WHITEPAPER

OP ZOEK NAAR DE MENS IN AI

BETREK DE BURGER EN EXPERIMENTEER OP VERANTWOORDE WIJZE

TNO innovation for life

INHOUDSOPGAVE

SAMENVATTING

Kunstmatige intelligentie (‘artificial intelligence’, AI) heeft de afgelopen jaren tot vele innovaties geleid. Daarom denken ook overheidsorganisaties na over hoe zij AI kunnen gebruiken voor maatschappelijke vraagstukken of het uitvoeren van overheidstaken. Zo kan AI worden ingezet voor algoritmische besluitvorming. Aan deze toepassing van AI kleven echter ook nadelen, bijvoorbeeld het risico op discriminatie. Om te voorkomen dat de inzet van AI dergelijke nadelige effecten heeft, richt Europees en nationaal beleid zich onder andere op het opstellen van normenkaders voor de ontwikkeling van AI en op regulering van AI toepassingen. In aanvulling daarop ontwikkelt TNO een methodiek voor het beproeven van AI-systemen voor algoritmische besluitvorming in een proeftuin: de *dynamische Impact Assessment*. Aan de hand daarvan wordt het mogelijk om effecten van de toepassing dynamisch, op korte en langere termijn, inzichtelijk te maken en om de belangen van verschillende stakeholders, waaronder burgers, mee te nemen.

1. INLEIDING

De verwachte innnovatiekracht van ‘artificial intelligence’ (AI) is groot. De Nederlandse overheid stelt in haar *Strategisch Actieplan voor AI* dat AI ‘stevig [zal] bijdragen aan economische groei, welvaart en welzijn van Nederland’.[1] Hoewel AI als technologie al sinds de jaren vijftig van de vorige eeuw bestaat, heeft de enorme toename aan data en rekenkracht voor een ongekende opleving gezorgd. Het gebruik van AI leidt naar verwachting tot doorbraken in medisch onderzoek en tot grotere verkeersveiligheid door (deels) zelfrijdende voertuigen. Geïnspireerd door deze voorbeelden wordt AI ook door overheden steeds vaker toegepast bij het verkennen van maatschappelijke vraagstukken en het maken van besluiten. Bijvoorbeeld voor de snellere afhandeling van visumaanvragen, om beter inzicht te krijgen in armoede- of schuldenproblematiek of voor onderhoud aan bruggen en sluizen. Dit wordt *algoritmische besluitvorming* genoemd, waarbij het (AI-)algoritme geheel of gedeeltelijk geautomatiseerd de uitkomst bepaalt.[2]

In deze voorbeelden heeft de technologie direct of indirect invloed op het leven van mensen, bijvoorbeeld wanneer zij in aanmerking komen voor schuldhulpverlening of wanneer zij een boete krijgen opgelegd. Dit leidt soms tot onbedoelde nadelige gevolgen (zie kader).[3]

1 Strategisch Actieplan voor AI (2019), Kamerbrief met Strategisch Actieplan voor Artificiële Intelligentie | Kamerstuk | Rijksoverheid.nl. Zie ook AI Strategies and Policies in Netherlands - OECD.AI.

2 Dit position paper richt zich op het gebruik van AI voor algoritmische besluitvorming in de publieke sector en dus bijvoorbeeld niet op de toepassing van AI in autonome systemen zoals zelfrijdende voertuigen of robots. Er wordt veel onderzoek gedaan, ook door TNO, naar de veiligheid van dergelijke AI systemen en hoe je dergelijke systemen op dat niveau ‘norm-compliant’ krijgt. Zie bijvoorbeeld: Aliman, N. M., Kester, L., Werkhoven, P., & Ziesche, S. (2019). Sustainable AI safety? Delphi, 2, 226.

3 Bronnen: Guide to AS and A level results for England, 2020 - GOV.UK (www.gov.uk); Why did the A-level algorithm say no? - BBC News.

