepassingen, waarin een enkel algoritme functioneert, tot zeer complexe toepassingen van ***machine learning*** of neurale netwerken. De risicoanalyse in deze rapportage maakt hier geen onderscheid in. De Directie Coördinatie Algoritmes (DCA) van de Autoriteit Persoonsgegevens (AP) monitort de mogelijke effecten van de inzet van algoritmes en AI voor publieke waarden en grondrechten. En rapporteert daar periodiek over. Zo draagt de AP bij aan verantwoordere inzet van algoritmes.

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 wordt opgemerkt dat de inzet van algoritmes en AI niet alleen risico's meebrengt maar ook een positieve bijdrage kan leveren, inclusief ter versterking van publieke waarden en grondrechten. Maar gegeven de rol van de AP in het algoritmetoezicht ligt de nadruk op (het wegnemen van) risico's. Ten tweede ligt de nadruk in deze periodieke rapportage op trends en ontwikkelingen in algoritme- en AI-risico's. Dit betekent dat accenten worden gelegd in de analyse, in aanvulling op structurele risico's die aanwezig zijn.

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.

In deze rapportage. Het eerste hoofdstuk van deze rapportage beschrijft op hoofdlijnen de voor Nederland belangrijkste recente overkoepelende ontwikkelingen bij de inzet van algoritmes en de risicobeheersing daarvan. Het tweede hoofdstuk gaat in op de ontwikkelingen en uitdagingen bij generatieve AI en ***foundation models***. Het derde en vierde hoofdstuk zijn thematisch en gaan in op algoritmes en AI op de werkvloer en in het onderwijs. Het vijfde hoofdstuk tot slot schenkt aandacht aan beleidsontwikkeling en institutionele kaders op nationaal en internationaal niveau.

Algoritmetoezicht in opbouw

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 algoritmetoezicht bij de AP en de periodieke systeemanalyses in deze RAN zijn daar een voorbeeld van. De DCA is dit jaar van start gegaan en is in opbouw. De eerste editie van de RAN is ingegaan op de werkzaamheden van de DCA.

Dit is de tweede 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.

Uw reacties op de RAN en suggesties voor verbeteringen zijn welkom. U kunt die mailen naar dca@autoriteitpersoonsgegevens.nl.

Kernboodschappen

1. Het overkoepelend risicobeeld blijft vragen om actie.

De voorzichtige inschatting is dat de risico's van algoritmes en AI in het afgelopen jaar over de volle breedte zijn toegenomen. Een belangrijke oorzaak hiervan is dat nieuwe (generatieve) AI-systemen snel op de markt komen en steeds meer worden ingezet. Hoewel wordt gewerkt aan meer en betere risicobeheersing, ligt het tempo bij algoritmes en AI lager dan de innovatiesnelheid. Dit komt mede doordat wettelijke kaders en productstandaarden nog in ontwikkeling zijn. Adequate beheersing van risico's is belangrijk om het vertrouwen in de technologie te verbeteren en innovatiemogelijkheden te ontsluiten. Dit geldt ook bij de inzet van nieuwe technologieën waarbij de risico's nog niet bekend zijn en burgers met mogelijk ongewenste effecten in aanraking kunnen komen.

2. We weten te weinig over waar en wanneer het mis gaat.

Slechts enkele toezichthouders ontvangen op dit moment signalen over incidenten met algoritmes en AI binnen hun toezichtgebied. Een oorzaak kan liggen in een gebrek aan transparantie over de rol van algoritmes en AI en in een gebrek aan rapportageverplichtingen. Tegelijkertijd wordt wereldwijd over meer AI-incidenten geschreven. De opkomst van generatieve AI speelt hierin een belangrijke rol. Voor meer transparantie is het (Nederlandse) Algoritmeregister voor overheidsorganisaties belangrijk. De AP observeert in de tweede helft van 2023 een significante toename van het aantal geregistreerde algoritmes. Het is belangrijk dat overheidsorganisaties nog sneller hun algoritmes registreren. En dat het denkwerk over verplichte registratie van toepassingen met een hoog risico concreter vorm krijgt.

3. Nederland voorbereiden op de toekomst via een deltaplan algoritmes en AI.

Voor nieuwe toepassingsmogelijkheden van algoritmes en AI is terecht veel aandacht. Het gebruik hiervan zal de komende jaren een verdere vlucht nemen. Maar er is meer samenhangende grip nodig op algoritmes en AI om daarvan als samenleving op een verantwoorde manier de vruchten te plukken. Dit gaat om meer dan alleen de inrichting van toezicht. Er zijn veel knoppen waaraan gedraaid kan en moet worden en het is van belang deze maatschappelijke agenda in samenhang te bezien. De AP adviseert om te denken in de richting van een Nederlands deltaplan voor algoritmes en AI, dat zich richt op vijf pijlers: (1) menselijke regie, (2) veilige applicaties en systemen, (3) organisaties in control, (4) nationaal ecosysteem & nationale infrastructuur en (5) internationale standaarden en samenwerking. Het deltaplan moet gezien worden in de context van de Europese markt en de open Nederlandse economie. Hoofdstuk 1 gaat hier nader op in.

4. De veelzijdige inzetbaarheid van generatieve AI vraagt om passend toezicht.

In korte tijd is het gebruik van generatieve AI diep doorgedrongen in de Nederlandse samenleving. Bij het gebruik van generatieve AI spelen vragen over de rechtmatigheid ervan. Ook nieuwe vormen van gebruiksrisico's en systeemrisico's zijn aandachtspunten. Het is van belang om verder te werken aan de concretisering van het Europese toezicht op ***foundation models*** (die de basis vormen voor toepassingen van generatieve AI) en de organisaties die deze modellen ontwikkelen. De AI-verordening biedt hiervoor de basis. Hoofdstuk 2 gaat hier nader op in.

5. Steeds meer werkenden worden aangestuurd door algoritmes.

Van buschauffeur tot servicemonteur tot thuiszorgmedewerker: algoritmes en AI spelen een steeds grotere rol in de aansturing van arbeid, zowel bij platformwerk als bij traditionele banen. Dit kan mensen ten goede komen, bijvoorbeeld als het hun fysieke veiligheid vergroot. Maar het kan ook negatieve gevolgen hebben voor werkdruk, autonomie en rechtvaardigheid in de arbeidsrelatie en bij de beoordeling van arbeidsprestaties. Europese wetgeving voor algoritmisch management in platformwerk is in de maak. De AP ziet ruimte om ook in bredere zin te werken aan meer transparantie over en betwistbaarheid van de rol van algoritmes in de aansturing van arbeid. Hoofdstuk 3 gaat hier nader op in.

6. Ook het leeraanbod voor scholieren en studenten wordt meer en meer beïnvloed door algoritmes en AI.

In de onderwijsaansturing wordt veel gebruik gemaakt van algoritmes en AI: van individuele adaptieve leersystemen op de basisschool tot analysetools in het beroeps- en wetenschappelijk onderwijs om de doorstroming van studenten te bevorderen. Er zijn echter allerlei oorzaken denkbaar waardoor de onderwijsprofilering of -voorspelling niet goed aansluit bij de situatie van een leerling. Zorgvuldige inbedding van AI in de onderwijsaanpak en bewustzijn onder docenten en schoolbesturen van de beperkingen van deze applicaties is daarom

cruciaal. De AP adviseert onderwijsinstellingen AI-inzet en AI-beheersingsprocessen mee te nemen in de inrichting van de ICT-strategie, met voldoende ondersteuning door interne of externe expertise. Ook het vergroten van AI-kennis onder docenten is een aandachtspunt. Hoofdstuk 4 gaat hier nader op in.

7. Het fundament voor toezicht op algoritmes en AI is bijna klaar, het is nu tijd de steigers te plaatsen.

Op 9 december is op Europees niveau een politiek akkoord gesloten over de AI-verordening. De komende maanden worden details verder uitgewerkt om de verordening medio 2024 in werking te laten treden, met daaropvolgend een overgangsperiode alvorens regels van toepassing worden. Daarmee is een belangrijke stap gezet die zal bijdragen aan de beheersing en controle van AI. De praktijk zal moeten uitwijzen of de verordening een adequate basis gaat vormen voor veilige algoritmes en AI. In Nederland bereiden toezichthouders zich gezamenlijk voor op de bijbehorende toezichttaak. Beheersing van algoritmes en AI vraagt echter om meer dan alleen toezicht op productniveau. Het is belangrijk dat ook wordt gewerkt aan intern toezicht plus externe controle op organisatorisch niveau. En dat wordt geïnvesteerd in samenwerking tussen meldpunten en loketten om algoritmische misstanden aan het licht te brengen. De AP ambieert vanuit haar coördinerende algoritmetaak deze ontwikkelingen aan te jagen en te faciliteren. Hoofdstuk 5 gaat hier nader op in.

1. Overkoepelende ontwikkelingen

Hoofdlijn

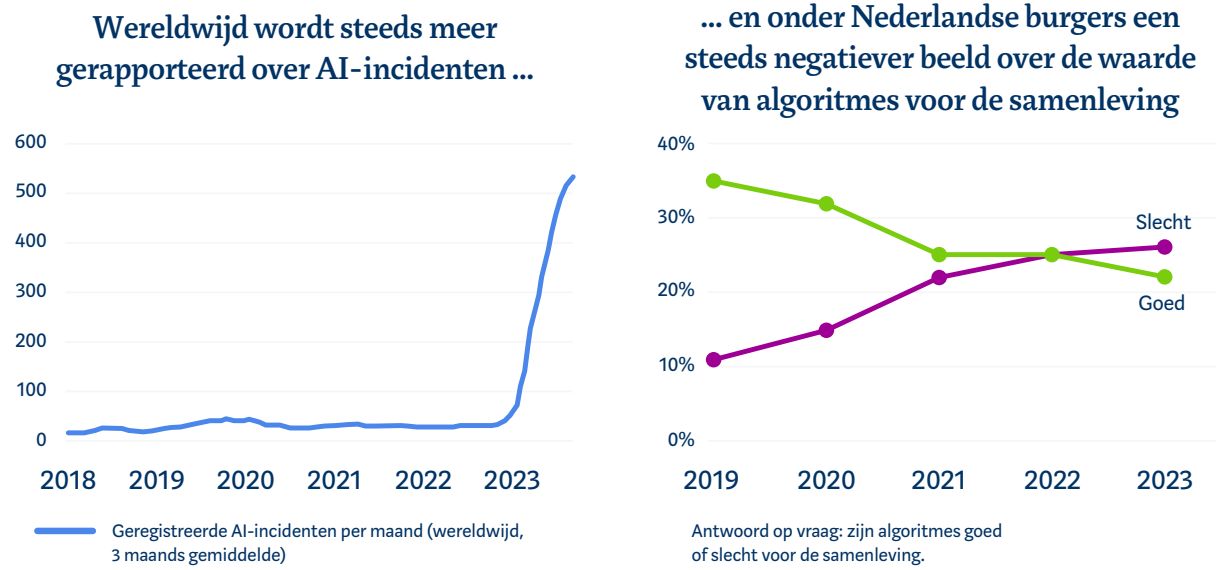
Risico's van algoritmes en AI blijven aanwezig in Nederland en nemen zelfs toe, maar er worden stappen gezet om deze risico's structureel aan te pakken. De ontwikkelingen gaan onverminderd snel. Waar een jaar geleden taalmodellen doorbraken voor het grote publiek, zijn deze modellen nu verfijnder en op steeds meer manieren toe te passen. Tegelijkertijd zien we door deze aandacht en brede toepassing dat de bekende, maar ook onbekende risico's zich daadwerkelijk materialiseren.

De OECD AI Incidents Monitor laat een significante toename zien. Het gaat bij deze OECD monitor om wereldwijde incidenten die in nieuwsartikelen zijn beschreven[1]. Figuur 1 toont de sterke groei van incidenten – een vertienvoudiging (961%) ten opzichte van vorig jaar. Het toenemende aantal bekende AI-incidenten toont de diverse risico's voor de publieke waarden en mensenrechten. Waarschijnlijk zijn er zoveel meer incidenten aan het licht gekomen in 2023 omdat generatieve AI steeds vaker wordt ingezet en dit in 2023 veel aandacht heeft gekregen. Het laat ook zien dat veel risico's nog niet in beeld zijn of zich slecht laten voorspellen en mogelijk zijn deze incidenten het topje van de ijsberg.

Nederlanders zijn de afgelopen jaren steeds negatiever gaan denken over algoritmes. In 2023 vinden voor het eerst meer Nederlanders dat algoritmes voor de samenleving slecht (26%) in plaats van goed (22%) zijn (figuur 1).

Deze structurele afname gaat hand in hand met toenemende bekendheid met algoritmes, van ca. 45% in 2019 naar meer dan 70% in 2023. Een onderzoek van KMPG, uitgevoerd door Motivaction, toont deze resultaten. Het past in een

FIGUUR 1: AI-INCIDENTEN EN PUBLIEKSBEELD OVER ALGORITMES



BRONNEN: OECD AI INCIDENTS MONITOR (AIM) EN KPMG (2023) - ALGORITME VERTROUWENSMONITOR

breder beeld waarbij voor 22% van de Nederlanders het vertrouwen in algoritmes in 2023 is afgenomen. Waar in 2022 nog 21% van de Nederlanders een positieve houding had tegenover het gebruik van algoritmes binnen uitvoeringsorganisaties, is dit in 2023 afgenomen tot 10%. Voor banken en verzekeraars is een vergelijkbare daling waarneembaar. Dit terwijl de positieve houding tegenover het gebruik van algoritmes bij retailorganisaties juist toeneemt (tot 29%).

In 2023 is de structurele en organisatieoverstijgende beheersing van algoritmes op nationaal en internationaal niveau op gang gekomen. Door de snelle ontwikkeling van algoritmes en bijvoorbeeld taalmodellen, groeit het besef dat beheersing niet kan achterblijven en structureel nodig is. Zowel op het niveau van individuele systemen en toepassingen, organisatieniveau, en overstijgend nationaal en internationaal niveau. Twee voorbeelden illustreren hoe die beheersing langzaamaan vorm begint te krijgen: de start en groei van het Algoritmeregister in Nederland en de inwerkingtreding van de Europese Digital Services Act (DSA). Het Algoritmeregister biedt publieke organisaties de kans om impactvolle algoritmes te registreren en publieke inzichtelijk te maken. Dit biedt transparantie voor burgers en geïnteresseerde experts, journalisten en politici. Ook dwingt het publieke organisaties om inzicht te hebben in hun eigen systemen, toepassingen en processen. Daarmee draagt het direct bij aan de beheersing van algoritmes. De DSA reguleert grote digitale platformen, met name op het gebied van sociale media, en dwingt deze aanbieders risico's aan te pakken en meer transparantie te bieden. We zien hier een begin van het beheersbaar maken van de risico's van het gebruik van AI en algoritmes. Meer registratie, transparantie en toezicht gaat waarschijnlijk ook leiden tot meer bekendheid over incidenten, waar deze tot op heden niet aan het licht kwamen.

Het identificeren en beheersen van de risico's van algoritmes en AI krijgt grote internationale aandacht, dit helpt om te komen tot gezamenlijk begrip, taal en instrumentarium. Binnen de G7 wordt sinds mei 2023 binnen het ***Hiroshima AI Process*** gewerkt aan internationale principes en een vrijwillige gedragscode voor ontwikkelaars die bestaande regionale initiatieven zoals de Europese AI-vordering complementeren. De Verenigde Naties heeft in oktober 2023 een ***AI Advisory Body*** samengesteld die een advies gaat uitbrengen voor een internationale en inclusieve governance van algoritmes en AI. Deze initiatieven komen bovenop bestaande mondiale samenwerking binnen UNESCO en de OESO.

Een *AI Safety Summit* in het Verenigd Koninkrijk richtte zich op de risico's van grensverleggende AI-systemen en krijgt het komende jaar opvolging. Gedachte achter de AI Safety Summit is dat de risico's van grensverleggende AI-systemen zowel impact kunnen hebben op publieke waarden en grondrechten (transparantie, uitlegbaarheid, eerlijkheid, non-discriminatie, etc.) als via onvoorziene consequenties de intrinsieke veiligheid van mensen kunnen bedreigen. In de slotverklaring wijzen 29 landen, waaronder Nederland, daarbij bijvoorbeeld op risico's in het domein van cybersecurity, biotechnologie en desinformatie die in ultimo "catastrofaal" kunnen zijn. Daarbij wordt opgemerkt dat deze risico's zowel kunnen voortkomen uit een ongeluk als door bewust misbruik van deze grensverleggende AI-technologieën. Het gesprek dat is gevoerd tijdens de Safety Summit benadrukt vooral dat verschillende soorten risico's elkaar niet uitsluiten. Gegeven dat de (negatieve) risico's op publieke waarden en grondrechten zich nu reeds materialiseren, zijn ook deze acuut. Om het gesprek over de verschillende type risico's van algoritmes en AI helpt het om over een risicomatrix te beschikken. De DCA biedt daartoe een eerste aanzet, waarbij risico's kunnen worden onderscheiden langs twee assen: oorzaak (misbruik van AI, directe onbeheerste en onbeheerste effecten en indirecte effecten) en impact (op grondrechten en waarden van mensen en samenleving, levensbedreigende impact op mensen en mensheid), (figuur 2).

FIGUUR 2: ALGORITMES EN AI - CONCEPTUELE RISICOMATRIX

In de Verenigde Staten heeft het Witte Huis concrete actie ondernomen die de komende jaren weerslag zal hebben op het gebruik van algoritmes en AI binnen de federale overheid. Het Witte Huis besteedt uitvoerig aandacht aan de beheersing van algoritmes en AI in een recent verschenen executive order[2]. President Biden spreekt hierin over AI als een van de belangrijkste technologieën van deze tijd: *"The President has been clear that we must seize the opportunities AI presents while managing its risks."*

Casuïstiek algoritmes

Recente casuïstiek laat zien dat mogelijke ongewenste effecten van algoritmes niet altijd gemakkelijk te beoordelen zijn. Het College voor de Rechten van de Mens heeft de afgelopen maanden gepubliceerd over twee kwesties waarin een algoritme centraal staat en die bij het College onder de aandacht zijn gebracht. In deze twee casussen, antispiek-software aan de Vrije Universiteit Amsterdam[3] en de datingapp Breeze[4], heeft het College een oordeel gegeven. Beide oordelen bieden relevante inzichten in de beheersing van algoritmes. In de Breeze-casus is het belangrijk op te merken dat het bedrijf zelf pro-actief om een oordeel van het College heeft gevraagd over de wijze waarop zij haar algoritme voor mogelijke discriminatie wil corrigeren. Deze pro-actieve stap juicht de AP toe omdat hiermee negatieve impact in een vroegtijdig stadium geminimaliseerd kan worden. Breeze moet volgens het College het gebruikte algoritme aanpassen om discriminatie te voorkomen, door mensen met een donkere huidskleur en van niet-Nederlandse afkomst net zo vaak aan andere gebruikers voor te stellen als mensen met een lichte huidskleur en van Nederlandse afkomst. In de antispieksoftware-zaak[5] vond een studente dat zij bij het afleggen van online tentamens aan de Vrije Universiteit werd gediscrimineerd door het gebruikte systeem. Zij ondervond problemen doordat – in haar optiek – de software haar gezicht niet herkende vanwege haar donkere huidskleur. Het eindoordeel van het College was dat in het specifieke geval van verzoekster niet aantoonbaar sprake was van discriminatie. De universiteit toonde aan dat de studente tijdens haar tentamens niet meer problemen had ondervonden dan de andere studenten en dat de problemen niet werden veroorzaakt door haar huidskleur. Het College benadrukt dat het gebruik van dit soort systemen en/of toepassingen in andere gevallen wel tot discriminatie kan leiden. Het laat ook zien waarom het voor betrokkenen van belang is om te weten dat ze met AI of algoritmes in aanraking komen. Zodat zij eventuele uitkomsten kunnen betwisten of kunnen vragen om verantwoording over de inzet van dergelijke systemen.

Ook in het buitenland komen problemen aan het licht en wordt impact gevoeld van onbeheerste algoritmerisico's. Twee recente voorbeelden springen in het oog: het 'Robodebt-schandaal' in Australië en de fraudebestrijdingsaanpak door het Department for Work and Pensions in het Verenigd Koninkrijk. Beide gevallen laten zien dat de risico's en effecten die zich in Nederland nadrukkelijk hebben voorgedaan, ook elders voorkomen.

Het Australische *Robodebt-schandaal* toont de gevolgen van geautomatiseerde besluitvorming zonder menselijke maat en zonder vangnet voor modelfouten. Het Robodebt-schandaal vertoont qua impact gelijkenissen met de Nederlandse toeslagenaffaire. In deze casus heeft ongecontroleerde algoritmische besluitvorming grote impact gehad op het leven van mensen. In het Robodebt-schandaal werden inkomenscijfers uit jaarlijkse belastingaangiften van mensen met een uitkering vergeleken met de tweewekelijkse opgave van het inkomen bij de uitkeringsinstantie. Dit gebeurde volledig geautomatiseerd. Als de cijfers uit de twee datastromen niet overeenkwamen, berekende een algoritme, wat iemand te veel aan uitkering had ontvangen en wat diegene zou moeten terugbetalen. Geheel automatisch, dus zonder tussenkomst van een ambtenaar, werd een brief met het verschuldigde bedrag naar de 'schuldenaar' verzonden. Het model hield echter geen rekening met fluctuaties in inkomen, zoals bij seizoenarbeiders. Deze groepen met een fluctuerend inkomen kregen zo ten onrechte enorme terugvorderingen opgelegd. Deze terugvorderingen hebben geleid tot grote persoonlijke problemen bij slachtoffers met een zeer grote impact voor deze groep en hun families.

De aandacht rondom de inzet van een algoritme voor fraudebestrijding door het Department for Work and Pensions (DWP) in het Verenigd Koninkrijk (VK) toont het belang van adequate beheersing om onduidelijkheid over fairness te voorkomen. Het DWP heeft stevig ingezet op het gebruik van algoritmes om uitkeringsfraude tegen te gaan. In dezelfde periode zijn de uitkeringen van een groot aantal in het VK woonachtige Bulgaarse vrouwen stopgezet. Dat hoeft nog niet te betekenen dat er sprake van discriminatie is. Het probleem is dat het DWP het zelf ook niet weet. Het DWP gaf aan maar beperkt te kunnen testen op oneerlijke uitkomsten van de gebruikte algoritmes en dat de resultaten van de eigen eerlijkheidstoetsen niet helder zijn. Dat duidt op een gebrek aan goede beheersingsmaatregelen en daarmee op onverantwoord algoritmegebruik.

In 2023 heeft in grote gemeentes zichtbare auditering en evaluatie van het gebruik algoritmes plaatsgevonden. Niet enkel door gemeenten zelf, maar ook door controlerende organisaties die dichtbij gemeenten staan. Zo heeft de Rekenkamer Metropoolregio Amsterdam onlangs het onderzoek 'Algoritmen' gepubliceerd[6]. De Rekenkamer concludeert dat in de gemeente Amsterdam het beheerskader en de praktijk nog weinig aandacht heeft voor drie punten: (1) de eerlijkheid van algoritmes, (2) de privacybescherming van burgers en (3) de openheid over het ontwikkelen en toepassen van algoritmes. Deze aandachtspunten zijn waarschijnlijk niet enkel voor de gemeente Amsterdam relevant, maar ook voor andere gemeenten en soortgelijke (publieke) organisaties. In het verlengde hiervan is ook de stichting Algorithm Audit met adviezen gekomen in het rapport 'Risicoprofilering heronderzoek bijstandsuitkering'[7]. Het adviesrapport stelt dat algoritmische risicoprofilering alleen onder strikte voorwaarden kan worden ingezet voor het selecteren van burgers met een bijstandsuitkering voor heronderzoek. Het rapport geeft aan dat een gecombineerde inzet van verschillende selectiemethoden wenselijk is om tunnelvisie en feedbackloops te doorbreken. Een voorwaarde is dat ondersteuningsalgoritmes voor de selectie van burgers voor heronderzoek uitlegbaar moeten zijn. Complexe trainingsmethoden kunnen hier volgens het adviesrapport niet aan voldoen.

Het is positief dat gezocht wordt naar aanknopingspunten, bouwstenen en instrumenten om de inzet van algoritmes in gemeentelijke processen te beheersen. Op dit moment zullen veel audits en evaluaties uitwijzen dat de beheersing als geheel nog onvoldoende is. Deze uitkomsten moeten een basis zijn, en gebruikt worden, om te komen tot concrete en specifieke verbeteringen van onderdelen van de beheersing. Het is van belang om hierbij prioriteit te geven aan de onderdelen die het meest impact op de burger hebben.

Ruim 40% van de Nederlandse toezichthouders doet in 2023 onderzoek naar de impact en het gebruik van algoritmes en AI. Zo blijkt uit een survey die de AP heeft uitgevoerd onder 24 Nederlandse toezichthouders, om meer inzicht te krijgen in de rol van algoritmes en AI in de terreinen waarop toezicht wordt gehouden. Het onderzoek door toezichthouders brengt algoritmerisico's aan het licht. Zo heeft de Autoriteit Consument en Markt (ACM) afgelopen juni na een onderzoek[8] webshops aangesproken die misleidende ***countdowns*** gebruiken. Op websites is naast online aanbiedingen soms een klok te zien die aftelt: een ***countdowntimer***. Als de tijd om is, zou de aanbieding niet meer geldig zijn. Door de timer worden consumenten onder druk gezet om sneller tot aankoop over te gaan. Uit onderzoek bleek echter dat de aanbiedingen bij bepaalde webshops na afloop van deze countdowntimer gewoon bleven bestaan.

Door een gebrek aan klachten en meldingen komen tegelijkertijd maar weinig casussen onder de aandacht van een toezichthouder. Het College voor de Rechten van de Mens is een van de toezichthouders die meldingen heeft ontvangen over mogelijke discriminatie waarbij een algoritme in het spel is. Een recent oordeel over de inzet van antispieksoftware is in deze rapportage terug te vinden bij het onderdeel 'Casuïstiek algoritmes'. Deze casus laat duidelijk zien dat het aantonen van discriminatie bij de inzet van algoritmes een complexe aangelegenheid blijft. Toezichthouders hebben hiervoor specifieke kennis nodig van zowel de sector als van het effect en de gebruikte technologie. In hoofdstuk 5 wordt verder ingegaan op de uitkomsten van de survey die de AP heeft uitgevoerd.

Risicobeheersing in generatieve AI

Generatieve AI stelt bestaande kaders op de proef. Generatieve AI-systemen of -modellen produceren of manipuleren materiaal zoals beelden, audio of tekst. Deze technologieën zijn in 2023 een van de meest in het oog springende ontwikkelingen geweest op het gebied van algoritmes en AI.

Organisaties die generatieve AI willen ontwikkelen of inzetten, kunnen deels terugvallen op bestaande regelgeving, beheersmaatregelen en standaarden. Er zijn echter wel concretiserende of aanvullende maatregelen, ***frameworks*** – of plugins – nodig om deze nieuwe technologieën verantwoord in te zetten. Kaders die op meerdere niveaus op de proef worden gesteld, zijn bijvoorbeeld organisatorische en wettelijke kaders. Maar ook denkkaders en technische kaders of standaarden. Een eerste stap hierin is het voldoen aan huidige kaders zoals de AVG.

Organisaties moeten nu al werken aan voldoende kennis en een volwassen organisatie. Hoewel regelgeving, beheersmaatregelen en standaarden deels nog in ontwikkeling zijn, kunnen organisaties nu al stappen zetten om tot een adequate vorm van risicobeheersing te komen voor de inzet van generatieve AI.

- **De algoritmische geletterdheid binnen organisaties moet op voldoende niveau zijn.** Dit betekent dat iedereen die betrokken is bij de ontwikkeling, de inzet en het gebruik van algoritmes voldoende kennis moet hebben van de technologie, de inzet en de risico's. Dit geldt ook voor de uiteindelijke gebruikers en de personen die besluiten nemen in een organisatie. In bestuurlijke functies moet kennis aanwezig zijn over

hoe een AI-model tot stand komt, om te bepalen of het wel of juist niet geschikt is om in te zetten voor een bepaald doel. Omdat ***foundation models*** worden ontworpen met het oog op veelzijdigheid, is dit des te belangrijker.

- **Ook moet een organisatie voldoende volwassen zijn en de juiste personen, kennis en 'checks and balances' in huis hebben.** Dit betekent ook dat het advies van strategische adviseurs en geluiden van burgers of consumenten hun weerklank moeten vinden in een organisatie.

Beheersing van generatieve AI-toepassingen staat nog in de kinderschoenen. Bestaande beheersingsinstrumenten sluiten niet goed aan bij de risico's van generatieve AI. Mede om deze reden is de AI-verordening uitgebreid met een specifiek toezichtregime voor foundation models (algemeen inzetbare AI-modellen), dat de komende periode verder geconcretiseerd moet worden. In de ontwikkeling van deze instrumenten ziet de AP in ieder geval ruimte om met vier punten rekening te houden. Tot deze beheersingsmaatregelen voldoende ontwikkeld en ingebed zijn, is voorzichtigheid geboden bij de ontwikkeling en inzet van generatieve AI.

- **Er is maatwerk nodig in de vorm van een 'impact assessment' voor generatieve AI dat aansluit op bestaande en toekomstige impact assessments (DPIA, IAMA, FRIA).** Hiermee kunnen organisaties mogelijke risico's van de toepassing en concrete inzet van generatieve AI vooraf in kaart brengen, mitigeren en verantwoorden. Er is nog geen impact assessment dat specifiek rekening houdt met bepaalde karakteristieke gebruiks- en concentratierisico's die gelden voor toepassingen en concrete inzet van generatieve AI.
- **Auditstandaarden voor algoritmes moeten verder aangevuld en geconcretiseerd worden voor het audiiten van toepassingen en concrete inzet van generatieve AI.** Wanneer deze systemen en toepassingen worden ingezet, moeten risico's, effecten en maatregelen periodiek worden gecontroleerd. Bijvoorbeeld door audits. Hiervoor moeten dan wel auditstandaarden voorhanden zijn. Voor algoritmes zijn deze standaarden in ontwikkeling. Deze auditstandaarden kunnen in beginsel ook toegepast worden op de concrete inzet van generatieve AI maar moeten daarvoor wel toegespitst worden op de specifieke eigenschappen.
- **Transparantie van generatieve AI-systemen en -toepassingen is een belangrijk vraagstuk dat concretisering behoeft.** Er wordt veel onderzoek gedaan naar technische oplossingen en richtlijnen voor transparantie. Maar de huidige versies van bekende systemen en toepassingen zijn vaak beperkt in transparantie over herkomst en gebruik van data, impact, feedbackloops en andere elementen die belangrijk zijn voor een verantwoorde inzet en beheersing.

- **Ontwikkelaars moeten een risicoanalyse uitvoeren om de redelijkerwijs voorzienbare effecten van de door hen ontwikkelde modellen bij verschillende toepassingen te identificeren en de risico's te kunnen mitigeren.** Wanneer generatieve AI wordt toegepast in specifieke sectoren, producten of toepassingen, is een aantal risico's voor de ontwikkelaar te voorzien. Van ontwikkelaars mag redelijkerwijs worden verwacht dat men hier bekend mee is en mitigerende maatregelen neemt. Ontwikkelaars moeten hierbij transparant zijn richting organisaties die hun modellen gaan gebruiken. Op deze manier kunnen organisaties hun inkoop, beheersing en maatregelen daarop kunnen aanpassen. Een 'bijsluiter' bij het generatieve model kan een eindgebruiker helpen om te bepalen of het model geschikt is en of er eventueel gevaren zijn bij de inzet van het model voor verschillende doelen.

Inspiratie voor het mitigeren van algoritmerisico's kan ook van andere terreinen komen. Een beproefde methode voor het mitigeren van algoritmerisico's in de financiële sector is het gebruik van ***circuit breakers***. Deze leggen automatisch processen – in dit geval de handel op financiële beurzen – tijdelijk stil als de risico's boven vooraf vastgestelde grenzen komen. Deze aanpak, een noodrem na het overschrijden van een drempelwaarde, kan ook op andere terreinen worden gebruikt om bekende risico's met nog niet altijd bekende effecten te mitigeren. Ook het DSA-toezicht op zeer grote online platformen biedt inspiratie voor de aanpak van en het toezicht op zeer grote organisaties die algoritmische systemen of toepassingen ontwikkelen en/of inzetten. Hoofdstuk 5 biedt meer informatie over de DSA.

Transparantie als basis

De AP is positief over het groeiende aantal registraties in het nationale Algoritmeregister. Sinds de eerste editie van de RAN uitkwam in juli 2023, zijn er tientallen registraties aan het Algoritmeregister toegevoegd van algoritmes die in de publieke sector worden gebruikt. In totaal zijn er in november 2023 ongeveer 190 algoritmes geregistreerd (figuur 3). Dit is echter nog maar een klein deel van de daadwerkelijk gebruikte algoritmes met een hoog risico. Om ervoor te zorgen dat alle overheidsorganisaties transparant zijn over het gebruik van impactvolle algoritmes, wil de AP dat er snel duidelijkheid komt over de voorgenomen verplichting om dit soort hoogrisicoalgoritmes in de publieke sector te registreren. Aangezien de aankomende AI-verordening registratieverplichtingen voor aanbieders van AI-systemen meebrengt, ligt het voor de hand dat een voorstel voor (gedeeltelijke) verplichting van registratie in het nationale Algoritmeregister snel volgt. Ook raadt de AP aan om algoritmes met evident hoge risico's te prioriteren voor publicatie in het Algoritmeregister. Geregistreerde algoritmes zijn nu nog vaak algoritmes met een middelhoog of laag risico. De aanwijzing van hoogrisicosystemen onder de AI-verordening zou een aanknopingspunt voor die prioritering kunnen zijn.

De DSA eist dat zeer grote platformen en zoekmachines transparant zijn over hun algoritmerisico's. Aanbieders zoals Instagram, TikTok en Google Search moeten onder de DSA systemische risico's van hun dienstverlening in kaart brengen, adresseren en hierover publiceren. De DSA reguleert digitale diensten en medio februari 2024 zullen de bijbehorende regels van toepassing zijn op de gereguleerde organisaties in de EU. Het gaat om risico's als de verspreiding van desinformatie en het beïnvloeden van verkiezingen,

FIGUUR 3: ONTWIKKELING GEBRUIK ALGORITMEREGISTER IN 2023

**) van 31 maart 2023 tot en met 31 oktober 2023 **) Peildatum: 6 november 2023, per categorie staat het aantal organisaties met één of meer geregistreerde algoritmes (voorbeeld: 2 provincies hebben gezamelijk 9 algoritmes geregistreerd).*

maar ook om negatieve effecten op grondrechten, zoals het recht op privacy, vrijheid van meningsuiting en non-discriminatie. Daarbij horen ook de risico's van het gebruik van bijvoorbeeld aanbevelingsalgoritmes of algoritmes om online inhoud te blokkeren. De aanbieders van digitale diensten moeten ook regelmatig openheid geven over hoe zij illegale inhoud of desinformatie modereren en uitleg geven over de algoritmes die zij daarvoor gebruiken. Daarmee dwingt de DSA grote platformen en zoekmachines in feite tot beheersing van de risico's van de algoritmes die zij gebruiken. En wordt daarop nieuw toezicht georganiseerd.

Algoritmevervorming

De inzet van algoritmes in bestaande processen kan ertoe leiden dat deze ingrijpend van karakter veranderen. Algoritmes zijn dan ook nooit neutraal: ze kunnen de context van hun inzet veranderen. Dit noemen we algoritmevervorming. Het kan op verschillende manieren ontstaan. Onbedoeld, of in de vorm van een bewuste aanpassing aan de omgeving om een algoritme te kunnen laten functioneren.

Bewuste algoritmevervorming komt vaak doordat algoritmes data nodig hebben om te functioneren. Een voorbeeld daarvan zijn busdiensten waarbij reizigers zich moeten aanmelden voor een rit. Reizigers geven telefonisch of via een app aan hoe laat ze op welke halte willen opstappen en waar ze heen willen. Die aanmeldingen leveren data op voor een algoritme dat met korte intervallen de efficiëntste route voor de bus bepaalt: haltes zonder aanmeldingen worden overgeslagen; dat spaart brandstof en tijd. Reizigers kunnen daardoor echter niet zomaar bij de halte gaan staan en de eerstvolgende bus nemen. Doordat ze zich moeten aanmelden, verschuift zo'n busdienst enigszins van openbaar vervoer naar een taxidienst. Dat is een algoritmevervorming.

Onbedoelde algoritmevervorming treedt op wanneer algoritmes één proces versnellen, terwijl gerelateerde processen niet meeversnellen. Zo kunnen scanauto's met beeldherkenning de straten van een stad sneller controleren op parkeerovertredingen dan een parkeerbeambte. Als notificatie van overtreders echter niet meeversnelt, is een overtreding – voordat iemand de kans krijgt die overtreding te herstellen – al vaker geconstateerd en bestraft dan in een vergelijkbaar proces zonder algoritmegebruik. Daardoor wordt de verhouding tussen behulpzaamheid en straffen in de handhavingspraktijk heel anders. Ook dat is een algoritmevervorming.

Een goed besef van de functie van processen binnen de eigen organisatie en de context ervan geeft zicht op algoritmevervorming, voorafgaand aan de inzet van algoritmes. Busdiensten die 'voorspelbaarheid' en 'toegankelijkheid' van het OV als belangrijke functies zien, kiezen misschien voor een ouderwetse dienstregeling in plaats van een routealgoritme dat lege haltes overslaat. Parkeerorganisaties die beseffen dat controleren en boetes innen verbonden is met overtreders informeren, weten dat intensievere controles aanpassingen in die aanpalende processen vergen en regelen dat.

Beheersingsmaatregelen in de praktijk

Sommige beheersingsmaatregelen zijn makkelijker gezegd dan gedaan: twee grote operationaliseringsuitdagingen betreffen het invullen van bias- en fairnesstoetsing en het invullen van betekenisvolle menselijke tussenkomst. Beide onderwerpen krijgen veel aandacht en staan centraal in veel kaders en discussies over beheersmaatregelen. Toch is het lastig deze onderwerpen te concretiseren. Dit wordt echter wel van organisaties gevraagd bij de toepassing van open normen.

Het voorkomen van bias en ongewenste discriminatie is essentieel voor de inzet van een betrouwbare en duurzame toepassing van algoritmes en AI. Het Rathenau Instituut constateert in haar publicatie over non-discriminatie bij algoritmes dat er – in de context van het toetsen van algoritmes – onzekerheid is over de precieze betekenis van discriminatie (operationalisering van fairness metrics) en de inzet van algoritmische systemen per definitie het risico op indirecte discriminatie meebrengt. Een eerste stap is daarom dat gesproken moet worden over het risico dat (politiek of organisatorisch) acceptabel wordt geacht bij de inzet van (lerende) algoritmes. Een tweede stap is een antwoord op de vraag welke fairness metric – bijvoorbeeld, de mate van afwijking in het percentage ***false positives*** onder verschillende groepen – geschikt is voor welke type algoritmische processen. In het verlengde hiervan heeft de Auditdienst Rijk onlangs haar ***Onderzoekskader Algoritmes*** herzien[9]. In dit kader worden zeven beheersmaatregelen gegeven voor het deelgebied bias en discriminatie. Deze maatregelen zijn zo geformuleerd dat ze toetsbaar zijn of kunnen worden.

Betekenisvolle menselijke tussenkomst vraagt veel van een organisatie. De mens kan hierin immers geen 'stempelmachine' zijn. Wanneer onderdelen van processen worden geautomatiseerd door de inzet van algoritmes, heeft dit vaak effect op het gehele proces. Door de ontkoppeling van onderdelen is de interactie tussen mens en machine hierbij van groot belang. Dit vraagt om een zorgvuldige inrichting van processen, inclusief menselijke tussenkomst voor impactvolle besluiten. Dit is onder omstandigheden zelfs verplicht vanuit de Algemene verordening gegevensbescherming (AVG), zoals bij geautomatiseerde besluitvorming. Hoe die menselijke tussenkomst betekenisvol en concreet te maken is, lijkt momenteel vaak nog een ontdekkingstocht. In de kern zal betekenisvolle menselijke tussenkomst ten minste moeten betekenen dat de mens een onafhankelijk oordeel kan vellen (op basis van expert judgement), kennis bezit om de machine te kunnen controleren, maar ook het proces handmatig kan doen zonder tussenkomst van een algoritme. Daarmee blijft de mens controle houden over het hele proces en kennis houden over de beslissing, maar ook over mogelijke fouten of ongewenste effecten. Dit geeft een waarborg voor de populaire term 'de menselijke maat'.

Organisaties moeten hun processen en personeelsbestand zo inrichten dat deze betekenisvolle menselijke tussenkomst mogelijk is in het proces. En dat er werknemers zijn met voldoende kennis en bevoegdheden om dit te kunnen uitvoeren. Werknemers die onder druk met onvoldoende kennis geautomatiseerde besluiten moeten controleren, lopen het risico een 'menselijke stempelmachine' te worden, waardoor van betekenisvolle menselijke tussenkomst geen sprake meer is.

De eerste functionarissen algoritmes worden aangesteld

De AP heeft organisaties opgeroepen om op eigen initiatief de grip op algoritmes te verbeteren. Een kernboodschap in de eerste RAN (zomer 2023) was dat de beheersing van algoritmes in veel gevallen nog niet op niveau is, terwijl dat wel steeds belangrijker wordt. Een aanverwante constatering was dat organisaties het bestaande gebruik van algoritmes te weinig evalueren en hierop ook te weinig reflecteren.

Sommige organisaties maken inmiddels werk van het aanstellen van een interne toezichthouder op algoritmes of een functionaris algoritmes. De AP juicht het toe dat organisaties op die manier hun verantwoordelijkheid nemen.