Gebruik van een algoritme voor het bepalen van eindexamencijfers
In het Verenigd Koninkrijk zijn eindexamencijfers ('A-level'-scores) bepalend voor de kans dat scholieren worden toegelaten tot de universiteit van hun keuze. Toen vanwege de Covid-19-pandemie de eindexamens niet doorgingen is een algoritme gebruikt om de scores te voorspellen. Het algoritme maakte daarbij gebruik van een combinatie van individuele prestaties en de gemiddelde schoolresultaten van het voorgaande jaar. Dit leidde ertoe dat getalenteerde scholieren van slechter presterende scholen werden benadeeld en scholieren van private scholen – die vaak kleiner zijn en meer aandacht kunnen bieden aan individuele leerlingen – bevoordeeld. Het gebruik van het algoritme voor het bepalen van de A-level-scores heeft dan ook flinke kritiek gekregen en heeft geleid tot protesten tegen het systeem.

Niet alleen in het Verenigd Koninkrijk heeft de toepassing van algoritmen door de overheid tot maatschappelijke verontwaardiging geleid. De vraag is dan ook hoe ongewenste gevolgen voorkomen kunnen worden. De rol van de overheid is hierin tweeledig. Ten eerste past zij zelf algoritmen toe en moet dit op een verantwoordelijke manier doen. Ten tweede kijkt de overheid welke bestaande wetgeving hiervoor kan zorgen en wordt er, waar nodig, nieuw beleid gemaakt. Een belangrijk onderdeel van bestaande wetgeving is de Algemene Verordening Gegevensbescherming (AVG), gericht op het waarborgen van (data)privacy. Daarnaast worden er normenraamwerken voor de ontwikkeling van AI opgesteld. Hierin zijn principes voor verantwoordelijke toepassing van AI geformuleerd, gebaseerd op mensenrechten.[4] Een voorbeeld van een normenkader is de set van zeven voorwaarden voor verantwoordelijke AI van de Europese High-level Expert Group voor AI (AI HLEG).[5] Op basis van o.a. dit normenkader heeft de Europese Commissie recentelijk een voorstel voor AI regulering gepubliceerd.[6]

De voorgestelde regulering stelt voor om hoog-risicotoepassingen te reguleren door ze voorafgaand aan de toepassing aan te laten tonen dat ze voldoen aan verplichtingen ten aanzien van bijvoorbeeld menselijke controle. Er wordt daarbij geen onderscheid gemaakt tussen toepassingen in de private of publieke sector. Om voldoende ruimte te geven voor innovaties met AI, wordt er daarnaast aandacht besteed aan het belang van experimenteren. Op basis van een overzicht van hoe AI op dit moment wordt gereguleerd en een analyse van bestaande normenkaders en methodieken, introduceert dit paper daarom een methodiek die zich richt op het beproeven van de effecten van AI: een *dynamische impact assessment* voor AI-systemen gericht op maatschappelijke vraagstukken.

De opbouw van het paper is als volgt: ten eerste worden de toepassing van AI in algoritmische besluitvorming en de bijbehorende risico's beschreven. Daarna volgt een overzicht van het huidige beleid dat toeziet op AI, gevolgd door een analyse van normenkaders en methodieken om waardengebaseerde technologie te ontwikkelen. Ten slotte wordt de methodiek geïntroduceerd voor het beproeven van verantwoorde AI-systemen voor algoritmische besluitvorming.

4 Kamerbrief over artificiële intelligentie, publieke waarden en mensenrechten | Kamerstuk | Rijksoverheid.nl
5 Europese Commissie High-level Expert Group on AI, Shaping Europe's digital future (europa.eu).
6 Europese Commissie (2021), Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future (europa.eu).