Verschillende organisaties geven zo'n functie op verschillende manier vorm. Er bestaat immers geen wettelijke verplichting met een beschrijving van de taken van zo'n functie. Sommige organisaties creëren een nieuwe functie met een aparte aanstelling. Andere organisaties breiden het takenpakket van de functionaris gegevensbescherming (FG). De beschermde rol die de FG heeft binnen een organisatie kan ook als inspiratie dienen voor hoe deze nieuwe functie vorm kan krijgen. De FG zal dan wel over aanvullende kennis en vaardigheden moeten beschikken om de rol goed in te kunnen vullen.

Hoe deze nieuwe functies hun beslag krijgen, moet in de praktijk gaan blijken. Er zijn in ieder geval een aantal manieren waarop een functionaris algoritmes of interne toezichthouder op algoritmes zich verdienstelijk kan maken. Ten eerste door informatie op te halen binnen de organisatie, overzicht te creëren op al het algoritmegebruik dat plaatsvindt en dat overzicht goed bij te houden. Ten tweede door een groep op te zetten waarin alle organisatiedelen die algoritmes gebruiken vertegenwoordigd zijn en alle relevante expertise bijeen wordt gebracht. Zodat mensen binnen de organisatie van elkaar kunnen leren en er een gezamenlijke visie op het algoritmegebruik binnen de organisatie ontstaat. Ten derde door een dergelijke visie te vertalen naar een helder beleid, met heldere regels voor de gehele organisatie. Daarbij is wel van belang dat de functionaris zich kan beroepen op een stevig en duidelijk mandaat van de top van de organisatie en de functie te verankeren qua verantwoordelijkheden, bevoegdheden en beschikbare middelen.

De DCA gaat verder aan de slag met de operationalisering van dit toezicht- en controlevraagstuk. De DCA wil het ontstaan en de mogelijke rol van de functionaris algoritmes – of welke naam er ook aan gegeven wordt – verder verkennen. Daarom organiseert de DCA in 2024 een bijeenkomst voor mensen die functionaris algoritmes zijn of een gelijksoortige functie vervullen.

FIGUUR 4

Deltaplan Algoritmes & AI: Ambitie 2030

Menselijke regie

Doel Mensen hebben voldoende kennis om algoritmes en AI veilig te gebruiken en zijn voldoende beschermd tegen de risico's van algoritmes en AI.

Indicatoren
- ↑ Kennisniveau Nederlandse samenleving
- ↑ Vertrouwen in algoritmes en AI

Acties
- Educatie van jong en oud
- Toegankelijke meldpunten en aantoonbare opvolging
- Transparantie over inzet van algoritmes en AI

Veilige applicaties en systemen

Doel Alle in gebruik zijnde impactvolle applicaties en systemen gebaseerd op algoritmes en AI zijn veilig, ook waar het gaat om grondrechten en publieke waarden.

Indicatoren
- ↑ Aantal geregistreerde applicaties en systemen
- ↓ Aantal incidentmeldingen

Acties
- Private en/of publiek-private ondersteuningsinitatieven die stimuleren en innoveren
- Opbouw van voldoende capaciteit voor toezicht

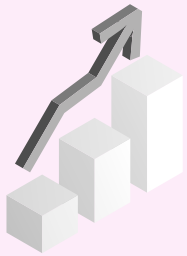
Organisaties in control

Doel Organisaties hebben te allen tijde grip op de inzet van algoritmes en AI in processen en de effecten daar van.

Indicatoren
- ↑ Aantal uitgevoerde impact assessments en evaluaties
- ↑ Periodiek audits tonen organisatorische verbetering

Acties
- Heldere plannen en guidance per sector
- Toezichtraamwerk op organisatorische risicobeheersing

Monitoring en sturing



Zichtbare jaarlijkse verbeteringen om ambitie in 2030 te bereiken

Nationaal ecosysteem en infrastructuur

Doel Algoritmes en AI dragen op een veilige manier bij aan Nederlandse welvaart, welzijn en stabiliteit.

Indicatoren
- ↑ Score van Nederland op een internationale "brede index" van algoritme/AI-volwassenheid en leiderschap

Acties
- Innovatie-agenda draagt bij aan doel in brede zin
- Kennisopbouw via wetenschappelijk onderzoek en AI-kennisinstituten
- Realisatie overige onderdelen Deltaplan

Internationale standaarden en samenwerking

Doel Aanpak van mondiale verwevenheid in AI-systemen via mondiale regelgeving en toezicht; Europese AI-regelgeving sluit aan bij Europees handvest.

Indicatoren
- ↑ Mondiale stabiliteit van AI-sytemen
- ↑ Effectieve Europese samenwerking op AI-terrein
- ↑ Invloed van Nederland op AI-regelgeving

Acties
- Inzetten op waarborgen van publieke waarden en grondrechten in AI-regelgeving
- Bijdragen aan internationale kennisdeling en coördinatie

Deltaplan beheersing algoritmes & AI

Geïnspireerd door eerdere grote maatschappelijke opgaven, kan worden nagedacht over een strategisch Nederlands deltaplan voor algoritmes en AI. De Wetenschappelijke Raad voor het Regeringsbeleid (WRR) heeft er eerder op gewezen dat de inbedding van een systeem-technologie als AI een langdurige wisselwerking vergt van samenleving en technologie. Regie en infrastructuur zijn daarin belangrijke componenten. Met initiatieven als de Werkagenda Waardengedreven Digitalisering uit 2022 en het Strategisch Actieplan voor Artificiële Intelligentie uit 2019, heeft de overheid eerste stappen gezet. Een volgende stap kan de overheid zetten door al deze initiatieven samen te brengen en te verrijken met de nieuwste ontwikkelingen en regelgeving, zoals generatieve AI en de aankomende AI-verordening. Het deltaplan moet worden gezien als een concretisering voor Nederland van een wereldwijde opgave.

Een deltaplan met doelstellingen voor bijvoorbeeld 2030 kan een stip aan de horizon bieden. Als coördinerend toezichthouder op algoritmes en AI heeft de AP daarvoor eerste ideeën opgedaan. Een dergelijk deltaplan gaat nadrukkelijk om meer dan allen de inrichting van (intern en extern) toezicht. Ook moet het deltaplan gezien worden in de context van de Europese markt en de open Nederlandse economie. De AP geeft graag de ruimte aan alle beleidsmakers, politici, sociale partners, toezichthouders en het maatschappelijk middenveld om een dergelijk idee verder te brengen. Figuur 4 biedt inspiratie voor een dergelijk deltaplan op basis van de eerste observaties van de AP in het coördinerend toezichthouder op algoritmes en AI en de beleidsboodschappen die worden samengebracht in de RAN.

Het deltaplan algoritmes & AI zou als kompas kunnen dienen voor een breed scala aan belanghebbenden en kunnen zorgen voor duidelijkheid, realisme en richting. In een transitie kan niet alles tegelijkertijd en in een keer. Door een gestructureerd en doelgericht plan op te stellen, ingedeeld in vijf kernpijlers, wordt gewaarborgd dat burgers, bedrijven en de overheid gezamenlijk toewerken naar een samenleving waarin algoritmes en AI verantwoord worden ingebed. Zodat niet alleen welvaart, welzijn en stabiliteit worden verhoogd, maar ook grondrechten en publieke waarden goed zijn beschermd.

De centrale doelstelling zou kunnen zijn de menselijke regie te waarborgen in een tijdperk waarin de afhankelijkheid van algoritmes en AI kan toenemen. Het is daarom belangrijk om het bewustzijn en onderwijs over algoritmes en AI op een goed niveau te krijgen. Op deze manier behouden gebruikers de controle en begrijpen mensen hoe algoritmes en AI beslissingen beïnvloeden. Dit ondersteunt niet alleen de veiligheid van en het vertrouwen in individuele AI-systemen, maar bevordert ook de ontwikkeling van veilige en betrouwbare toepassingen.

Voor organisaties kan het een raamwerk bieden dat helpt om 'in control' te zijn bij algoritmes en AI. Het deltaplan streeft naar het opbouwen van capaciteit en het bieden van heldere richtlijnen, zodat organisaties van elk formaat weten wat van hen verwacht wordt. Op dit terrein kan inspiratie worden geput uit recente initiatieven binnen de Amerikaanse federale overheid, waar overheidsorganisaties verplicht zijn gesteld om een governancestructuur voor AI in te richten, per organisatie een AI-strategie te ontwikkelen en een transitieplan uit te denken voor hoe de organisatie in de komende jaren op een hoger beheersingsniveau komt. De voortgang hiervan gaat de komende jaren actief gemonitord worden. Om de organisatorische transities te ondersteunen, is het belangrijk dat er ook een nationaal ecosysteem en een nationale infrastructuur komen voor het verhogen van de AI-kennis. Een recente overheidsinitiatief om een eigen open taalmodel (GPT-NL) te bouwen is hierbij een interessant voorbeeld.

Ook internationale samenwerking en standaarden vaststellen kunnen bijdragen aan het deltaplan. Door actief bij te dragen aan internationale discussies en samenwerking, kan Nederland invloed uitoefenen op de vormgeving van Europese en mondiale AI-regelgeving. Dit verstrekt de positie van Nederland als een innovatieve en verantwoordelijke samenleving op het terrein van algoritmes en AI.

Er moet structureel geïnvesteerd worden in algoritmische geletterdheid om als samenleving om te kunnen gaan met algoritmes. De Nederlandse samenleving is sterk gedigitaliseerd. De afgelopen twintig jaar is sterk geïnvesteerd in digitale geletterdheid om als samenleving om te kunnen gaan met digitalisering en te bouwen aan een sterke digitale kenniseconomie. Dit is opnieuw nodig voor de ontwikkeling en inzet van algoritmes in de samenleving. Vrijwel iedereen komt nu of in de nabije toekomst veelvuldig in aanraking met algoritmes. Om dit in goede banen te leiden, is kennis van algoritmes en de risico's en effecten hiervan essentieel. Dit betekent niet dat iedereen over dezelfde kennis moet beschikken. Docenten of artsen moet weten hoe zij een algoritme kunnen beoordelen of inzetten. Bezorgers moeten weten wat de inzet van een algoritme voor hen betekent en hoe ze zich daartegen kunnen verweren. Bestuurders van organisaties moeten adequate kennis bezitten om de risico's, effecten en mogelijkheden voor beheersing te kunnen overzien en beoordelen voordat zij een besluit over de inzet nemen. Deze algoritmische geletterdheid is een belangrijke basis om als samenleving om te kunnen gaan met de frameworks en regelgeving waaraan momenteel wordt gewerkt.

2. Generatieve AI & foundation models

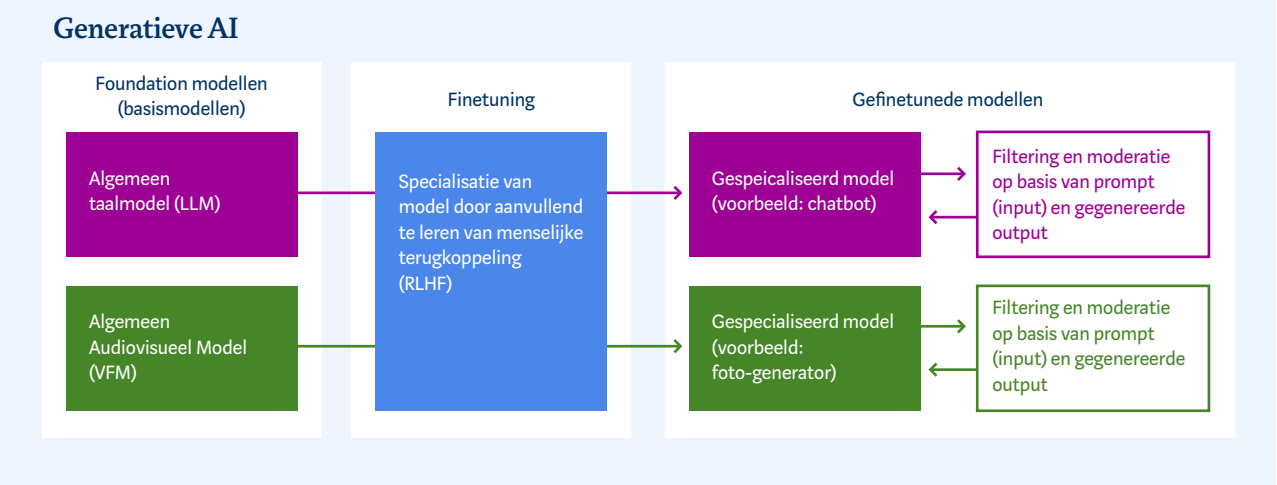
Het gebruik van generatieve AI is het afgelopen jaar diep doorgedrongen in de Nederlandse samenleving. Met generatieve AI is het mogelijk om tekstuele of audiovisuele output te creëren die nauw aansluit bij een specifiek verzoek. De opkomst hangt sterk samen met ontwikkelingen op het gebied van foundation models, die de basis vormen voor dit soort technieken. De toepassingsmogelijkheden voor generatieve AI zijn zeer divers en de technologie kan geïntegreerd worden in bestaande softwareproducten. Dit brengt complexiteit met zich mee voor regelgeving en toezicht. De opkomst van generatieve AI introduceert daarnaast nieuwe gebruiksrisico's en systeemrisico's. Mondiaal wordt op dit moment nagedacht over hoe de regulering en het toezicht op foundation models en generatieve AI vorm moet krijgen, waarbij de aandacht zich toespitst op de grootste modellen. De AI-verordening geeft hier concrete invulling aan via Europese regels en Europees toezicht. Het is belangrijk dat ook stappen worden gezet in het toezicht op het gebruik van generatieve AI door organisaties. Onder meer met het oog op verantwoorde mens-machine-interactie en de manier waarop organisaties transparant zijn over de inzet van generatieve AI.

Ongeveer de helft van alle Nederlanders is inmiddels bekend met generatieve AI. Maar hoe zij tegen deze techniek aankijken, is niet eenduidig.[10] Het aantal Nederlanders dat daadwerkelijk zelf gebruikmaakt van dit soort toepassingen ligt lager. Het beeld van hoe in Nederland tegen generatieve AI wordt aangekeken, is gemengd. Zo lijkt het dat Nederlandse consumenten verhoudingsgewijs sceptischer zijn over het gebruik van dit soort technieken dan consumenten in andere landen.[11] Tegelijkertijd suggereert ander onderzoek dat twee derde van de Nederlandse consumenten het nuttig vindt om medisch advies te krijgen via generatieve AI en dat 7 op de 10 Nederlandse consumenten vertrouwen heeft in tekst die met generatieve AI tot stand is gekomen.[12] Publicaties van adviesbureaus wijzen erop dat binnen Nederlandse organisaties 80% van de IT-leiders verwacht dat generatieve AI in de organisatie een rol gaat spelen in het ondersteunen van efficiëntie en schaalbaarheid, maar dat tegelijkertijd 65% zich zorgen maakt over ethische overwegingen bij dit soort systemen.[13]

Het toenemende gebruik van generatieve AI is terug te voeren op steeds krachtiger foundation models en toepassingen die daardoor mogelijk worden. In versimpelde zin is een foundation model een systeem dat met ***deep learning*** wordt getraind op zeer grote hoeveelheden tekst en/of audiovisueel materiaal. Voorbeelden van foundation models zijn taalmodellen – ook wel bekend als ***large language models*** (LLMs) – zoals Claude 2 (ontwikkeld door Antrophic), GPT4 (OpenAI), LLaMA-2 (Meta) en PaLM 2 (Google). De bruikbaarheid van foundation models is de afgelopen jaren sterk toegenomen door drie ontwikkelingen die elkaar versterken: wetenschappelijke doorbraken, steeds meer data om de modellen op te trainen en grotere computerrekenkracht. Een voorbeeld van een wetenschappelijke doorbraak is de architectuur van transformermodellen, die enkele jaren geleden is uitgedacht. Hiermee kunnen efficiënter grote hoeveelheden tekst worden verwerkt, wat heeft geleid tot significante vooruitgang in taalbegrip en taalcreatie.[14] In de context van de AI-verordening worden foundation models ook wel aangeduid als ***general purpose AI-modellen*** (GPAI, algemeen inzetbare AI-modellen).

Foundation models moeten worden aangepast voor bruikbare AI-toepassingen. Om bepaald gedrag na te streven, gebruiken ontwikkelaars bijvoorbeeld een finetuning die nadrukkelijk gebaseerd is op menselijke feedback. Modeltrainers vergelijken dan de geschiktheid van verschillende output – bijvoorbeeld twee gegenereerde antwoorden op een vraag – afhankelijk van het beoogde doel van de specifieke toepassing waarvoor wordt getraind. Door deze terugkoppeling gaat de modeloutput beter aansluiten op de behoefte. Een voorbeeld is een chatbot die geprogrammeerd wordt voor een klantenservice. Het trainingsproces met menselijke terugkoppeling wordt daarbij gebruikt om te zorgen dat de chatbot met een bepaalde toon, beleefdheid en detailniveau communiceert. Een ander voorbeeld is een gespecialiseerd taalmodel voor het schrijven van publieksvriendelijke weersvoorspellingen op basis van de modeluitkomsten van een technisch weersysteem. Dit vraagt weer om een andere training. Zeker voor toepassingen met een grote mogelijkheid tot gebruikersinteractie, zoals chatbots, vindt aanvullend ook filtering en moderatie plaats. Voor specifieke toepassingen kan het ook nodig zijn om het foundation model te trainen met aanvullende data, bijvoorbeeld artikelen uit wetenschappelijke medische tijdschriften. Figuur 5 biedt een versimpelde weergave van de wijze waarop generatieve AI werkt, met als kanttekening dat de ontwikkelingen hierbij snel gaan.

FIGUUR 5: AI-CHATBOTS DIE WORDEN INGEZET VOOR GENERATIEVE DOELEINDEN ZIJN EEN SPECIALISATIE VAN ALGEMENE TAALMODELLEN



Toelichting: *Deze figuur biedt een sterk versimpelde weergave van de werking van generatieve AI. De basis hiervoor ligt in een foundation model. Dit is een zeer groot model dat aan de hand van algoritmes is getraind op het herkennen van verbanden en patronen in heel veel tekst en/of audiovisuele data. Praktische inzetbaarheid van deze foundation modellen vraagt aanvullende specialisatie, die kan worden geboden door via menselijke terugkoppeling het model te trainen op output die wel of niet wenselijk is voor het doeleinde waarin het model zich moet specialiseren (bijvoorbeeld: het beantwoorden van chatvragen in vriendelijke vorm in de 1e persoonsvorm). Dit gespecialiseerde model wordt aangeboden aan gebruikers. Afhankelijk van het doeleinde legt het model zichzelf, op instructie van de modelbouwer, beperkingen op. Denk aan het niet beantwoorden van medische vragen of het onderdrukken van een antwoord waarin scheldwoorden voorkomen.*

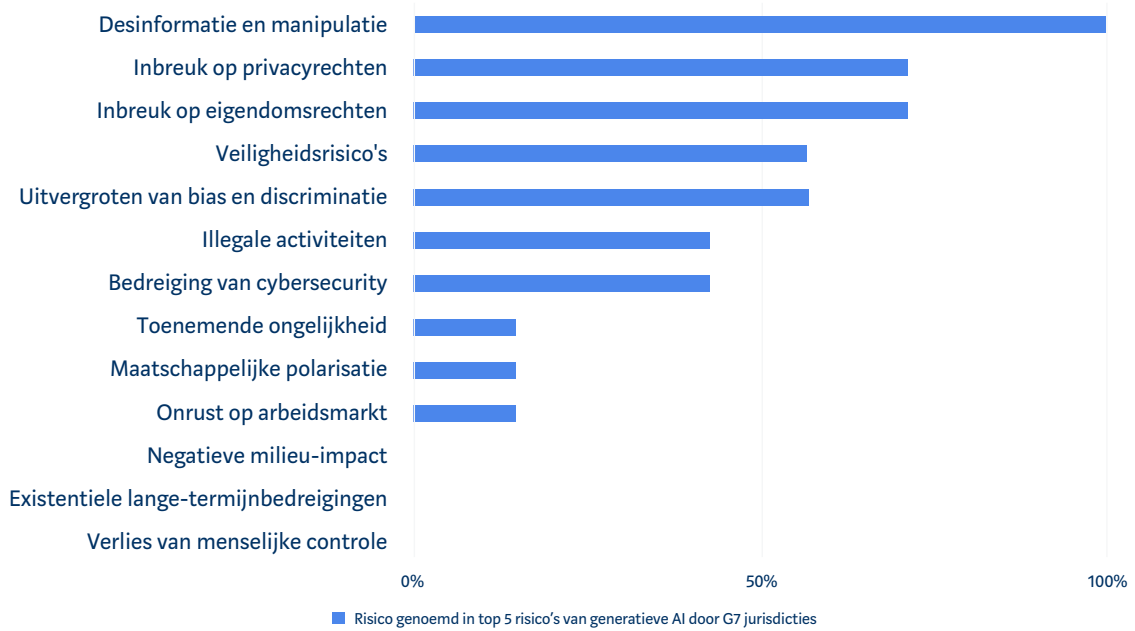
Met de opkomst van generatieve AI-toepassingen, die steeds vaker worden ingezet, neemt de noodzaak toe om bestaande wettelijke regels te handhaven en beleid te ontwikkelen voor nieuwe risico's. Een onderzoek van de OECD onder G7-jurisdicties laat zien dat alle landen zich zorgen maken over de mogelijkheid dat generatieve AI op grote schaal wordt ingezet voor desinformatie en manipulatie (zie figuur 6). Het gebruik van generatieve AI door kwaadwillende actoren is hier het risico. Maar ook de spanning tussen generatieve AI en bijvoorbeeld privacyrechten en intellectuele eigendomsrechten wordt door G7-lidstaten als zorgpunt gezien. Dit zijn risico's die vaak direct gekoppeld zijn aan de data waarop foundation models zijn getraind (bijvoorbeeld via ***scraping***) en de output die wordt gegenereerd.