2. ALGORITMISCHE BESLUITVORMING

AI sinds de opkomst van de computer in de jaren vijftig wordt gebruik gemaakt van algoritmen om besluiten te nemen – ook in overheidsprocessen. Zo kunnen algoritmen bepalen wie er belastingplichtig is en wie niet. Een aantal goed gedefinieerde, wettelijk vastgelegde regels ligt ten grondslag aan het algoritmische besluit. Bijvoorbeeld: wanneer iemand inkomen heeft uit werk is diegene belastingplichtig voor inkomensbelasting. Vervolgens kan deze regel worden gebruikt om volledig geautomatiseerd, zonder tussenkomst van menselijk handelen, te bepalen of iemand een brief toegestuurd krijgt waarin wordt opgeroepen om belastingaangifte te doen. De algoritmen die hiervoor worden gebruikt zijn in de loop der jaren steeds verder verfijnd als gevolg van verbeterde technologie.

Een recente studie van TNO laat zien dat het gebruik van AI binnen de publieke dienstverlening de afgelopen twee jaar is toegenomen.[7] De Algemene Rekenkamer onderzocht onlangs (AI-)algoritmen die *voorspellend*, waarbij het algoritme een risicovoorspelling doet, en *voorschrijvend*, waarbij het algoritme geautomatiseerd een besluit neemt, door de overheid worden toegepast.[8] Een voorbeeld van het eerste is een op AI gebaseerd model dat op basis van sensordata voorspelt wanneer een brug of sluis een onderhoudsbeurt moet krijgen of aan vervanging toe is. Een voorschrijvende toepassing is het eerdergenoemde voorbeeld waarin geautomatiseerd wordt bepaald wie er belastingplichtig is en dus een brief krijgt met een oproep om aangifte te doen.

Een vergissing in deze voorbeelden kan vervelend uitpakken, bijvoorbeeld wanneer een brug onnodig dicht blijft of iemand abusievelijk als belastingplichtig wordt aangemerkt. De Algemene Rekenkamer vond in haar onderzoek alleen relatief eenvoudige toepassingen die geheel geautomatiseerd plaatsvonden. Toch kunnen ook simpele besluiten of besluiten die niet geheel geautomatiseerd worden genomen een grote impact hebben op individuen. Bijvoorbeeld bij het gebruik van AI voor het bepalen van eindexamenscores. En dit kan ook leiden tot ongewenste effecten zoals ongelijkheid. Zodoende kan men zich afvragen of de toepassing van AI geschikt is voor algoritmische besluitvorming en of de voordelen van de technologie opwegen tegen de nadelige gevolgen.

Onderzoek van het Rathenau Instituut,[9] de Algemene Rekenkamer en TNO wijst op de risico's en nadelige gevolgen van algoritmische besluitvorming. Er zijn meerdere redenen om daar aandacht aan te besteden. Zo kan de vooringenomenheid in data of in een algoritme veranderen in ongewenste ongelijke behandeling van individuen of groepen. Bovendien wordt bij een algoritme mogelijk de vooringenomenheid van de programmeur ingebouwd, terwijl die – ten opzichte van de gebruikelijke besluitvormers – weinig inhoudelijke kennis en ervaring heeft met de specifieke toepassing van het algoritme. Het is echter juist voor maatschappelijke vraagstukken cruciaal dat de belangen van degenen over wie een besluit wordt genomen worden meegenomen en afgewogen ten opzichte van de beslissers.

7 TNO (2021), Quickscan AI in publieke dienstverlening II | Rapport | Rijksoverheid.nl.
8 Algemene Rekenkamer (2021), Aandacht voor algoritmes | Rapport | Algemene Rekenkamer.
9 Rathenau Instituut (2020), Nieuwe regels voor kunstmatige intelligentie? | Rathenau Instituut.

Tabel 1 geeft schematisch de nadelige gevolgen van op AI-systemen gebaseerde algoritmische besluitvorming weer. Sommige van deze gevolgen zijn niet uitsluitend verbonden met AI, maar hebben te maken met het (grootschalig) gebruik van (persoons)gegevens of juist met de toepassing van algoritmes in brede zin. AI kan hierbij een versterkende factor zijn. Daarnaast zijn er specifieke nadelige gevolgen verbonden aan algoritmes op basis van AI, met name bij de toepassing van 'machine learning' in besluitvormingsprocessen.