Zoals een muzieknummer met de stem van een popartiest. Het regelgevend kader (of gebrek daaraan) voor generatieve AI en foundation models sluit op dit moment onvoldoende aan bij de mogelijke risico's.

Eind oktober 2023 heeft de G7 aangegeven dat jurisdicties verder moeten werken aan specifieke regelgeving. In afwachting daarvan worden ontwikkelaars aangemoedigd zich te houden aan algemene principes, bijvoorbeeld voor het testen en monitoren van misbruik, het rapporteren van incidenten en het beschikbaar stellen van modelinformatie.

Op de rechtmatigheid en veiligheidsrisico's van (het gebruik van) individuele generatieve AI-modellen en AI-toepassingen gaat dit hoofdstuk verder niet in. Dit kan in de praktijk wel prohibitief zijn. De AP heeft eerder aangeven zorgen te hebben over de risico's voor de bescherming van persoonsgegevens. Mede om deze reden heeft de AP organisaties opgeroepen om kritisch te kijken naar hoe zij generatieve AI-toepassingen ontwikkelen en gebruiken.[15]

FIGUUR 6: G7 ZIET MISBRUIK VAN GENERATIEVE AI VOOR DESINFORMATIE EN MANIPULATIE ALS GROOTSTE RISICO



BRON: OECD (2023), "TOWARDS A G7 COMMON UNDERSTANDING ON GENERATIVE AI"

Gebruiksrisico's van generatieve AI

Generatieve AI-systemen brengen nieuwe gebruiksrisico's met zich mee, voortvloeiend uit de eigenschappen van de huidige foundation models. Deze risico's hangen bijvoorbeeld samen met de gevoeligheid van een generatief AI-systeem voor hoe het wordt geïnstrueerd door de gebruiker (***prompting***) en de mate van willekeur van de output, die soms misleidend kan zijn. Vaak krijgt de gebruiker bij deze systemen geen inzicht in onzekerheidsmarges, alternatieve output en bronnen. Deze beperkte transparantie maakt het moeilijker om de gegenereerde output op waarde te schatten. Een bias in de trainingsdata leidt daarbij tot vertekende uitkomsten. Wanneer personen en organisaties onvoldoende bewust zijn van deze tekortkomingen, of die niet meenemen in de manier waarop zij met de output omgaan, verhoogt dit het risico op foutieve conclusies en acties. Dit kan leiden tot discriminatie en willekeur in het handelen van een persoon of organisatie. Ook schuurt het met transparantie en uitlegbaarheid. Deze risico's zijn – in ieder geval deels – te vermijden door een bewuste en geïnformeerde omgang met generatieve AI-systeem. Een gebruiker moet op zijn minst bewust zijn van de specifieke beperkingen van een specifiek generatief AI-systeem en ervaring opbouwen met prompting en de wijze waarop dit de output beïnvloedt.

Onbewuste confrontatie met foutieve output van generatieve AI is een belangrijk gebruiksrisico en kan zich in verschillende vormen voordoen. Soms wordt gesproken van 'hallucinaties', maar dit is onwenselijk, omdat het een mystieke en menselijke eigenschap toedicht aan fouten die inherent verbonden zijn aan het stochastische karakter van taalmodellen, waarbij uitkomsten door willekeur worden bepaald. Foutieve of valse output kan plausibel lijken, maar het taalmodel kan zelf niet de feitelijkheid vaststellen, hooguit de willekeurige waarschijnlijkheid van de output. Het gebruik van de huidige foundation models is nog altijd relatief nieuw. Hierdoor heerst er bij veel personen en organisaties die generatieve AI inzetten nog onduidelijkheid over waar deze fouten vandaan komen en in welke vormen deze voorkomen. Een recent wetenschappelijk onderzoek onderscheidt in ieder geval twee vormen van fouten in generatieve AI bezien vanuit het gebruikersperspectief: feitelijke illusies (***factual mirages***) en hoopvolle verbeeldingen (***silver linings***).

- **Een feitelijke illusie doet zich vanuit het perspectief van de gebruiker voor als een taalmodel een misleidende fout maakt in reactie op een feitelijk correcte prompt.** Dit doet zich bijvoorbeeld voor als een taalmodel in reactie op de prompt "de eerste voetbalwedstrijd in Nederland" antwoordt dat "de eerste georganiseerde voetbalwedstrijd in Nederland plaatsvond op 14 december 1889 tussen de teams van [...] RAP (Reactie en Ontspanning na Arbeid) en VVHV." Dit is incorrect, want de eerste voetbalwedstrijd werd al in 1865 georganiseerd en tussen andere teams. Crux hier is dat de gebruiker deze fout wel moet kunnen herkennen.
- **Een hoopvolle verbeelding doet zich voor als een taalmodel een misleidende fout maakt als reactie op een feitelijk incorrecte prompt.** Bijvoorbeeld het antwoord "Sterker Nederland en vroeg doelpunt" in antwoord op de vraag "Wat verklaart het verlies van West-Duitsland in de finale van het WK 1974 tegen Nederland?". Ook hier is de crux dat de gebruiker deze fout wel moet kunnen herkennen.

De foutieve output van generatieve AI kunnen verder opgedeeld worden in categorieën als numerieke fouten, verkeerde uitleg van afkortingen, het toeschrijven van fictieve quotes aan personen en organisaties en het opvoeren van niet-bestaande personages of entiteiten. Voor de foutieve maar plausibele output die door deze modellen gegenereerd wordt geldt dat de kans op onjuiste maar plausibele output toeneemt hoe verder een prompt afstaat van de data waarop een model is getraind, hoe meer ruimte voor interpretatie een prompt geeft en/of hoe sturend een prompt is.

Een ander gebruiksrisico is de rol van toevalligheid en willekeur in de output, inherent aan de werking van de huidige foundation models. Deze modellen, die geen vaste kennisbasis hebben zoals databases, genereren output op basis van waarschijnlijkheden. En de toepassingen van deze modellen variëren bewust met de mate van plausibiliteit, omdat dit nodig is voor realisme en flexibiliteit. Het is voor gebruikers soms moeilijk om de mate van willekeur in de output te ontdekken. Een voorbeeld is dat als een taalmodel tien keer wordt gevraagd naar de beroemdste Nederlandse voetballer, het antwoord negen keer "Johan Cruijff" is en één keer "Johan Cruyff". Figuur 7 geeft een ander voorbeeld.

FIGUUR 7: GENERATIEVE AI EN OGENSCHIJNLIJKE WILLEKEUR
De output van generatieve AI is iedere keer een nieuwe combinatie van ogenschijnlijke willekeur en waarschijnlijkheden op basis van patronen geïdentificeerd in de data tijdens de eerdere trainingen

Toelichting: *antwoorden gegenereerd door ChatGPT op 9 november 2023. De antwoorden zijn telkens gegenereerd in een nieuw chatgesprek op basis van dezelfde vraag. Het taalmodel is gevoelig voor de precieze woordkeuze in de input. Wanneer in het prompt het woord "wat" in "noem" werd verandert, kwamen namen van – telkens verschillende – restaurants terug in het antwoord.*

De mate van variabiliteit in de output is in beginsel een modelparameter die aangepast kan worden. Maar, omwille van gebruikersvriendelijkheid is deze variatie in sommige populaire tools voor generatieve AI echter niet beschikbaar. De variabiliteit in de output van generatieve AI benadrukt de noodzaak dat gebruikers zich hiervan bewust zijn en van zorgvuldige interpretatie. In situaties waarin precisie essentieel is, is het inzetten van generatieve AI niet passend

Net als bij alle algoritmes en AI, moeten gebruikers bovendien rekening houden met bias die ontstaat in de modeltraining. Dat AI-systemen bias kunnen vertonen, is inmiddels breed bekend. Een belangrijke oorzaak is trainingsdata die onvoldoende representatief zijn voor de huidige samenleving of zoals deze gewenst is. In relatie tot generatieve AI springen hierbij nadrukkelijk beeldmodellen in het oog, die bijvoorbeeld – heel eenzijdig – artsen onevenredig vaak als man visualiseren en verplegers onevenredig vaak als vrouw. Daarbij presenteren AI-modellen voor beeldcreatie een gebruiker vaak verschillende varianten van de output, waardoor de bias nadrukkelijker naar voren komt. Bij taalmodellen is uiteraard eenzelfde bias aanwezig, maar is deze minder zichtbaar. Bijvoorbeeld doordat deze via finetuning is onderdrukt (maar onderliggend nog steeds aanwezig is).

De gebruiksrisico's van generatieve AI benadrukken de noodzaak van zorgvuldige mens-machine-interactie... In traditionele algoritmische processen functioneert AI autonoom, met eventueel menselijke tussenkomst om de uitkomsten te beoordelen. Generatieve AI daarentegen vereist veelal actieve input van gebruikers. Dit creëert een andere dynamiek. Hierdoor is het essentieel dat gebruikers de werking en beperkingen van het gebruikte AI-model goed begrijpen. Een eerste vuistregel kan zijn dat een generatief AI-model niet als zoekmachine, feitendatabase of rekenmachine gebruikt kan worden – al is hier een nuance op zijn plaats als het model gebruikmaakt van daarvoor geschikte plug-ins en het gebruik daarvan transparant en verifieerbaar is in de output. Een tweede vuistregel kan zijn dat een gebruiker voorbereid moet zijn op foutieve of valse output, net als willekeur en bias in de output.

...ook omdat te veel afhankelijkheid van generatieve AI het verlies riskeert van waardevolle menselijke inzichten en creativiteit. Generatieve AI-modellen bieden efficiëntie en nieuwe perspectieven. Dit betekent echter niet automatisch dat generatieve AI altijd de eerste stap moet zijn in het denk- en creatieproces. Een aanpak waarbij menselijk denkwerk in eerste instantie vooropstaat, waarborgt de aanwezigheid van menselijke expertise en nuance die AI niet kan repliceren. Generatieve AI kan vervolgens behulpzaam zijn voor verfijningen (bijvoorbeeld het uitschrijven van een brief), controle en herschrijven (mist er iets of kan het duidelijker?) of variatie (opties).

Systeemrisico's van foundation models

De schaalgrootte die nodig is voor de huidige generatieve foundation models introduceert concentratierisico's. De afgelopen jaren is de hoeveelheid data waarop foundation models getraind worden, steeds verder toegenomen. Het zijn op dit moment de grootste modellen die de indrukwekkendste vaardigheden laten zien bij taal- en beeldcreatie. Het trainen van deze modellen kost steeds meer data, rekenkracht en tijd. De hoge initiële investeringskosten zorgen voor toetredingsbarrières en maken dat het op dit moment vooral bigtechbedrijven zijn die de hiervoor benodigde financiële investeringen opbrengen. Een ander gevolg is de kans dat straks enkele foundation models dominant zijn, waardoor het ecosysteem beperkt is. De afhankelijkheid van een klein aantal foundation models brengt concentratierisico met zich mee (***single point of failure***). De modelmatige fouten (bias), veiligheidsissues (cybersecurity) en organisatorische risico's (governance en kans op falen hiervan) die kunnen samenhangen met een foundation model, kunnen doorwerken in alle diensten en gebruikers die van een model afhankelijk zijn.

Wetenschappers en toezichthouders wijzen daarnaast op het risico van homogenisatie en kuddegedrag in AI-modellen, dat door foundation models kan ontstaan. Homogenisatie houdt in dat op dit moment vrijwel alle krachtige foundation models afgeleid zijn van (de techniek achter) enkele modellen die een paar jaar geleden zijn ontwikkeld. Het Center for Research on Foundation Models (Stanford University) heeft reeds in 2021 gewezen op het risico dat al deze modellen dus kwetsbaar zijn voor dezelfde risico's en bijvoorbeeld gevoelig voor hetzelfde type bias.[16] Deze toenemende homogenisatie is volgens het instituut een gevaarlijke zaak. Vanuit ethisch- en veiligheidsoogpunt moet het terugdringen van de gezamenlijke risico's daarom de centrale uitdaging zijn in de verdere ontwikkeling van deze modellen. In het verlengde van deze gedachtegang heeft de Amerikaanse Securities and Exchange Commission (SEC) gewezen op het risico dat marktpartijen in de financiële sector zich baseren op dezelfde (generatieve) AI-modellen en hierdoor kuddegedrag gaan vertonen. Dit kan gelijktijdig leiden tot monocultuur en grotere netwerkverwevenheid in het financiële systeem, wat systeemrisico's met zich meebrengt.[17]

Het gevolg van homogeniteit van AI-modellen kan zijn dat bepaalde groepen personen consequent worden benadeeld. En dat dit pas meetbaar wordt als we naar het ecosysteem als geheel kijken. Recent onderzoek heeft dit bij traditionelere machine learning modellen aan het licht gebracht. Zo'n stelselmatige benadeling kan zich bijvoorbeeld voordoen bij AI voor werving en selectie of voor huidonderzoek. Het onderzoek laat hiermee ook zien dat biasreductie binnen een individueel systeem meestal ten goede komt aan personen die al correct beoordeeld worden door andere systemen.[18]

Het toenemende gebruik van generatieve AI kan via terugkoppelingseffecten op termijn ook zijn weerslag hebben op de prestaties van foundation models (*model collapse*). Dit risico ontstaat als foundation models worden getraind op (synthetische) data die weer de uitkomst zijn van eerdere foundation models. Hoe meer de output van generatieve AI zijn weg vindt in de echte wereld, hoe moeilijker het wordt om deze gegenereerde teksten of beelden te onderscheiden van authentieke teksten of beelden.[19] Door het trainen op gegenereerde data vervormt een foundation model zich als het ware, net zoals een foto van een foto minder van kwaliteit is dan het origineel. Om de prestaties van foundation models daarom in stand te houden, neemt de behoefte aan authentieke data toe.

Desinformatie

Generatieve AI-modellen zijn zeer geschikt voor het genereren van beelden. Dit vraagt om een andere benadering van de betrouwbaarheid van beelden. Deze gegenereerde beelden verschillen in kwaliteit en gelijkheid van 'echte' beelden, zoals nieuwsfoto's. Door de snelle ontwikkeling van deze modellen worden de kwaliteit en gelijkheid echter steeds beter. Dat heeft dit jaar al met regelmaat geleid tot beelden die als 'echt' werden ingeschat, maar in werkelijkheid gegenereerd waren. Ook is toegang tot deze modellen steeds laagdrempeliger, niet alleen voor het genereren van beelden maar ook van geluid (stemmen). Het risico hiervan is dat mensen fictieve, gegenereerde beelden voor echt zullen gaan aanzien. Met bijvoorbeeld als gevolg de verspreiding van des- en misinformatie en de daarmee gepaard gaande bedreiging voor het functioneren van onze democratie.

Om zulke risico's het hoofd te bieden, zal de AI-verordening het 'watermerken' of labelen van gegenereerde beelden in bepaalde gevallen verplicht stellen. Ook heeft een aantal ontwikkelaars van generatieve AI-modellen aangegeven dit watermerken of labelen onderdeel te willen maken van hun modellen. Door het snel groeiende aantal toepassingen van AI-modellen kunnen steeds meer gebruikers beelden genereren. Maar ook worden deze modellen aangepast voor specifiek gebruik, waardoor watermerken of labelen bij een steeds grotere groep eindgebruikers komt te liggen. Waar mensen nu beelden vertrouwen, is een heel pakket aan waarborgen en educatie nodig om te voorkomen dat mensen vertrouwen verliezen in de betrouwbaarheid van beelden en dat dit maatschappelijke schade met zich meebrengt.

Het verifiëren van de herkomst van beelden wordt steeds belangrijker. Er zijn al initiatieven om daar handen en voeten aan te geven. Een initiatief dat een andere benadering heeft gekozen, is het Content Authenticity Initiative. Dit is een coalitie van mediabedrijven, technologiebedrijven en NGO's, die gezamenlijk technische standaarden hebben ontwikkeld die het onder meer met cryptografie mogelijk maken om de herkomst van beelden te verifiëren. Een aantal bedrijven gebruiken de C2PA-standaard (Coalition for Content Provenance and Authenticity) waarop in het najaar van 2023 ook al camera's zijn gelanceerd waar deze standaard is ingebouwd. Dit maakt twee belangrijke waarborgen mogelijk voor de betrouwbaarheid van beelden. Ten eerste dat organisaties de herkomst van beelden kunnen verantwoorden en hierop hun processen kunnen inrichten. Ten tweede dat het voor iedereen die beelden 'consumeert' mogelijk wordt om het denken over en de kennis van de betrouwbaarheid van beelden te veranderen of te ontwikkelen.

Vertrouwen op je zintuigen is niet meer altijd mogelijk, dit zal gaan verschuiven naar het vertrouwen op de bron en verifieerbaarheid van beeld. De herkomst en betrouwbaarheid kunnen controleren van beelden die een grote maatschappelijke waarde of impact hebben, is een waardevolle bijdrage aan het beheersen van algoritmes en de effecten die deze kunnen hebben.

Toezicht op foundation models en generatieve AI

In Europa zal de AI-verordening een toezichtsregime introduceren voor foundation models, in de verordening ook wel '*general purpose AI*' (GPAI) modellen genoemd. De AP is positief over de toezichtsbevoegdheden die worden gegeven aan een nieuw op te richten Europees AI Office binnen de Europese Commissie. Hiermee ontstaat ook een wendbare structuur om in de toekomst in te spelen op verdere ontwikkelingen in foundation models en hun capaciteiten. De AP zal zich inzetten om met risicosignalering en adviezen bij te dragen aan de verdere ontwikkeling van regels voor GPAI-modellen.

De AI-verordening zal voorzien in een getrapt regime voor foundation models. Er zullen enkele algemene verplichtingen gaan gelden voor alle GPAI-modellen. Voor GPAI-modellen met systeemrisico's zullen aanvullende en specifieke eisen gelden. Aanbieders van hoog risico AI-systemen die voortbouwen op GPAI-modellen zullen zich aan de reguliere bepalingen van de verordening moeten houden en hun AI-systemen bijvoorbeeld op conformiteit moeten toetsen.

Voor alle aanbieders van foundation models zullen verplichtingen gelden die helpen om risico's verderop in de AI-keten te mitigeren. Aanbieders van GPAI-modellen moeten (1) hun model documenteren en (2) informatie bieden aan 'downstream' aanbieders van AI-systemen. Verder moeten aanbieders mogelijk een beleid ontwikkelen om auteursrechten te respecteren en een samenvatting geven over de data die is gebruikt om het model te trainen. Ook moeten aanbieders van modellen die kunnen worden gebruikt om inhoud te genereren, zoals afbeeldingen en tekst, een digitaal watermerk opnemen in de inhoud.

De AI-verordening bevat een aanvullend regime voor foundation models met systeemrisico's (systeemrelevante GPAI-modellen). Van een dergelijk model is volgens de AI-verordening sprake als het op basis van technische mogelijkheden een grote impact kan hebben of als het een aanzienlijke impact heeft op de interne markt. Om dat te bepalen kan worden gekeken naar kwantitatieve criteria zoals de (grote) hoeveelheid rekenkracht waarmee het model is getraind, het aantal parameters en het aantal zakelijke gebruikers van het model. Een foundation model is in ieder geval systeemrelevant vanaf een bepaalde benodigde rekenkracht voor training, met 10^{25} ***floating point operations per second*** (FLOPs) als vermoedelijke ondergrens. Voorlopige inschattingen die circuleren in de media zijn dat de meeste GPAI-modellen op dit moment niet voldoen aan de grens, met ***Gemini*** (Google) als mogelijke uitzondering.

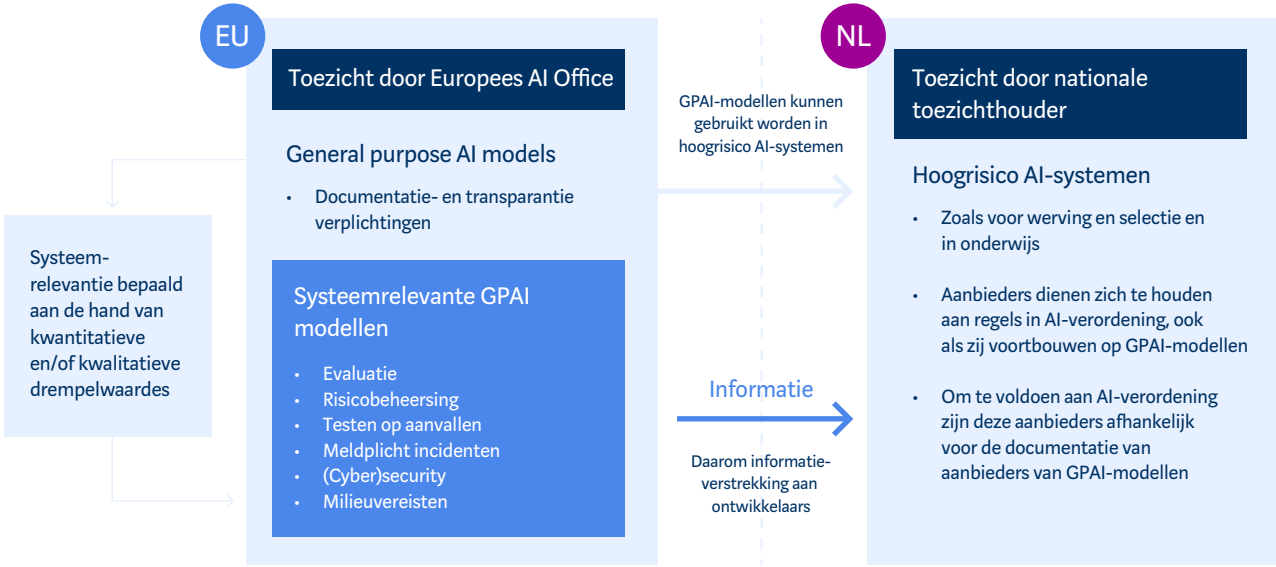
Voor aanbieders van systeemrelevante foundation models gelden strengere eisen. De aanvullende eisen waaraan aanbieders van deze modellen moeten voldoen zijn dat zij (1) hun modellen evalueren (2) mogelijke systeemrisico's van hun modellen identificeren en aanpakken, (3) hun modellen testen op aanvallen, (4) incidenten melden, (5) veiligheidsmaatregelen moeten nemen, zowel op het gebied van cybersecurity als met het oog op de fysieke infrastructuur van het model en (6) het voldoen aan bepaalde milieuvereisten.

Een nieuw op te richten Europees AI Office zal toezicht houden op foundation models. Dit AI Office zal de regels over foundation models in de verordening kunnen handhaven. De AI Board speelt een rol via betrokkenheid in de verdere uitwerking van standaarden voor foundation models (zie verder hoofdstuk 5). Daarnaast kan het AI Office systeemrelevante foundation models ook op basis van kwalitatieve criteria aanwijzen. Figuur 8 geeft een schematische weergave van de beoogde regulering van GPAI-modellen in de AI-verordening.

Naast de AI-verordening is er behoefte aan kaders voor veilig gebruik van generatieve AI door organisaties, de AP wil hier op inzetten. Doordat de mens-machine-interactie bij generatieve AI een andere vorm heeft dan bij overige algoritmes en AI, is er behoefte aan duidelijkheid over de wijze waarop organisaties generatieve AI kunnen inbedden in hun processen. Dit kan voortbouwen op de verantwoordelijkheid die producenten onder de AI-verordening hebben om gebruikers te voorzien van, simpel gezegd, een handleiding voor veilig gebruik. Maar het valt niet uit te sluiten dat ook een organisatiebreed perspectief nodig is. Bijvoorbeeld een specifieke uitwerking voor generatieve AI die past binnen de overkoepelende internationale standaarden voor AI-risicobeheersing. De AP zet erop in om komend jaar met nationale en internationale partners na te denken over hoe dit vorm moet krijgen.

Ook buiten Europa wordt gewerkt aan nieuwe kaders voor foundation models. De president van de Verenigde Staten heeft vanwege de nationale veiligheid onlangs via een ***executive order*** in feite een toezichtregime geïntroduceerd voor de grote foundation models. De gedachte achter deze maatregelen is dat foundation models in ieder geval in theorie ook misbruikt kunnen worden door kwaadwillenden en het daarom belangrijk is er een bepaalde mate van grip op te krijgen. Vanaf een bepaalde modelomvang – die op dit moment waarschijnlijk nog niet wordt bereikt – moeten ontwikkelaars van deze foundation models op continue basis de federale overheid voorzien van informatie over deze modellen. Daarbij moeten zij rapporteren over de training, ontwikkeling en productie van nieuwe foundation models en de beschermende maatregelen, zowel de fysieke als de op cybersecurity gerichte maatregelen. Hiertoe behoort ook transparantie over het eigenaarschap en bezit van de modelgewichten. Technische uitwerking van deze maatregelen volgen in de komende periode.

FIGUUR 8: REGULERING VAN GENERAL PURPOSE AI-MODELLEN



3. Algoritmes en AI op de werkvloer

Algoritmes en AI worden op steeds meer terreinen ingezet en raken zo verweven met veel aspecten van onze samenleving. Ook waar het niet direct zichtbaar is, worden algoritmes ingezet voor diverse doelen. Zoals bij het verdelen en managen van arbeid. Deze algoritmes op de werkvloer kunnen zorgen voor meer efficiëntie, maar kunnen ook leiden tot ongewenste effecten voor werknemers.

Ontwikkeling arbeidsmarkt

De opkomst van algoritmes en AI raakt veel arbeidssectoren, op verschillende manieren. In veel sectoren worden grote veranderingen verwacht door de inzet van algoritmes en AI.[20] De ontwikkeling van nieuwe technologieën, waaronder AI, zorgt volgens het Centraal Planbureau (CPB) zowel voor het verdwijnen en veranderen van bestaande banen als voor de introductie van nieuwe banen.[21] De gevolgen die AI op de werkvloer teweegbrengt, zijn nog in beweging en daardoor maar beperkt scherp. Duidelijke risicobeelden zijn essentieel om effectieve regelgeving te maken voor de inzet van AI op de werkvloer.

Een van die verschuivingen is te vinden in de platformwerksector. Platformwerk gaat hand in hand met de inzet van algoritmes. Deze sector is met 65% toegenomen van 2015 tot 2019.[22] Op dit moment zijn er ongeveer 100.000 mensen in Nederland werkzaam via een platform.

Maar algoritmes spelen ook een steeds grotere rol bij arbeid die niet via platformen verloopt. Zo wordt er bij de aansturing van personeel, maar ook bij de werving en selectie van nieuw personeel, steeds vaker gebruikgemaakt van algoritmes.[23] Het gebruik van algoritmes voor werving en selectie kent veel risico's terwijl de werking van deze systemen en toepassingen maar zelden bewezen is.

Algoritmes en AI op de werkvloer

Meer dan 75% van de bedrijven in Nederland zet in de komende vijf jaar AI-applicaties in op de werkvloer.[24] **Daarnaast zet 40% van de grote bedrijven op dit moment al AI-systemen in op de werkvloer.**[25] Dit leidt tot automatisering van de aansturing van arbeid, het zogenoemde 'algoritmisch management' zoals beschreven door de ***International Labour Organization***.[26] Hoewel modellen voor het optimaliseren van werk en productie al langer worden toegepast, slagen organisaties er vaak in om op basis van data en met nog meer precisie de efficiëntie van werknemers te verhogen. Het inzetten van AI-systemen op de werkvloer zal worden gereguleerd door de AI-verordening, voor zover het gaat om werving en selectie en beslissingen die worden genomen over mensen die werken.

Op internationaal niveau worden de effecten van algoritmes en AI op de werkvloer hoog op de wetgevingsagenda gezet. De uitdagingen van AI-systemen op de werkvloer worden specifiek benoemd als aandachtspunt voor de komende jaren door de Europese Commissie.[27] In de Verenigde Staten zal de overheid ook verder onderzoek doen naar de effecten van algoritmes en AI op de werkvloer.[28] Begin 2024 worden daar de eerste rapporten gepubliceerd, waarop beleidsvoorstellen kunnen volgen. Het in kaart brengen van risico's van algoritmes en AI op de werkvloer is uiteraard een essentiële eerste stap in een wetgevingstraject. Waar risico's al wel scherp in beeld zijn, zijn wetgevingstrajecten in gang gezet. De Europese Platformwerkrichtlijn is hier een voorbeeld van.

De inzet van algoritmes en AI om arbeidsproductiviteit te bevorderen is in sommige sectoren eerder regel dan uitzondering. Bijvoorbeeld in de bezorgsector, waar arbeidsproductiviteit en strakke planning van groot belang zijn. Maar ook in andere arbeidsrelaties worden algoritmische managementsystemen ingezet om de efficiëntie te verhogen en werknemers in staat te stellen meer werk te verrichten in minder tijd. Algoritmes kennen een grote ***span of control*** en maken aansturing van werknemers op grotere schaal mogelijk.

Werkmonitoring door algoritmes kan ervoor zorgen dat mensen als arbeidsproduct worden benaderd. Sommige werknemers moeten in hun dagelijkse werk steeds meer vertrouwen op een algoritmisch managementsysteem. Hierin schuilen risico's die veroorzaakt kunnen worden door het gebruik van algoritmes. Zo kunnen werknemers zich continu bekeken en geanalyseerd voelen door een algoritmisch systeem. Ook kan een vrees ontstaan voor slechte dagen met een verminderde productiviteit. Door algoritmisch management verschuift de algehele gezagsverhouding op de werk-

vloer zich van mens-tot-mens naar algoritme-tot-mens. Hoe groter de rol van algoritmes in de arbeidssturing, hoe belangrijker dus de rol van een manager die een menselijke maat kan waarborgen.

Risico's van algoritmische managementsystemen

Algoritmes en AI worden in specifieke beroepsgroepen ingezet en kennen daarmee ook sectorspecifieke risico's, naast de bekendere algemene risico's. Een systeem dat een groep van magazijnmedewerkers aanstuurt, kan voor een fastfoodrestaurant minder bruikbaar zijn dan voor bijvoorbeeld een webwinkel. Ook kunnen de risico's anders zijn. Algoritmes inzetten om werknemers te managen, kan dus in verschillende sectoren en bedrijven anders uitpakken. Grip krijgen op de risico's van de inzet van een algoritmisch managementsysteem vergt daarom maatwerk.

Algoritmische systemen kunnen zorgen voor meer werkdruk. Deze toename in werkdruk kan ervoor zorgen dat werknemers sneller opgebrand raken dan bij menselijke managementsystemen als er onvoldoende oog is voor persoonlijke omstandigheden. Omdat deze werknemers te maken hebben met een algoritmisch systeem, kan het voor deze werknemers bovendien lastiger zijn om arbeidsdruk te bespreken.

Algoritmes en AI hebben niet altijd perfect inzicht in de fysieke wereld. Dat kan leiden tot inefficiëntie, last en frustratie bij werknemers. De werkdruk van algoritmisch gemanaged werk kan ook juist te laag liggen. Zoals bij enkele bezorgdiensten in Nederland. Als pakketten binnen een beperkt tijdvak geleverd moeten worden, dan kan het voorkomen dat bezorgers juist verplicht moeten wachten voor zij aan hun levering in het volgende bezorgvenster kunnen beginnen. Soms moeten bezorgers voor een huis stilstaan en kunnen zij een pakket nog niet afleveren, omdat het systeem geen duidelijkheid verschaft over welk pakket voor het huis bestemd is, tot het tijdvak van bezorging is ingegaan.

De inzet van algoritmische systemen kan leiden tot gedragsveranderingen bij werknemers. In callcenters worden bijvoorbeeld algoritmische systemen gebruikt die het inkomend telefoonverkeer verdelen over de beschikbare werknemers en die tegelijkertijd werknemers monitoren. Zo zijn er systemen die de emotie in de stem van een callcenterwerknemer analyseren en de werknemer wijzen op een gebrek aan hoorbare empathie.[29] Algoritmische managementsystemen kunnen werknemers dus aanmoedigen of zelfs aansporen om hun gedrag te veranderen om productiever of effectiever te werken. Bijvoorbeeld door de toon van de stem te veranderen. Op die manier kunnen algoritmes op de werkvloer het gedrag van werknemers in verregaande mate beïnvloeden.

Algoritmes en AI kunnen de autonomie van werknemers aantasten. Bij primaire aansturing door een manager kunnen werknemers aangeven dat zij processen op een andere manier dan het algoritme willen uitvoeren. Met een algoritmisch managementsysteem kan dit minder makkelijk. Opgedane kennis van werknemers vertaalt zich niet snel door in algoritmisch management. Daarom neemt de mogelijkheid om in te spelen op toekomstige uitdagingen en gebruikmaken van ervaring en individuele kwaliteiten mogelijk af bij de inzet van een algoritmisch managementsysteem. Werknemers ervaren hierdoor minder autonomie en meer controle op de werkvloer. Werknemers hebben zo steeds minder de regie over hun eigen werkzaamheden.

Negatieve effecten op publieke waarden en grondrechten. De effecten van de inzet van algoritmische systemen op de werkvloer hebben ook impact op publieke waarden en grondrechten. Bijvoorbeeld aantasting van het recht op privacy, maar ook van het recht op gelijke behandeling. Daarnaast kunnen algoritmische managementsystemen invloed hebben op de mate waarin en de manier waarop werknemers informatie tot zich kunnen nemen.

Algoritmische managementsystemen verzamelen veel data en halen hier inzichten uit, die voor de werknemers mogelijk niet toegankelijk zijn. Deze ongelijke informatiepositie kan leiden tot een zwakkere positie van werknemers.

Platformwerk

Platformwerkers ondervinden grotere risico's dan anderen door algoritmisch management. Deze platformwerkers staan vaak niet sterk tegenover de aanbieder van het platform. Er opereren bovendien veelal geen menselijke managers op de werkvloer. Daardoor zijn de risico's op een onredelijke werkdruk, minder autonomie, minder veiligheid en gedragsbeinvloeding voor platformwerkers in potentie nog groter. Een goede bescherming bieden aan deze platformwerkers is dus een essentieel onderdeel van een duurzame structurering van de arbeidsmarkt waarin algoritmes en AI een steeds grotere rol zullen spelen. De Europese Platformwerkrichtlijn, waarover op 13 december een politiek akkoord is bereikt, zal hierin ook een belangrijke rol spelen.

Risico's inzet algoritmes en AI voor platformwerkers

Bij veel platformwerk is het verweren tegen de risico's lastig. Bij platformwerk ontbreekt vaak de sociale component die verzet of verweer tegen de aanbieder of het platform kan bespoedigen. Als de aard van het werk meebrengt dat platformwerkers weinig of geen contact hebben met collega's, dan zullen platformwerkers ook maar beperkt ervaringen uitwisselen. Dat maakt het lastig om samen veelvoorkomende problemen te ontdekken en hiervoor samen gericht oplossingen af te dwingen bij een platform.

Opdrachten die platformwerkers via een platform krijgen aangeboden, kunnen zij lang niet altijd weigeren zonder consequenties. Het weigeren van opdrachten die weinig opleveren, lang duren of extra werk met zich meebrengen, kan gevolgen hebben voor het aanbod van toekomstige opdrachten en dus voor toekomstige inkomsten. Platformwerkers die meer opdrachten accepteren, zullen sneller in aanmerking komen voor opdrachten met goede verdiensten of prettige werkzaamheden. Platformwerkers zijn daardoor erg afhankelijk van de algoritmes van het arbeidsplatform.

Algoritmische ratingsystemen kunnen leiden tot gedragsverandering bij platformwerkers. ***Rating systems*** zijn erg populair bij aanbieders van platformwerk. Daarnaast hebben goede ratings een positief effect op veel aspecten van de arbeid die een platformwerker uitvoert. En platformwerkers zijn voor het krijgen van werk afhankelijk van een goede rating. De ratings worden namelijk door algoritmes meegenomen bij het verdelen van opdrachten over de beschikbare platformwerkers. De verhoudingen tussen platformwerkers en afnemers zijn erop ingericht dat platformmedewerkers ambiëren een zo hoog mogelijke score te ontvangen.

Toekomstige regulering

De Platformwerkrichtlijn zal eisen stellen aan de inzet van algoritmisch management door digitale arbeidsplatformen. De richtlijn zal meer duidelijkheid scheppen over de arbeidsstatus van mensen die via een platform werken. Maar de richtlijn zal ook eisen stellen op het punt van transparantie en grenzen stellen aan het algoritmisch management door platformaanbieders.

In afwachting van wetgeving kunnen organisaties zelf aan de slag om algoritmische managementsystemen op een gedegen manier te laten verlopen. Hierbij zijn voldoende organisatorische maatregelen van groot belang. Het aansturen van arbeidskrachten met algoritmes en AI kan – mits op een verantwoorde manier geregeld – allerlei voordelen meebrengen, ook voor werknemers en werkgevers.[30] De introductie van nieuwe technologieën heeft de arbeidsmarkt keer op keer veranderd.[31] Algoritmes en AI zullen hierop geen uitzondering vormen, maar de effecten hiervan op werknemers moeten scherp in de gaten worden gehouden. Organisaties moeten zo worden ingericht dat algoritmes en AI op de werkvloer niet enkel efficiëntiedoeleinden dienen en risico's van algoritmes worden ondervangen.

Interview: Trainen van algoritmes en AI is nieuwe werkvorm

Dit interview is in november 2023 afgenomen bij een platformwerker in Nederland die onder andere werkt aan het trainen van AI-modellen, zoals de taalmodellen die in 2023 veel aandacht hebben gekregen. Wereldwijd werken naar schatting tienduizenden mensen als platformwerker aan het beoordelen en verbeteren van de uitkomsten van algoritmes en AI-modellen.

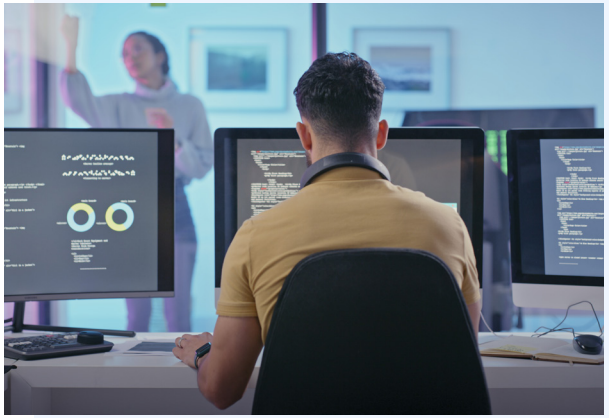
Kun je je werkzaamheden als AI-trainer beschrijven? "Ik verricht werk voor een aantal online aanbieders, waarbij ik grotendeels algoritmes train op juistheid en relevantie, bijvoorbeeld van zoekresultaten. Recent is daar het herschrijven en beoordelen van antwoorden van AI-modellen bijgekomen. Die beoordeel ik op juistheid en kwaliteit van de antwoorden. En ik check of het model zich niet als een persoon voordoet. Maar ook het bedenken van prompts en het herschrijven van antwoorden van het AI-model behoren tot de opdrachten die ik doe."

Wat is je overweging geweest om deze vorm van werk te gaan doen? "Doordat al het werk op online platformen gebeurt, past dit goed bij mij. Het geeft vrijheid, het is te combineren met andere creatieve werkzaamheden en is niet afhankelijk van je verblijfplaats."

Zitten er ook nadelen aan deze vrijheid? "Jazeker, deze vorm van werk geeft geen regelmatig inkomen en het aanbod van werk varieert sterk. Er zijn geen garanties voor mij als platformwerker, niets om op terug te vallen, zoals een contract. Je kunt van de ene op de andere dag je werk en dus je inkomen verliezen. Ook worden vergoedingen plotseling naar beneden bijgesteld. Betalingen die niet gedaan worden – wat voorkomt – zijn zonder een menselijk aanspreekpunt moeilijk of niet te krijgen."

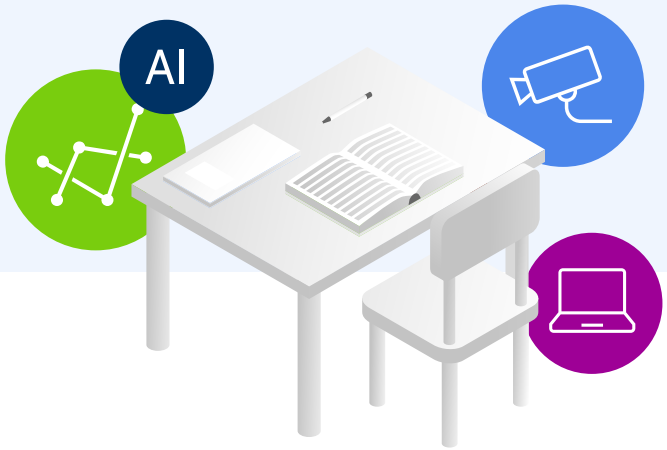
Wat doe je precies bij het trainen van algoritmes? "Ik werk bij dit soort projecten met duizenden mensen die wereldwijd deze modellen en systemen verbeteren en voorzien van antwoorden, teksten en beoordeling. Dat steeds meer mensen van over de hele wereld aan deze AI-modellen werken, zorgt wel voor verslechtering van de taal in deze modellen. Dit komt door continue druk op de verdiensten van mensen en de kwaliteit die mensen moeten leveren, zonder dat ze enige bescherming daarin hebben."