Tabel 1: Nadelige gevolgen van algoritmische besluitvorming en mogelijke oplossingen

Nadelige gevolgen	Oorzaken; oorsprong	(Mogelijke) oplossingen
Ongewenste of onbedoelde impact van voorspellingen; discriminatie of uitsluiting systemische ongelijkheid	Incomplete, incorrecte of vooringenomen data Misclassificatie van gevallen op basis van (zelflerende) algoritmen	Representatieve en kwalitatief hoogwaardige datasets Herstelmechanismen zoals klachtenprocedures, democratische controle, rechtszaken, protest
Onduidelijkheid over controleerbaarheid van besluiten of van systemen die besluiten nemen	Gebrek aan transparantie van het algoritme of systeem, gebrek aan uitlegbaarheid van (de uitkomsten uit) algoritmische besluitvorming	Transparantie, bijv. via algoritmeregisters; uitlegbaarheid en verantwoording over de werking van algoritmen en systemen
Ontmenselijking	Bureaucratisering van besluitvorming ('computer says no')	'Meaningful human control', toezicht en controle

Een eerste nadelig gevolg van geautomatiseerde besluitvorming is discriminatie en uitsluiting als gevolg van onvolledige, incorrecte of vooringenomen data. Voor het ontwikkelen van een AI-systeem is het allereerst belangrijk dat er voldoende bruikbare data voorhanden is om van te kunnen 'leren'. Dit geldt vooral voor '(un)supervised machine learning'. Wanneer er onvoldoende data beschikbaar is wordt het lastig om een algoritme te gebruiken voor het oplossen van een probleem. Naast de hoeveelheid data spelen ook juistheid en vooringenomenheid van de data een grote rol. Dat laatste wil zeggen dat de dataset die wordt gebruikt om het algoritme te trainen mogelijk niet representatief is voor de populatie waarvoor je je AI-systeem gaat inzetten, of wel representatief is, maar daarmee sociaal-maatschappelijke ongelijkheid bevestigt. Een voorbeeld hiervan is het COMPAS-systeem in de Verenigde Staten (zie kader).[10] Ditzelfde nadelige effect kan ook het gevolg zijn van een algoritme dat verkeerde patronen of inschattingen heeft aangeleerd. Autonome algoritmen kunnen namelijk toevallige afwijkingen opsporen en die foutief als patronen worden aanmerken. Dit kan leiden tot misclassificatie of uitsluiting van personen.

10 Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin (2016). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica.

COMPAS-systeem in de Verenigde Staten
In het COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) systeem werd een algoritme ingezet om de hoogte van een straf te bepalen bij een herhaaldelijke overtreding. Omdat de dataset waarop dit algoritme werd getraind vooral Afro-Amerikaanse mannen bevatte en omdat het juridische apparaat in de Verenigde Staten historisch gezien (veel) hogere straffen geeft aan deze groep dan aan anderen voor dezelfde overtreding, leidde dit algoritme tot een extreme versnelling van dit effect. Dit gebeurde zonder tussenkomst van een mens.

Een tweede nadelig gevolg betreft de uitlegbaarheid en beperkte controleerbaarheid van besluiten of van de systemen die besluiten nemen. Hierbij gaat het zowel om begrip van het algoritme zelf als de uitkomsten. Zowel de Algemene Rekenkamer als TNO concluderen dat in veel gevallen waarin de overheid besluiten neemt over burgers, de gegevensverwerking en de toepassing van het algoritme worden uitgevoerd door commerciële partijen. Dit betekent vaak dat het exacte ontwerp van een algoritme wordt gezien als eigendom van de externe partij, waardoor er niet altijd inzicht wordt geboden in de precieze werking. Dit maakt het lastiger voor een overheidsorganisatie om controle op een algoritmisch systeem uit te voeren en hierover verantwoording af te leggen.[11]