Welke invloed hebben algoritmes op jouw werkzaamheden? "Door algoritmische beoordeling van de kwaliteit van je opdracht of de tijd die je eraan hebt besteed, kun je soms je toegang tot projecten verliezen. Bijvoorbeeld als je te snel een opdracht afrondt. Of als de kwaliteit van je opdracht niet als voldoende wordt beoordeeld. Alleen is vaak niet duidelijk wat de regels of richtlijnen zijn. Dus je moet bij sommige projecten voldoen aan regels die je niet kent. Daardoor voel je je als werker inwisselbaar. Eigenlijk



wordt systematisch, maar ondoorzichtig, beoordeeld of jij recht hebt op je inkomen die maand. Maar er zijn ook projecten of aanbieders waarbij juist een overvloed aan informatie is op basis waarvan je aan opdrachten moet werken."

Heeft dit werk invloed gehad op je vertrouwen in algoritmes? "Ik heb niet direct meer of minder vertrouwen gekregen in algoritmes. Maar dit werk helpt mij wel om algoritmes en AI op waarde te schatten. Als platformwerker zie je de kwaliteit van die AI-modellen. Dit helpt enorm om de hype te relativeren. Tegelijkertijd zie je ook het blinde vertrouwen dat mensen in de modellen stellen. Dat blijkt uit de interacties tussen mensen en de modellen, zoals de prompts die mensen geven en die ik te zien krijg. Ik maak mij zorgen over hoeveel vertrouwen mensen in modellen stellen die in essentie nog zoveel mensenwerk bevatten."



4. Algoritmes en AI in het onderwijs

De onderwijssector (primair, voortgezet, middelbaar beroeps- en hoger onderwijs) maakt steeds meer gebruik van algoritmes en AI. Zo gebruiken onderwijsinstellingen adaptieve leersystemen, die de lesstof geautomatiseerd en op de individuele leerling aangepast aanbieden. Systemen met 'learning analytics' geven nieuwe inzichten op basis van allerlei data, met als doel de doorstroming van studenten in het onderwijs of de onderwijskwaliteit te verbeteren. In de kern voorzien deze systemen veelal in een profilering van en voorspelling over leerlingen of studenten. Er zijn allerlei mogelijke oorzaken waardoor profilering of voorspelling niet goed aansluit bij de situatie van individuele leerlingen. Zorgvuldige inbedding van algoritmes en AI in het onderwijs en kennis van de beperkingen ervan is cruciaal, omdat het om kinderen en jongvolwassenen gaat. Leerlingen en studenten gebruiken veel generatieve AI. Dat doen ze om een deel van hun 'maakopdrachten' uit te besteden. Dat daagt het onderwijs uit om het gebruik van generatieve AI op een goede manier te verwerken in de doelen die men nastreeft. Het onderwijs heeft zoiets eerder succesvol gedaan bij de opkomst van het internet en Wikipedia. De AI-verordening zal een aantal AI-systemen voor het onderwijs als hoogrisicotoepassingen reguleren waarvoor aan productstandaarden moet worden voldaan, maar deze zullen vooral gelden voor ontwikkelaars en niet altijd voldoende concreet zijn. De DCA adviseert daarom om beleidsstrategieën en beheersingsprocessen voor algoritmes en AI in te richten binnen onderwijsinstellingen, onder begeleiding van de aanwezige externe ondersteuningsorganisaties. In aanvulling daarop moet de onderwijssector collectief heldere productstandaarden met ontwikkelaars van AI-systemen overeenkomen. Ook het vergroten van AI-kennis onder leerkrachten en docenten is een aandachtspunt, om te zorgen voor een zorgvuldige inbedding en beheersing van algoritmes in het onderwijs.

Het onderwijs maakt momenteel een intensivering van digitalisering en algoritmegebruik door. De onderwijssector stapte in de jaren voor de coronapandemie al in toenemende mate over naar digitale lesmiddelen en leeromgevingen. De pandemie heeft die digitalisering in het onderwijs geïntensiveerd.[32] Dit maakt het verzamelen van data over de prestaties van leerlingen gemakkelijker. De aanwezigheid van meer data betekent vaak dat het beter mogelijk is om algoritmes te gebruiken. Voor de huidige intensivering van digitalisering en de inzet van algoritmes in het onderwijs zijn niet alleen meer data, maar ook meer financiële middelen beschikbaar. Vanuit het Nationaal Groeifonds is er 80 miljoen euro beschikbaar gesteld voor het Nationaal Onderwijslab AI, dat onderzoek doet naar de kansen en risico's van AI in het onderwijs.[33][34] Ook hebben veel scholen een deel van het budget van het Nationaal Programma Onderwijs gebruikt om digitale leermiddelen aan te schaffen.[35]

Om algoritmes verantwoord in te zetten, moet de onderwijssector beter kunnen inschatten wat de sector met de inzet van digitale middelen kan en wil bereiken. Uit cijfers van het Kohnstamm Instituut[36] en Npuls magazine[37] blijkt, zoals weergegeven in figuur 9, bijvoorbeeld dat de helft van de scholen in het primair onderwijs geen strategisch ICT-beleid heeft. Verder heeft meer dan de helft van de leerkrachten geen kennis over AI.

Ook is er bij slechts enkele hogescholen en mbo's beleid voor AI-gebruik door studenten. Het gros van de studenten aan universiteiten maakt inmiddels echter al gebruik van Large Language Models. In het geval van de Erasmus Universiteit zelfs meer dan 90 procent.[38] Deze cijfers onderstrepen de noodzaak van visie op en beleid voor digitalisering in het onderwijs.

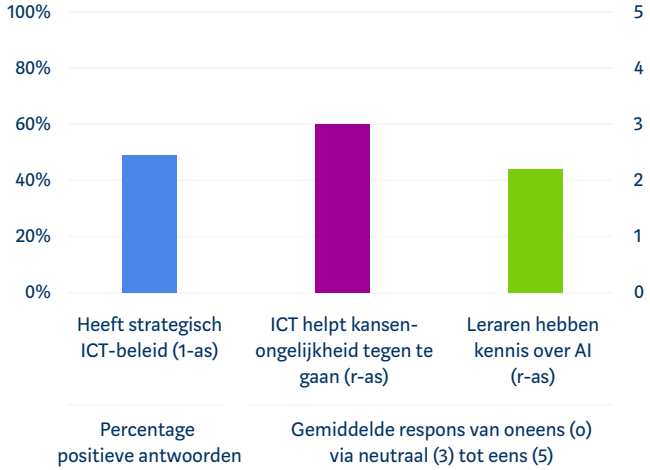
Twee bekende algoritmetoepassingen in het onderwijs zijn adaptieve leersystemen (AL) en *learning analytics* (LA). Er zijn nog veel meer algoritmetoepassingen in het onderwijs, van toepassingen die docenten lessen helpen samenstellen in het voortgezet onderwijs tot algoritmes die helpen met het controleren van toetsen of huiswerk. AL en LA zijn echter de bekendste voorbeelden van commercieel beschikbare systemen die onderwijsinstellingen gebruiken en die de risico's van algoritmegebruik in het onderwijs goed zichtbaar maken. AL komt vooral in het primair onderwijs voor, LA in het hoger onderwijs. Dat zijn de deelsectoren die het meest van algoritmes gebruikmaken, het middelbaar beroepsonderwijs en het voortgezet onderwijs gebruiken vooralsnog minder algoritmes.

Adaptief leren

Adaptieve leersystemen selecteren oefenstof voor individuele leerlingen op basis van hoe goed een leerling eerdere oefeningen heeft gemaakt. Meer dan de helft van de leerlingen in het primair onderwijs oefent dagelijks met de lesstof in zo'n systeem. Sterk versimpeld werkt AL als volgt: een leerling maakt drie opgaven achtereen goed en snel. Het AL-systeem legt dat vast en interpreteert het als een teken dat de leerling aan een hoger niveau toe is. Op grond daarvan selecteert het systeem daarna drie opgaven van een hoger begripsniveau. Een leerling die de opgaven langzamer of onjuist maakt, krijgt opgaven van een lagere moeilijkheids-

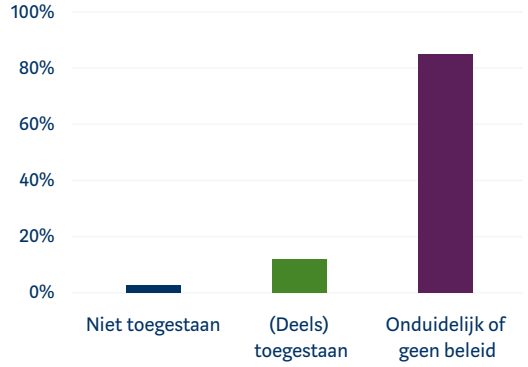
FIGUUR 9: ONDERWIJSINSTELLINGEN VOEREN BEPERKT BELEID OVER INZET ICT EN AI-TECHNIEKEN

In het primair onderwijs ontbreekt het vaak aan strategische visie op het gebruik van ICT en lerarenkennis over AI is laag...*



Respons van schoolleiders in primair onderwijs

... en in het mbo en hoger onderwijs ontbreekt het aan beleid over het gebruik van AI door studenten**



Beleid in Nederland bij onderwijsinstellingen voor mbo en hoger onderwijs

)BRON**: KOHNSTAMM INSTITUUT (2023). *)BRON**: NPULS(2023)

graad voorgeschoteld. Het systeem profileert de leerling dus en stelt het profiel na elke opdracht bij. Het voorspelt op basis van dat profiel welke opdracht de leerling de beste kans op goede voortgang biedt en selecteert die. Veel AL-toepassingen gebruiken daarvoor een algoritme dat ooit is ontwikkeld om schakers te rangschikken. Dat AL vooral in het primair onderwijs voorkomt, heeft ermee te maken dat de helder te beschrijven vaardigheden die leerlingen in het primair onderwijs leren zich goed lenen voor het toepassen van AL. Bovendien is het primair onderwijs een grote markt, doordat elke school toewerkt naar dezelfde leerdoelen.[39] Het primair onderwijs is daarom aantrekkelijker voor aanbieders van algoritmetoepassingen dan het in lesstof meer gedifferentieerde hoger onderwijs.

Leraren gebruiken dashboards van adaptieve leersystemen, die realtime inzicht geven in de interactie tussen het systeem en de leerling. Dashboards zijn digitale omgevingen waarin de voortgang van individuele leerlingen wordt getoond, op kortere of langere termijn. Dashboards komen voor in meerdere delen van de onderwijssector en zijn niet uitsluitend aan AL-systemen gekoppeld. In het voortgezet onderwijs zijn ze dat vaak wel. Een dashboard dat aan een AL-systeem is gekoppeld, helpt leraren om snel het leerproces van leerlingen te kunnen bijsturen. Het dashboard toont namelijk data waarmee de leerkracht kan zien welke leerling op dat moment goed vordert en wie aandacht nodig heeft. Er zijn dashboards die vooral data tonen, maar er zijn ook dashboards die de data ook analyseren en interpreteren. Bijvoorbeeld door leerlingen zichtbaar te categoriseren of zelfs een bepaalde interventie door een leerkracht voor te stellen.

Ondanks de intentie van onderwijs op maat voor elk kind, kunnen adaptieve leersystemen tot kansenongelijkheid leiden. De modellen aan de basis van AL-systemen zijn namelijk vaak minder complex dan de werkelijkheid in de klas. Dat betekent dat leerlingen die structureel makkelijk leren, meestal oefeningen krijgen die hen stimuleren om zich te ontwikkelen, of die hen aan het werk houden op hun niveau. De behoeftes van leerlingen die de lesstof nog onvoldoende beheersen, stelt het systeem moeilijker vast. Het systeem houdt namelijk geen rekening met verschillen anders dan de verschillen in succesratio's tussen leerlingen, zoals de thuissituatie of de concentratiespanne, terwijl dergelijke verschillen juist de soms uiteenlopende behoeftes van kinderen met toch vergelijkbare aantallen goede en foute antwoorden kunnen verklaren. Als een systeem zulke achterliggende verschillen tussen leerlingen moeilijk herkent, worden sommige leerlingen verkeerd geprofileerd en selecteert het systeem daardoor minder goed passende opgaven voor die leerlingen.

Om gelijke kansen bij het gebruik van adaptieve leersystemen te waarborgen, moeten leraren en adaptieve leersystemen elkaar beter aanvullen. Als leerlingen door een AL-systeem onnauwkeurig geprofileerd worden, krijgen ze door het systeem vaak dezelfde opgaven voorgeschoteld, terwijl ze verschillende behoeftes hebben. Om ook die leerlingen te bieden wat ze nodig hebben, moeten leraren op basis van hun dashboard en hun eigen waarneming inschatten wat die leerlingen nodig hebben. Dat vergt enige kennis over de werking van het AL-systeem en het verschil tussen model en praktijk. Dat heet 'algoritmische geletterdheid'. Die is niet altijd aanwezig. Verder moeten leraren, als zij algoritmegeletterd zijn, hun vaardigheid ook daadwerkelijk kunnen inzetten in de ondersteuning van leerlingen. Scholen kunnen AL-systemen dus niet onbeperkt inzetten om leraren vrij te spelen voor andere taken.

Schoolbesturen, leerkrachten en ontwikkelaars van AL-systemen delen de verantwoordelijkheid om het risico van kansenongelijkheid door AL-systemen te beperken en om de kansen op gepersonaliseerd onderwijs te verwezenlijken. Leraren moeten kunnen ingrijpen wanneer een AL-systeem onnauwkeurige, niet passende of onjuiste keuzes voor hun leerlingen maakt. Ze hebben daarvoor 'algoritmegeletterdheid' nodig. Schoolbesturen moeten hun beleid zodanig vormgeven dat leraren de benodigde algoritmegeletterdheid kunnen opdoen en inzetten, bijvoorbeeld door trainingen. Verder is het raadzaam dat schoolbesturen een onverantwoorde aankoop van een AL-systeem moeten voorkomen door vóór de aankoop te bepalen welke implementatiemogelijkheden de school heeft. Daarbij is het ook van belang om een implementatie- en gebruiksorganisatie neer te zetten met een heldere verantwoordelijkheidsverdeling.

Onderzoek en pilots zijn nodig om de interactie tussen leerling, AL-systeem en leraar beter te begrijpen. Dat moet resulteren in een betere wisselwerking tussen AL-systemen en leraren, zodat er synergie tussen mens en machine ontstaat. Daarvoor staan zowel ontwikkelaars, besturen als leraren aan de lat. Daarbij zijn ondersteunende organisaties met expertise op dit gebied, zoals Kennisnet en SURF, onontbeerlijk.

Learning analytics

Learning analytics (LA) is het gebruik van data om inzicht te krijgen in de voortgang van studenten en leerlingen en in de kwaliteit van het onderwijs. LA valt uiteen in twee toepassingsdoeleinden: het verbeteren van de doorstroom en de begeleiding van studenten en het verbeteren van onderwijs. Een adaptief leersysteem dat docenten realtime data over leerlingen toont, maakt in essentie dus ook gebruik van LA. Echter, wanneer gesproken wordt over 'learning analytics', wordt doorgaans gedoeld op het verkrijgen van inzicht uit data over langere perioden, gericht op structurele ingrepen zoals aanpassingen in het curriculum of het preciezer bepalen waardoor studenten vertraging oplopen. LA-systemen worden nog niet massaal gebruikt. Vooral instellingen in het hoger onderwijs onderzoeken echter actief de toepassingsmogelijkheden. Daarin staat SURF hen bij.

LA wordt als een grote belofte beschouwd[40] en is beschikbaar, maar onderwijsinstellingen hebben over het algemeen geen visie op of beleid voor de inzet van LA-toepassingen. Grote leermanagementsystemen, zoals Canvas en BlackBoard, bieden geïntegreerde LA aan. De Universiteit van Utrecht stelde bijvoorbeeld al een team LA in.[41] Veel onderwijsinstellingen weten echter nog niet goed genoeg wat zij precies met een LA-toepassing kunnen en willen op zo'n manier dat zowel studenten als docenten en bestuur er tevreden mee zijn. Het verzamelen van data zonder duidelijk doel kan ertoe leiden dat data op onverantwoorde manieren worden gebruikt. Bovendien is dit niet toegestaan wanneer het persoonsgegevens betreft. Het is daarom raadzaam voor onderwijsinstellingen om learning analytics niet op grote schaal te in te zetten voordat een duidelijke visie op het gebruik ervan is ontwikkeld.

Ondoordachte inzet van LA kan een grote impact op individuele studieloopbanen en privacy hebben. De inzet van LA om individueel studiesucces van studenten te bevorderen, vraagt om profilering en de verwerking van persoonsgegevens. Dat betekent altijd een inbreuk op hun privacy. De inzet van LA zonder duidelijk doel en heldere kaders vergroot de kans op onrechtmatige of onevenredige inbreuken aanzienlijk. Het gebruik van LA in de besluitvorming over studieloopbanen vraagt ook om een goed doordacht plan. Data en profielen gebruiken zonder duidelijkheid over wat de data wel en niet betekenen en welke conclusies eruit getrokken mogen worden, vergroot het risico op onrechtvaardige behandeling van studenten.

De onderwijssector moet voor het adopteren van LA heldere, veilige *use cases* ontwikkelen, waarin het belang van de student voorop staat. Het hoger onderwijs werkt, samen met onder andere SURF, stapsgewijs aan kennis en kunde over LA. Dat is een behoedzame werkwijze die is aan te moedigen. Het belang van studenten moet altijd voorop staan bij de ontwikkeling en het gebruik van LA. Daarom is het aan te raden dat onderwijsinstellingen studentenpopulaties actief te betrekken bij het ontwikkelen van uiteindelijke use cases. De eerste ervaringen uit het hoger onderwijs leren al dat studenten daarvoor openstaan en dat hun betrokkenheid de use cases naar een hoger plan tilt door ze voor alle betrokkenen nuttiger en betekenisvoller te maken.

Generatieve AI

Generatieve AI wordt inmiddels alomtegenwoordig en vraagt een duidelijk beleid van onderwijsinstellingen, vergelijkbaar met dat voor het gebruik van internetbronnen. Uit een kleine enquête van de Erasmus Universiteit blijkt dat 92 procent van de studenten ChatGPT gebruikt voor verschillende doeleinden.[42] Ook in het voortgezet onderwijs gebruiken leerlingen generatieve AI voor huiswerkopdrachten. Instellingen in het mbo en hoger onderwijs hebben echter nog amper beleid voor wat wel en niet mag met generatieve AI. Evidente risico's van generatieve AI voor de onderwijssector zijn ongeoorloofd gebruik ervan door leerlingen en studenten, zoals plagiëren, en misinformatie door foutieve maar plausibele output. Daarnaast is het momenteel nog de vraag of en hoe generatieve AI rechtmatig gebruikt kan worden. Mede vanwege dat rechtmatigheidsvraagstuk is het op dit moment verstandig voor onderwijsinstellingen zelf om terughoudend te zijn met de inzet van generatieve AI. Daardoor kan de sector zich concentreren op de beheersing van de inzet van generatieve AI door studenten. Dat heeft het onderwijs eerder succesvol gedaan bij het gebruik van internetbronnen, zoals Wikipedia.

Het onderwijs kan generatieve AI niet uitbannen en kan studenten en leerlingen beter leren er goed mee om te gaan. Het is te voorzien dat generatieve AI, net als Google en Wikipedia, alomtegenwoordig en dagelijks gebruikt zal worden. Een overweging voor het onderwijs is een tweesporenplan te volgen. Enerzijds moeten leerlingen en studenten leren omgaan met generatieve AI door het te gebruiken. Dat betekent dat het onderwijs toets- en lesmethoden nodig heeft die leerlingen en studenten bij een gecontroleerd gebruik van generatieve AI begeleiden. Dit moet wel gebeuren op een manier die de rechtmatigheidsvragen rondom foundation models en afgeleide toepassingen voldoende in acht neemt. Anderzijds moeten leerlingen en studenten generatieve AI kunnen beoordelen doordat ze zelf kunnen wat generatieve AI kan.

Beleidsaanpak

De AI-verordening stelt in bepaalde gevallen eisen aan profilerende en beoordelende algoritmetoepassingen in het onderwijs. De sector moet die eisen wel aanvullen. Dat betekent dat de systemen zelf aan strenge eisen moeten voldoen om op de Europese markt te worden toegelaten. Het duurt alleen nog even voordat de AI-verordening van toepassing is. Bovendien zijn de eisen in die verordening algemeen en niet onderwijs-specifiek en gelden ze niet voor alle actoren in de keten. Voor een verantwoorde inzet van algoritmetoepassingen moet (partners in) de onderwijssector dus ook zelf bepalen aan welke eisen systemen moeten voldoen en die eisen bij ontwikkelaars neerleggen.

Onderwijsinstellingen moeten hun organisatie voorbereiden op het integreren van algoritmetoepassingen in het onderwijs voorafgaand aan de aankoop ervan. Nieuwe regelgeving zoals de AI-verordening gaat namelijk nog niet over de wijze waarop gebruikers algoritmetoepassingen daadwerkelijk in hun organisaties inbedden. Een juiste inbedding is van groot belang voor het beperken van risico's. Docenten en leraren moeten voldoende algoritmegeletterd zijn om de uitkomsten te interpreteren van systemen waarmee ze werken. Onderwijsinstellingen moeten verantwoordelijkheden duidelijk beleggen en vooraf nadenken over de evaluatie van algoritmetoepassingen. Verder is het van belang dat onderwijsinstellingen mogelijk negatieve effecten van algoritmetoepassingen in kaart brengen en vooraf bepalen hoe die gedekt kunnen worden.

5. Beleid en regelgeving

Beleid en regelgeving voor algoritmes en AI vormt zich. Nieuwe wetgeving is in werking getreden of zal dat binnen afzienbare tijd doen. Er is een politiek akkoord bereikt over de AI verordening die daarmee in 2024 in werking kan treden. Daarmee is een belangrijke stap gezet die zal bijdragen aan de beheersing en controle van AI-systemen. De Digital Services Act (DSA) is een belangrijke Europese verordening die recentelijk al in werking is getreden en aanpak van algoritmerisico's eist van zeer grote platformen en zoekmachines. Er is ook een akkoord bereikt over de Platformwerkrichtlijn, die regels zal geven over algoritmische besluitvorming en monitoring van mensen die werken via een platform. Deze nieuwe wetgeving draagt in grote mate bij aan bescherming en houvast bij de beheersing van algoritmerisico's. Het is met deze en andere Europese wetgeving wel steeds zoeken naar de samenhang en hoe deze wetgeving zich tot elkaar verhoudt. Dat kan onduidelijkheid scheppen over regels en het toezicht daarop en daarmee ook de beheersing van algoritmerisico's.

AI-verordening

Er is een politiek akkoord over de AI-verordening. Publieke, private en toezichthoudende organisaties moeten daarom alert zijn in hun voorbereidingen. De verordening zal waarschijnlijk op zijn vroegst medio 2026 van toepassing zijn. Het verbod op bepaalde AI-systemen zal waarschijnlijk na 6 maanden (jaareinde 2024) van toepassing zijn; de bepalingen over generatieve AI-modellen en de governance waarschijnlijk na 1 jaar (medio 2025). Nu er een politiek akkoord ligt kan worden gesteld dat de structuur van de verordening niet wezenlijk anders is dan omschreven in een eerdere RAN (juli 2023). De AI-verordening zal dus met name de ontwikkeling van AI-systemen uitdrukkelijk gaan reguleren. Toch zijn er een aantal nieuwe onderdelen toegevoegd die naar waarschijnlijkheid impact zullen hebben op de effectiviteit van de verordening in het beheersen en controleren van algoritmerisico's. Deze zullen in de volgende alinea's worden toegelicht. Met het aanbreken van deze nieuwe fase is het van belang nogmaals te benadrukken dat de verordening alleen zal kunnen slagen als overheden, private partijen en toezichthouders werken aan duidelijke en adequate standaarden, een consistente interpretatie van

regelgeving en een effectieve structuur voor toezicht en samenwerking op het gebied van AI.

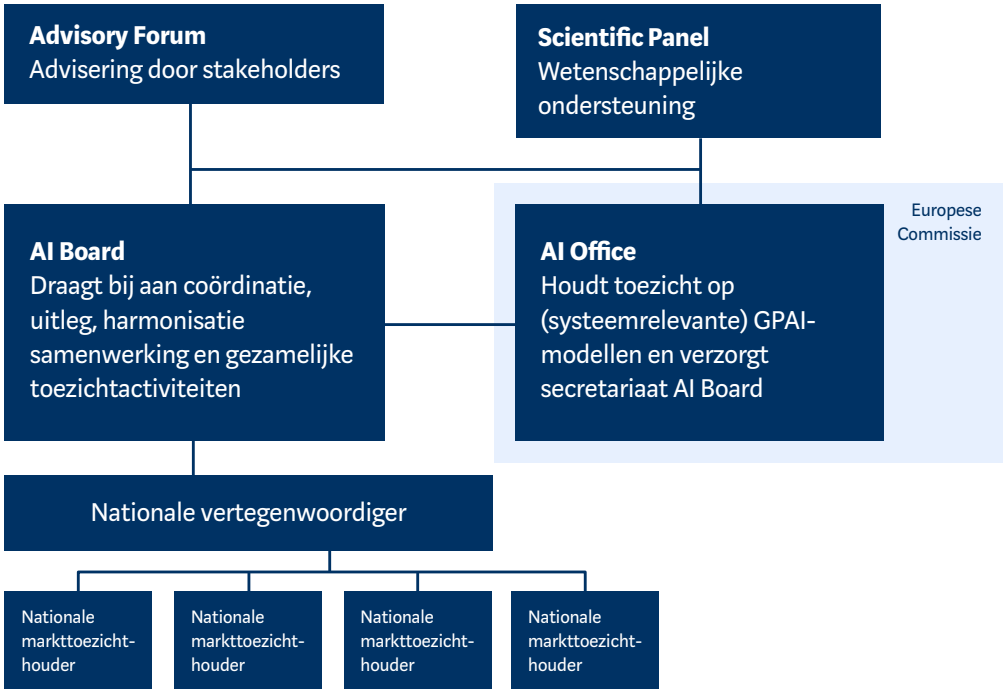
De *'carve out'* voor hoog risicosystemen die beperkt zijn tot voorbereidende of ondersteunende taken heeft het regelgevend kader van de AI-verordening minder eenduidig gemaakt. In de AI-verordening worden bepaalde systemen als 'hoog risico' geclassificeerd, zoals systemen die zijn bedoeld om beslissingen te nemen over de werving en selectie van personeel. Dergelijke AI-systemen mogen alleen op de markt komen als ze voldoen aan de eisen in de AI-verordening. De verordening zal niet van toepassing zijn op AI-systemen met voorbereidende of ondersteunende taken ('carve-out'). Met name voor AI-systemen die bedoeld zijn om een voorbereidende taak uit te voeren of die een aanvullende rol hebben bij de menselijke beoordeling, bijvoorbeeld om een beslissing te bevestigen. Deze carve-out is niet zonder risico's omdat het een wetmatigheid is dat mensen geneigd zijn teveel te vertrouwen op voorbereidende beoordelingen van AI-systemen, een principe dat bekend staat als '***overreliance***'. Ook als een AI-systeem slechts een voorbereidende of ondersteunende taak heeft kunnen de risico's op bijvoorbeeld discriminatie zich materialiseren, ondanks dat er mensen betrokken zijn in de besluitvorming. Daarnaast bestaat er het risico dat aanbieders hun systemen ten onrechte als 'voorbereidende' of 'ondersteunende' systemen in de markt zetten. Het is dan aan de markttoezichthouder daarop te handhaven.

De verordening geeft rechten aan mensen die in aanraking komen met (de uitwerkingen van) AI-systemen. Mensen zullen het recht hebben om een klacht in te dienen over AI-systemen. Ook krijgen zij het recht op uitleg van beslissingen die zijn gebaseerd op hoog risico AI-systemen, zoals systemen voor werving en selectie van personeel of voor in het onderwijs. Ook zullen organisaties die een AI-systeem gebruiken zelf in een grondrechtenimpactassessment (***fundamental rights impact assessment***, FRIA) moeten nagaan welke negatieve effecten dat systeem kan hebben en zullen zij maatregelen moeten nemen om die te voorkomen.

Ook overheidsgebruik van hoog risico AI-systemen moet worden geregistreerd in een Europese database. Het Europese AI-register wordt hierdoor een mix van een productregister (welke producten hebben een CE-markering en zijn toegelaten op de Europese interne markt?) en een gebruiksregisters dat op punten mogelijk gelijkenissen zal vertonen met het Algoritmeregister zoals dat wordt ontwikkeld door de Nederlandse overheid.

Europese problemen vragen om Europese oplossingen. AI-systemen kunnen in alle Europese samenlevingen impact hebben. En er moet slagvaardig opgetreden kunnen worden onder de AI-verordening. Daarom zal een deel van het toezicht Europees worden ingericht. Deze toezichtstructuur bestaat uit verschillende onderdelen. Allereerst wordt er een '***AI Board***' opgericht welke zal bestaan uit nationale vertegenwoordigers en een adviserende en coördinerende rol heeft. Het Board kan bijvoorbeeld opinies en aanbevelingen geven over welke AI systemen als hoog risico aangewezen

FIGUUR 10: EUROPESE GOVERNANCE TOEZICHT AI-VERORDENING (VERSIMPELD)



moeten worden, de standaarden en de ontwikkeling van het AI-landschap in Europa. Daarnaast komt er een '***Advisory forum***' en een '***Scientific panel***', waarbij laatstgenoemde ook een rol heeft bij het toezicht op general purpose AI. Het toezicht op general purpose AI wordt belegd bij een op te richten '***AI Office***' dat onderdeel zal zijn van Europese Commissie. Figuur 10 geeft een schematische weergave van de beoogde governancestructuur.

De AI-verordening zal voorzien in een gelaagd regime voor zogeheten general purpose AI modellen (foundation models). Er zullen daarbij horizontale verplichtingen gaan gelden voor alle general purpose AI-modellen. Daarbij zullen aanvullende eisen gaan gelden wanneer deze modellen systeemrisico met zich meebrengen. Hoofdstuk 2 gaat dieper in op foundation models en het toezicht daarop.

AI-standaarden

Normen ('*standards*') zullen invulling geven aan de eisen in de AI-verordening. De verordening zal algemene regels ('essentiële eisen') geven waaraan ontwikkelaars moeten voldoen bij de ontwikkeling van hun AI-systeem. Het gaat dan bijvoorbeeld om risicomanagement, de kwaliteit van datasets, transparantie, menselijk toezicht en cyberveiligheid. Normalisatie-organisaties CEN en CENELEC werken op verzoek van de Europese Commissie al aan normen die ontwikkelaars houvast moeten geven.[43] De NEN, de organisatie die in Nederland normen beheert, neemt deel aan deze organisaties en vertegenwoordigt de belangen van de Nederlandse markt en deelnemende organisaties. Ontwikkelaars zijn niet verplicht om zich aan de normen te houden. Maar als zij dat wel doen, dan wordt verondersteld dat hun AI-systemen conform de essentiële eisen in de AI-verordening zijn. In de praktijk zullen de normen dus een grote rol spelen in het aantonen van compliance en de beoordeling van conformiteit.

De ontwikkeling van AI-normen moet op evenwichtige en transparante wijze gebeuren, met inachtneming van alle belangen en met het oog op de bescherming van grondrechten. De Nederlandse overheid heeft hierin een bijzondere verantwoordelijkheid. Normen zijn flexibel, bieden duidelijkheid en worden mede opgesteld door organisaties met veel technologische kennis. Tegelijkertijd vindt de ontwikkeling van normen voor een groot deel plaats buiten het zicht van de samenleving en is het gevaar dat niet alle belangen goed worden vertegenwoordigd. Het is daarom van belang dat niet alleen bedrijven die AI ontwikkelen vertegenwoordigd zijn in organisaties zoals NEN, maar ook overheidsinstellingen en organisaties die staan voor de bescherming van grondrechten en de belangen van burgers. Gezien de publieke belangen die op het spel staan, moeten overheidsinstellingen openheid geven over hun eigen inzet en inbreng in de normering. Dat bevordert de legitimiteit van het overheidsoptreden en geeft meer inzicht in dit anderszins private en gesloten proces. De normen zelf moeten ook openbaar zijn. Ook moet de Europese Commissie scherp toezien op een eerlijke belangenafweging en de bescherming van grondrechten zodra de Commissie normen harmoniseert en daarmee in feite goedkeurt. Tot slot moeten markttoezichthouders in de praktijk checken of een geharmoniseerde norm in een concreet geval ook daadwerkelijk recht doet aan de eisen die de AI-verordening stelt, bijvoorbeeld eisen aan transparantie of het voorkomen van discriminatie.

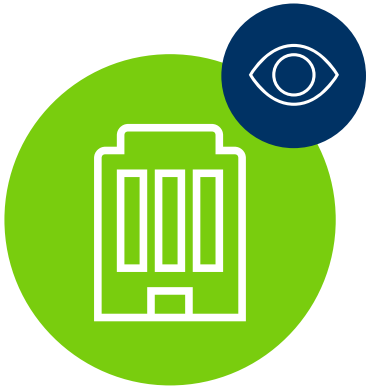
Digital Services Act

Naast de AI-verordening is er ook andere Europese (sectorspecifieke) regelgeving die inmiddels eisen stelt aan de manier waarop en algoritmes en AI worden ingezet. Een voorbeeld daarvan is de Digital Services Act (DSA). De DSA reguleert internetdiensten als hostingdiensten, socialemediadiensten, online marktplaatsen en zoekmachines, met het oog op de bescherming van gebruikers van online platforms, veiligheid en de verantwoordelijkheid van platforms voor hun diensten. De bepalingen in de DSA die gericht zijn op zeer grote platformen en zeer grote zoekmachines, zijn deels al van toepassing. De rest van de DSA wordt in februari 2024 van toepassing.

De DSA geeft regels voor hoe aanbieders van digitale diensten moeten optreden tegen illegale inhoud. Grote online platformen moeten bijvoorbeeld optreden tegen des- en misinformatiecampagnes. Zeker als de verspreiding van dergelijk materiaal risico's oplevert voor het verloop van verkiezingen. Een zorg hierbij is de toenemende rol van generatieve AI om, op heel grote schaal, goedkoop en snel, enorme hoeveelheden inhoud te produceren, die vervolgens via aanbieders van digitale diensten verspreid kan worden. Generatieve AI kan namelijk ook worden gebruikt voor betrouwbaar of realistisch ogende mis- en desinformatie. Tegelijkertijd kunnen aanbieders van digitale diensten zelf ook AI-systemen en algoritmes gebruiken om gegenereerde inhoud te identificeren en soms ook te verwijderen. De bestrijding van mis- en desinformatie wordt zo als het ware een wapenwedloop van AI-systemen.

Aan het vinden en verwijderen van illegale inhoud en mis- en desinformatie met hulp van AI-systemen kleven ook risico's. Het onterecht verwijderen van inhoud door een algoritme levert spanning op met fundamentele rechten, zoals de vrijheid van meningsuiting en gelijke behandeling. De kans bestaat dat verwijderde inhoud helemaal niet illegaal is of niet als des- of misinformatie kan worden bestempeld. Daarbij kan een rol spelen dat aanbieders van digitale diensten het zekere voor het onzekere nemen en doorschieten in het verwijderen van online inhoud. Ook het gebrek aan menselijke controle is problematisch. Uit de transparantierapportage van X (voorheen: Twitter) uit 2023 is af te leiden dat in het Content Moderation Team maar één medewerker de Nederlandse taal machtig is.[44] Het onterecht verwijderen van legitieme inhoud met algoritmes en AI is dus een risico dat serieus moet worden genomen.

De Europese Commissie is de toezichthouder op zeer grote platformen en zoekmachines en op de bijzondere eisen die voor deze dienstverleners gelden. Deze dienstverleners hebben in de DSA een bijzondere verantwoordelijkheid om maatregelen te nemen tegen risico's die hun diensten met zich meebrengen. Meerdere platformen, zoals X, TikTok en Meta, hebben op grond van de DSA van de Europese Commissie al informatieverzoeken ontvangen over de verspreiding van ongewenste inhoud op deze platformen. De Commissie zet hiermee een eerste stap om zeer grote platformen te houden aan de DSA, zodat zij maatregelen nemen tegen de verspreiding van desinformatie via aanbevelingsalgoritmes op hun platforms. De DSA is daarmee een waardevol instrument om specifieke risico's van algoritme- en AI-gebruik inzichtelijk te maken en te beperken. Ook stelt de DSA samenlevingen in staat om meer grip te krijgen op door algoritmes gedreven verspreiding van des- en misinformatie.



Toezicht en controle in Nederland

Algoritmes en AI zijn een systeemtechnologie en raken verweven door alle delen van de samenleving, dit geldt ook voor toezicht en controle. De verwachting is dat vrijwel alle toezichthoudende organisaties met algoritmes en AI in aanraking zullen komen bij het uitoefenen van toezicht of zelfs bij de uitvoering van hun taken. Algoritmes worden altijd in een specifieke context toegepast. Daarom is de kennis die sectorale toezichthouders van die context hebben van belang. Anderzijds moet er worden gewaakt voor gefragmenteerd toezicht. Samenwerking tussen toezichthouders is cruciaal om goed functionerend en samenhangend toezicht te organiseren op de ontwikkeling en inzet van algoritmes.

Nederland kent een groot aantal toezichthoudende en controlerende organisaties die toezien op deelaspecten van de ontwikkeling en inzet van algoritmes en AI, of deze controleren. Deze organisaties houden toezicht op algoritmes en AI, controleren de werking en rechtmatigheid ervan of behandelen klachten over publieke en private organisaties die algoritmes en AI inzetten. Colleges van staat, markttoezichthouders en inspecties voeren een groot deel van deze taken uit. Daarnaast vervullen de Nationale Ombudsman en lokale ombudspersonen een belangrijke rol. Ook zijn er andere publieke organisaties en private organisaties die een controlerende taak hebben, bijvoorbeeld bij specifieke certificering of kwaliteitstoezicht. Tot slot kunnen burgers en organisaties ook terecht bij de rechter om hun rechten af te dwingen.

Toezicht op algoritmes en AI vindt plaats op basis van uiteenlopende bestaande en nieuwe kaders. Toezichthoudende organisaties baseren zich voor hun activiteiten op verschillende (juridische) kaders en verschillen in onafhankelijkheid of positionering. Internationale verdragen, Europese verordeningen en richtlijnen, maar ook nationale wetgeving, vormen een stevige basis voor een groot deel van de toezichthoudende en controlerende organisaties. Deze wettelijke kaders bevatten alleen niet vaak bepalingen die zich puur richten op algoritmes en AI zelf en de ontwikkeling en inzet daarvan, het toezicht daarop of de controle daarvan. Toezicht en controle kunnen, naast de wetgeving, worden aangevuld en versterkt door (open) normen en standaarden. Enerzijds biedt dit soms meer flexibiliteit in veranderende omstandigheden. Anderzijds bieden deze normen en standaarden minder houvast voor ingrijpen en sanctioneren bij overtredingen.

Gelaagd toezicht is nodig voor toepassing van systeemtechnologie. Toezicht en controle zijn niet altijd vorm gegeven als een directe, een-op-een interactie tussen een toezichthoudende of controlerende organisatie en een organisatie die onder toezicht staat (de 'ondertoezichtgestelde'). Wanneer het aantal ondertoezichtgestelden beperkt is, kan direct toezicht of controle op de producten, diensten of pro-

cessen van een organisatie plaatsvinden. Bij een omvangrijk aantal ondertoezichtgestelden is het echter vaak niet meer mogelijk om op alles direct toezicht te houden of controles uit te voeren. Dan kan het effectiever zijn om gebruik te maken van gelaagd toezicht. Er kan bijvoorbeeld intern toezicht bij ondertoezichtgestelden worden georganiseerd, zoals door een functionaris gegevensbescherming (FG) aan te stellen. Zo'n interne toezichthouder kan vroegtijdig problemen signaleren en adviezen geven. En zo als strategisch adviseur een organisatie en burgers of consumenten beschermen tegen overtredingen of ongewenste risico's en effecten. Een ander voorbeeld van een 'toezichtslaag' is routinematige procedures of standaardproducten en -diensten certificeren en daarop laten toezien door een daarvoor aangewezen derde instantie, zoals een accountantskantoor.

Algoritmes en AI vragen om een interne structuur van toezicht en controle die verantwoorde technologische innovatie mogelijk maakt. Algoritmes en AI worden op steeds grotere schaal ingezet in steeds meer sectoren van de samenleving. Dit maakt direct toezicht op alle ontwikkeling en inzet onmogelijk. Juist voor algoritmes is gelaagd toezicht passender, omdat de risico's, de verhoudingen tussen actoren en hun belangen sterk verschillen en afhankelijk zijn van de specifieke inzet en context. Van de werkvloer tot aan het bestuur moet dus aandacht worden besteed aan het opsporen, verminderen en controleren van risico's bij de ontwikkeling en inzet van algoritmes. Bij deze opgave past het introduceren van interne posities als een algoritmefunctionaris, het versterken van kennis en expertise of het aanstellen van een ***AI governance board***. Dit is een positieve ontwikkeling in de beheersbaarheid van algoritmes.