Ontmenselijking is een derde nadelig gevolg van algoritmische besluitvorming. Dit fenomeen treedt op wanneer een systeem uitspraak doet en er geen rekening wordt gehouden met de menselijke situatie. Ook bij besluitvorming waarop menselijke controle wordt uitgeoefend, is het mogelijk dat besluitvormers zich vrijwel automatisch conformeren aan de uitkomst van het systeem. Dit maakt menselijke controle van het systeem dus lastig. Ook maakt dit het moeilijk voor burgers om uitleg te krijgen over een besluit dat is genomen. Hierdoor worden mensen soms niet alleen geconfronteerd met een (verkeerde) nadelige uitkomst, maar krijgen zij ook geen begrijpelijk antwoord over waarom dit besluit werd genomen.

Bovenstaande nadelige gevolgen laten zien wat geautomatiseerde besluitvorming voor gevolgen kan hebben voor individuele burgers. Dit kan leiden tot een vertrouwenscrisis ten aanzien van algoritmische besluitvorming of zelfs de overheid als geheel. De protesten tegen het gebruik van het algoritme voor de examenscores in het Verenigd Koninkrijk zijn hiervan voorbeelden. Individuele ambtenaren en beleidsmakers kunnen ook vooringenomen zijn en systemische discriminatie en uitsluiting als gevolg daarvan komen voor. Om de invloed van die vooringenomenheid zo klein mogelijk te houden, bestaan er democratische controlemechanismen. Door de inzet van algoritmische besluitvorming komen die ‘checks and balances’ echter onder druk te staan. Om nadelige gevolgen als systemische discriminatie en ongelijkheid door het gebruik van algoritmen te voorkomen, ontwikkelen Europese en nationale overheden, waaronder Nederland, regulering en beleid gericht op AI.

11 En zou nopen tot het herzien van het aankoopproces van AI systemen binnen de overheid. Zie Van Noordt, C., Misuraca, G., Mortati, M., Rizzo, F. and Timan, T., (2020). AI Watch - Artificial Intelligence for the public sector, Publications Office of the European Union, Luxembourg.

3. EUROPEES AI-BELEID

Het voornaamste doel van het Europese AI-beleid is om AI-ontwikkeling en -toepassing te stimuleren via twee assen: *vertrouwen* en *excellentie*.[12] De technologie wordt immers gezien als een belangrijke aanjager van innovatie en economische groei. Europa wil niet achterblijven bij de enorme investeringen die elders in de wereld worden gedaan. Zoals in de Verenigde Staten, waar de meeste grote software- en databedrijven zijn gevestigd en in China, waar de staat gigantische investeringen doet in de ontwikkeling van AI. Via regulering die de invloed van buitenlandse platformen aan banden moet leggen[13] en beleid voor het stimuleren van AI-startups en MKB in Europa probeert de EU meer grip te krijgen op het AI-innovatielandschap.

Maar er is nog een reden om AI-ontwikkeling te stimuleren in Europa. Van Amerikaanse of Chinese technologie kan nog lastiger worden bepaald of deze zich wel gedraagt naar 'Europese waarden' als waardigheid, gelijke behandeling, het recht op privacy en informatieveiligheid. Daarom is het tweede doel van de Europese AI-strategie het waarborgen van de verantwoordelijke toepassing van AI. Dit wil zeggen dat de ontwikkeling en toepassing van Europese AI-systemen worden gereguleerd om te zorgen dat dit verantwoordelijk gebeurt. Tabel 2 laat een overzicht zien van de belangrijkste Europese beleidsstukken voor AI die zich steeds richten op het stimuleren én het reguleren van AI.

12 Europese Commissie (2021), Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future (europa.eu).
13 Europese Commissie (2020), Digital Markets Act: Ensuring fair and open digital markets (europa.eu).