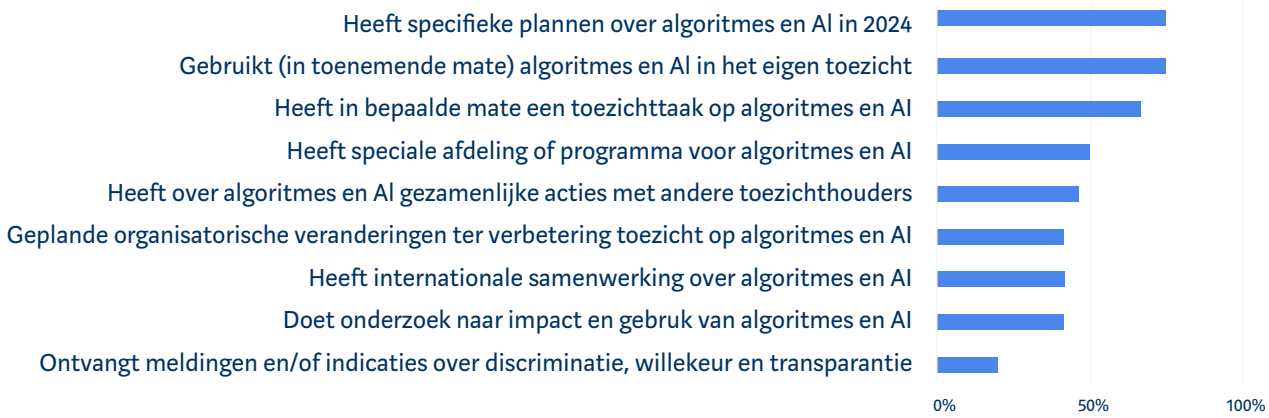
Burgers en organisaties weten niet altijd bij welke toezichthoudende of controlerende organisatie zij terechtkunnen. Voor burgers en organisaties zijn toezichthoudende en controlerende organisaties een belangrijke waarborg tegen inbreuken, overtredingen, risico's of ongewenste effecten en het zich verweren daartegen. Zijn er signalen dat normen overtreden worden? Of wordt een burger geconfronteerd met risico's of ongewenste effecten van algoritmes? Dan kan deze persoon daarvan melding maken bij een aantal loketten van toezichthouders of ombudspersonen. Een probleem hierbij is wel dat door het aanzienlijke aantal toezichthoudende organisaties in Nederland er onder veel burgers onduidelijkheid is over de toezichthoudende organisatie die zij moeten aanspreken. Het aanspreken van ***individuele*** toezichthouders is meestal de enige optie voor mensen om actie te ondernemen als zij een klacht hebben of een melding willen doen. De Rijksoverheid biedt namelijk beperkt overzicht van toezichthoudende of controlerende organisaties waarbij burgers effectief hulp zouden kunnen zoeken. Het ontbreken van informatie en overzicht kan leiden tot verminderde meldingsbereidheid en afbreuk van vertrouwen. Dit geldt des te meer als het een mogelijke misstand bij een overheidsorganisatie betreft. Een richtinggevend overzicht van toezichthoudende en controlerende organisaties, met bijbehorende loketten, taken en aandachtsgebieden, draagt bij om de toegang tot toezicht en correctief handelen voor iedereen te kunnen waarborgen. Het kunnen verweren tegen misstanden en onrechtmatigheden bij de inzet van algoritmes en AI is ook vanuit mensenrechtenperspectief een verplichting.

Het is van belang om te investeren in samenwerking tussen loketten om algoritmische misstanden aan het licht te krijgen. Burgers met klachten over algoritmes en AI moeten bij alle relevante loketten zonder belemmeringen terechtkunnen. Hoewel burgers momenteel nog niet in groten getale klachten indienen over door hen ervaren negatieve effecten van algoritmes en AI, moeten zij wel over deze mogelijkheid beschikken. Dat het nu nog niet veel gebeurt, komt vaak door een gebrek aan transparantie over de inzet van algoritmes. Daardoor zijn burgers zich er niet van bewust dat er een algoritme in het spel is. Ook kan het voor burgers onduidelijk zijn dat de oorzaak van een probleem eigenlijk in een algoritmisch systeem ligt. De DCA gaat in 2024 aan de slag met collega-organisaties die een loketfunctie hebben om de samenwerking tussen deze loketten te versterken. Het doel is algoritmeproblematiek eerder te herkennen en burgers en organisaties zo optimaal te kunnen helpen.

Inzicht in het toezicht op algoritmes en AI in Nederland

Om tot effectief toezicht op algoritmes en AI te komen, is het van groot belang om inzichtelijk te krijgen hoe toezicht op algoritmes momenteel wordt opgepakt door toezichthouders zelf. In de zomer van 2023 is een survey over algoritmegebruik en- risico's onder 33 toezichthouders uitgezet. Hiervan hebben 24 toezichthouders – met toezichtsbevoegdheden op landelijk niveau – de survey ingevuld. De uitkomsten geven inzicht in de staat van het algoritmelandschap in Nederland. Daarnaast komt er een terugkoppeling aan de toezichthouders, zodat zij de inzichten kunnen meenemen in hun strategie en focus. De uitkomsten van de survey laten zien dat een meerderheid van de toezichthouders zich bewust is van de risico's en kansen van algoritmes en AI, maar dat het per toezichthouder verschilt hoeveel algoritmes en AI onderdeel zijn van het dagelijkse werk. Figuur 11 geeft een overzicht van de survey-uitkomsten.

FIGUUR 11: HOE KOMEN ALGORITMES EN AI TERUG IN HET WERK VAN TOEZICHTHOUDERS?
UITKOMSTEN VAN EEN ENQUÊTE ONDER NEDERLANDSE TOEZICHTHOUDERS



De enquête is medio 2023 door Nederlandse toezichthouders ingevuld. De AP heeft een respons ontvangen van 24 toezichthouders met toezichtbevoegdheden op landelijk niveau, effectief alle aangeschreven toezichthouders.

Twee derde van de toezichthouders in de survey houdt in enige mate toezicht op algoritmes en AI. De impact van het gebruik van algoritmes en AI-systemen en/of -toepassingen verschilt per sector en daarmee ook per toezichthouder. Twee derde van de toezichthouders geeft aan in enige mate toezicht te houden op algoritmes. Een klein deel houdt structureel toezicht op het gebruik van algoritmes. Andere toezichthouders zijn voornamelijk extra alert op mogelijke risico's. Het deel van de toezichthouders dat geen toezicht houdt op algoritmes, geeft aan dat het op dit moment niet (genoeg) speelt in hun sector. Of dat expertise op dit vlak binnen hun organisatie ontbreekt.

Ruim 40% van de toezichthouders in deze survey doet onderzoek naar de impact en het gebruik van algoritmes en AI. Dit varieert van verkenningen, het verstrekken van gerichte informatie en onderzoeken naar potentiële algoritme-gerelateerde misstanden tot overtredingen waarbij een algoritme direct een rol speelt. Deze onderzoeken resulteren in guidance voor organisaties en burgers, verduidelijking en uitleg van normen, adviezen, waarschuwingen en sancties.

Vier toezichthouders in deze survey hebben concrete klachten of signalen van burgers ontvangen over de inzet van algoritmes en AI. Een oorzaak hiervan kan zijn dat burgers niet weten dat er een algoritme in het spel is. Mede om deze reden is het verbeteren van transparantie bij de inzet van algoritmes een van de kernthema's van de DCA. Daarnaast zijn de toegankelijkheid en samenwerking van toezichthouders en loketten van belang. Hierin zal de DCA in 2024 verder investeren.

De ondervraagde toezichthouders onderstrepen dat onderlinge samenwerking nodig is. De toename van het gebruik van algoritmes in de sectoren vraagt om meer expertise en samenwerking. Bijna de helft van de ondervraagde toezichthouders heeft contact met andere toezichthouders over systemen en/of toepassingen in het werkveld. En 40% van de ondervraagde toezichthouders heeft weleens contact over dit thema met internationale partners, vaak in al bestaande samenwerkingsstructuren. Een van de pijlers van de DCA is om de samenwerking tussen toezichthouders op dit vlak verder te versterken. Zo organiseert de AP workshops en kennissessies en kijkt de AP op basis van deze survey waaraan toezichthouders nog meer behoefte hebben.

De meerderheid van de toezichthouders heeft ambities en plannen om in 2024 meer te doen met algoritmes. Een aantal toezichthouders geeft aan te willen starten met (onderzoeks)pilots om de kansen en risico's van het gebruik van algoritmes (verder) te verkennen. Er is momenteel ook veel aandacht voor nieuwe wetgeving, zoals de AI-verordening en de impact hiervan op het werk van toezichthouders. Als reactie hierop krijgen medewerkers steeds vaker opleidingen en cursussen aangeboden om up to date te blijven over de ontwikkelingen van algoritmes en AI. Ook wordt er bij een deel van de toezichthouders nieuwe mensen geworven voor het toezicht op algoritmes en AI.

Toezichthouders zetten voor eigen werkzaamheden ook algoritmes en AI in. Driekwart van de toezichthouders in deze survey is bezig met (het uitbreiden van) de inzet van algoritmes om gerichter en effectiever toezicht te kunnen houden. Denk hierbij bijvoorbeeld aan systematische controles waarbij minder menselijke capaciteit nodig is of het detecteren van risico's en afwijkingen in een sector. Ook kunnen specifieke trends in sectoren sneller worden verzameld. Dit laat zien dat de wereld van de toezichthouders niet verschillend is van de rest van de samenleving, waar deze trends momenteel ook

gaande zijn. Uiteraard moeten ook toezichthouders investeren in verantwoorde inzet en adequate beheersing van algoritmes. Transparantie maakt hier onderdeel van uit, bijvoorbeeld door registratie van algoritmes in het Algoritmeregister.

Regulatory sandboxes

Voor de verantwoorde en veilige ontwikkeling van algoritmes en AI kunnen regulatory sandboxes worden gebruikt. Dat zijn omgevingen waarin nieuwe werkwijzen of technologieën onder toeziend oog van een toezichthouder kunnen worden uitgeprobeerd. Naar gelang het doel ervan, kan een regulatory sandbox verschillende vormen aannemen. In zo'n sandbox kan een toezichthouder bijvoorbeeld meekijken om snel te kunnen ingrijpen als risico's zich manifesteren. Een functie van de sandbox kan zijn om binnen de bestaande wet- en regelgeving algoritmen te ontwikkelen, en partijen guidance te geven en te begeleiden bij het voldoen aan de regels. Het doel van een sandbox is dan om ervoor te zorgen dat innovatie binnen de wettelijke kaders plaatsvindt. En om daarbij bijvoorbeeld lastige juridische vragen te identificeren en te beantwoorden, compliance te bevorderen en de drempel tot markttoegang te verlagen. De rol van toezichthouders is dan vooral om richting te geven aan ontwikkelaars. Opgedane kennis kan dan met de markt gedeeld worden, zodat er voor alle relevante marktpartijen meer duidelijkheid ontstaat over hoe te voldoen aan (doorgaans) nieuwe regelgeving. Bovendien kunnen wet- en regelgeving, richtsnoeren en (toezichts)beleid worden toegesneden op nieuwe ontwikkelingen en in een sandbox opgedane inzichten.

In Nederland wordt een regulatory sandbox voor de AI-verordening voorbereid. Op grond van deze verordening zullen EU-landen een sandbox moeten aanbieden. Hierin begeleiden en adviseren markttoezichthouders ontwikkelaars van AI-systemen, zodat zij hun systemen compliant kunnen maken. In Nederland verkennen toezichthouders samen met het ministerie van Economische Zaken en Klimaat (EZK) hoe zij de Nederlandse sandbox voor de AI-verordening kunnen inrichten. Uiteindelijk zullen er ook op Europees niveau nadere regels gesteld worden voor de werking van de sandboxes en de toegang daartoe door ontwikkelaars.

[1] OECD. (2023). OECD AI Incidents Monitor (AIM). OECD.AI Policy Observatory. https://oecd.ai/en/incidents

[2] Young, S. (2023). Proposed memorandum fort the heads of executive departments and agencies. [Memorandum]. Executive Office of the President – Office of Management and Budget. https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf

[3] College voor de Rechten van de Mens. (17 november 2023). 2023-111. https://oordelen.mensenrechten.nl/oordeel/2023-111

[4] College voor de Rechten van de Mens. (1 augustus 2023). 2023-82. https://oordelen.mensenrechten.nl/oordeel/2023-82

[5] College voor de Rechten van de Mens. (17 oktober 2023). Student niet gediscrimineerd door tentamensoftware Proctorio, maar VU had de klacht zorgvuldiger moeten behandelen [nieuwsbericht]. https://www.mensenrechten.nl/actueel/nieuws/2023/10/17/student-niet-gediscrimineerd-door-tentamensoftware-proctorio-maar-vu-had-de-klacht-zorgvuldiger-moeten-behandelen

[6] Rekenkamer Metropoolregio Amsterdam. (2023) Algoritmen https://www.rekenkamer.amsterdam.nl/onderzoek/algoritmen/

[7] Algorithm Audit. (2023) Adviesdocument Risicoprofilering heronderzoek bijstandsuitkering (AA:2023:02:A) https://algorithmaudit.eu/nl/algoprudence/

[8] Autoriteit Consument en Markt. (27 juni 2023). ACM spreekt webshops aan die misleidende countdowntimers gebruiken [nieuwsbericht]. https://www.acm.nl/nl/publicaties/acm-spreekt-webshops-aan-die-misleidende-countdowntimers-gebruiken

[9] Auditdienst Rijk. (2023). Onderzoekskader Algoritmes. https://open.overheid.nl/documenten/61b54381-d331-40ed-8fce-b2883b195f25/file

[10] Deloitte. (2023). Digital Consumer Trends Report 2023. https://www2.deloitte.com/nl/nl/pages/technologie-media-telecom/articles/digital-consumer-trends-23-report.html

[11] Consultancy.nl. (8 november 2023). Nederlandse consument staat niet zo positief tegenover generatieve AI [nieuwsbericht]. https://www.consultancy.nl/nieuws/49601/nederlandse-consument-staat-niet-zo-positief-tegenover-generatieve-ai

[12] Capgemini Research Institute. (2023). Why consumers love generative AI. https://prod.ucwe.capgemini.com/wp-content/uploads/2023/05/Final-Web-Version-Report-Creative-Gen-AI.pdf

[13] Salesforce. (24 juli 2023). Nieuw Salesforce-onderzoek onthult toenemende druk op IT-teams en een prominente rol voor generatieve AI [nieuwsbericht]. Emerce. https://www.emerce.nl/wire/nieuw-salesforceonderzoek-onthult-toenemende-druk-itteams-prominente-rol-generatieve-ai

[14] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., … Liang, P. (2021). On the Opportunities and Risks of Foundation Models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

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

[16] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., … Liang, P. (2021). On the Opportunities and Risks of Foundation Models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

[17] Gensler, G. (17 juli 2023). "Isaac Newton to AI" Remarks before the National Press Club [Speech]. U.S. Securities and Exchange Commission. https://www.sec.gov/news/speech/gensler-isaac-newton-ai-remarks-07-17-2023

[18] Toups, C., Bommasani, R., Creel, K.A., Bana, S.H., Jurafsky, D., Liang, P. (2023). Ecosystem-level Analysis of Deployed Machine Learning Reveals Homogeneous Outcomes. arXiv. https://doi.org/10.48550/arXiv.2307.05862

[19] Competition & Markets Authority. (2023). AI Foundation Models: Initial Report. https://www.gov.uk/government/publications/ai-foundation-models-initial-report

[20] OECD. (2023). Employment Outlook 2023: Artificial intelligence and jobs. https://oecd.org/employment-outlook/2023/

[21] Jongen, E., van den Berge, W., Goos, M., Kerkemezos, Y. (2023). Technologie, de arbeidsmarkt en de rol van beleid. Centraal Planbureau. https://www.cpb.nl/sites/default/files/omnidownload/CPB-Publicatie-Technologie-de-arbeidsmarkt-en-de-rol-van-beleid.pdf

[22] Klijs, B., Smits, W. (2023). Werkgelegenheid. In Monitor online platformen 2022 (hfst.4). Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/longread/rapportages/2023/monitor-online-platformen-2022/4-werkgelegenheid

[23] Vervliet, T. (2022). Digitale werving & selectie: algoritmegebruik in het werving- & selectieproces van werkgevers en het bewustzijn van risico's op uitsluiting en discriminatie (Rapport 2022-62). SEO Economisch Onderzoek. https://publicaties.mensenrechten.nl/publicatie/546c77bc-4db8-407e-ae52-41ea26510c68

[24] https://ec.europa.eu/commission/presscorner/detail/en/speech_23_5366

[25] Riso, S., Adăscăliței, D., Rodriguez Contreras, R. (2023). Ethical digitalisation at work: From theory to practice. Publications Office of the European Union. https://www.eurofound.europa.eu/en/publications/2023/ethical-digitalisation-work-theory-practice

[26] Baiocco, S., Fernandez-Macías, E., Rani, U., Pesole, A. (2022). The Algorithmic Management of work and its implications in different contexts. International Labour Organization, European Commission. https://www.ilo.org/wcmsp5/groups/public/---ed_emp/documents/publication/wcms_849220.pdf

[27] von der Leyen, U. (13 september 2023). 2023 State of the Union Address by President von der Leyen [Speech]. https://ec.europa.eu/commission/presscorner/detail/en/speech_23_4426

[28] Harris, L.A., Jaikaran, C. (2023). Highlights of the 2023 Executive Order on Artificial Intelligence for Congress (R47843). Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R47843

[29] Smith, C. (10 februari 2022). Using AI-Based Tools to Build Empathy in the Contact Center [Blog]. Genesys. https://www.genesys.com/blog/post/using-ai-based-tools-to-build-empathy-in-the-contact-center

[30] OECD. (2023). Employment Outlook 2023: Artificial intelligence and jobs. https://oecd.org/employment-outlook/2023/

[31] Jongen, E., van den Berge, W., Goos, M., Kerkemezos, Y. (2023). Technologie, de arbeidsmarkt en de rol van beleid. Centraal Planbureau. https://www.cpb.nl/sites/default/files/omnidownload/CPB-Publicatie-Technologie-de-arbeidsmarkt-en-de-rol-van-beleid.pdf

[32] Inspectie van het Onderwijs. (z.d.). Maatregelen van so-scholen voor het aanpassen van het onderwijs tijdens de coronacrisis. https://www.onderwijsinspectie.nl/onderwerpen/corona-onderzoeken/gevolgen-van-16-maanden-corona-voor-het-onderwijs/gevolgen-speciaal-onderwijs/maatregelen-scholen-voor-aanpassen-van-onderwijs

[33] Nationaal Groeifonds. (z.d.) Nationaal Onderwijslab. https://www.nationaalgroeifonds.nl/overzicht-lopende-projecten/thema-onderwijs/nationaal-onderwijslab

[34] Radboud Universiteit. (z.d.). Over het Nationaal Onderwijslab AI. https://www.ru.nl/nolai/over-nolai

[35] Moerland, S., Elings, M. (21 oktober 2021). Extra geld voor basisonderwijs vooral gebruikt voor aantrekken leerkracht. NOS Nieuws. https://nos.nl/artikel/2402556-extra-geld-voor-basisonderwijs-vooral-gebruikt-voor-aantrekken-leerkracht

[36] Karssen, M., Krepel, A., Stronkhorst, E., Lourens, J.M.P., Bruck, S., Van Kessel, M. & Saab, N. (2023). Monitor digitalisering primair en voortgezet onderwijs (rapport 1102). Kohnstamm Instituut. https://kohnstamminstituut.nl/rapport/monitor-digitalisering-primair-en-voortgezet-onderwijs/

[37] Scharwächter, V. (2023). ChatGPT: Overzicht AI-richtlijnen per onderwijsinstelling. In Slimmer onderwijs met AI, September 2023, 62-63. NPuls. https://npuls.nl/actueel/npuls-presenteert-het-magazine-slimmer-onderwijs-met-ai/

[38] Stokke, L., Gallagher, E. (29 maart 2023). ChatGPT is er, deal ermee. Ukrant. https://ukrant.nl/magazine/chatgpt-is-er-dus-deal-ermee-studenten-willen-regels-voor-ai-gebruik-blijkt-uit-enquete-ukrant/#:~:text=Toch%20lijken%20studenten%20nog%20zo,complete%20opdrachten%20hebben%20laten%20maken

[39] Onderwijsraad. (2022). Inzet van Intelligente Technologie https://www.onderwijsraad.nl/publicaties/adviezen/2022/09/28/inzet-van-intelligente-technologie

[40] Dijkstra, P. (14 juni 2022). Experimenteeromgeving learning analytics: aan de slag met studiedata met oog voor privacy. SURF Communities. https://communities.surf.nl/learning-analytics/artikel/experimenteeromgeving-learning-analytics-aan-de-slag-met-studiedata-met

[41] Universiteit Utrecht. (z.d.). LA in het onderwijs. https://www.uu.nl/onderwijs/learning-analytics/la-in-het-onderwijs

[42] Smaling, E., (16 november 2023). Bijna iedereen gebruikt ChatGPT, maar plagiëren gaat studenten te ver. Erasmus Magazine. https://www.erasmusmagazine.nl/2023/11/16/bijna-iedereen-gebruikt-chatgpt-maar-plagieren-gaat-studenten-te-ver/#:~:text=Van%20de%20154%20deelnemers%20aan,te%20zetten%20tijdens%20hun%20opleiding

[43] COMMISSION IMPLEMENTING DECISION 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, C(2023)3215. (22 mei 2023). Directoraat-generaal Communicatienetwerken, Inhoud en Technologie. https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=nl

[44] X. (2023). DSA Transparency Report. https://transparency.twitter.com/dsa-transparency-report.html

AUTORITEIT
PERSOONSGEGEVENS