Tabel 2: de Europese AI strategie.

EU AI-strategie	Stimulering technologieontwikkeling	Regulering verantwoordelijke toepassing
MEDEDELING VAN DE COMMISSIE Kunstmatige intelligentie voor Europa {COM(2018) 237 final}	– Technologische en industriële capaciteit van de EU vergroten en het gebruik van AI stimuleren in de private en publieke sector – Voorbereidingen treffen voor sociaaleconomische veranderingen als gevolg van AI	– Zorgen voor een passend ethisch en juridisch kader, op basis van de waarden van de Unie en in overeenstemming met het Handvest van de grondrechten van de EU
WITBOEK over kunstmatige intelligentie - een Europese benadering op basis van excellentie en vertrouwen {COM(2020) 65 final}	– Innovatie organiseren via Digital Innovation Hubs – Veilige toegang tot data voor AI – Implementeren van FAIR-principes rondom data	– Reguleren van hoog-risicotoepassingen – Zelfregulering via vrijwillige labels en certificering voor niet-hoog-risicotoepassingen – Samenwerkingsverbanden opzetten tussen nationale autoriteiten
Verslag over de gevolgen van kunstmatige intelligentie, het internet der dingen en robotica op het gebied van veiligheid en aansprakelijkheid {COM(2020) 64 final}	Additionele risicoanalyseprocedures voor producten die veranderen gedurende hun levenscyclus (bijv. door nieuwe software, nieuwe algoritmes of nieuwe data)	– Expliciete verplichtingen voor producenten omtrent immateriële schade voor kwetsbare gebruikers – Eisen inzake transparantie, robuustheid, verantwoordingsplicht en toezicht op algoritmen – Verplichtingen voor algoritmeontwikkelaars om ontwerpparameters en metagegevens openbaar te maken bij incidenten
Proposal for a REGULATION LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) {COM(2021) 206 final}	– Community of excellence – Nationale ‘regulatory sandboxes’ – Prioriteit voor MKB en startups om deze ‘proeftuinen’ te gebruiken	– Community of vertrouwen – Reguleren van hoog-risicotoepassingen – Europees en nationaal toezicht op hoog-risicotoepassingen

Een belangrijk onderdeel van de Europese AI-strategie gericht op het waarborgen van een op Europese waarden gebaseerde benadering van AI is het ontwikkelen van normenkaders voor verantwoordelijke AI. Om te komen tot principes en vereisten hiervoor heeft de Europese Commissie opdracht gegeven tot het aanstellen van een onafhankelijke High-Level Expert Group voor AI (AI HLEG) (zie kader).[14] Daarnaast hebben veel EU-lidstaten en industriepartijen eigen normenkaders ontwikkeld als een middel om verantwoordelijke AI te helpen ontwikkelen. Mede doordat normenkaders het gehele systeem beschouwen op basis van mensenrechten dienen ze als een startpunt voor beter begrip van systemen die leren en beslissingen nemen (gedeeltelijk) zonder menselijke tussenkomst.

De AI HLEG, die onder auspiciën van de Europese Commissie is opgericht, heeft een kader opgesteld met zeven vereisten voor betrouwbare AI. Deze zeven vereisten zijn gebaseerd op vier ethische grondbeginselen, zijnde respect voor menselijke autonomie, preventie van schade, rechtvaardigheid en verantwoording. De zeven vereisten zijn:

1. menselijke controle en toezicht;
2. technische robuustheid en veiligheid;
3. privacy en data-bestuur;
4. transparantie;
5. diversiteit, non-discriminatie en rechtvaardigheid;
6. maatschappelijk en milieuwelzijn;
7. aansprakelijkheid.

Een analyse van vijftien veelgebruikte normenkaders (zie kader) laat echter zien dat deze kaders veel verschillende normen bevatten, maar niet duidelijk maken wie bepaalt hoe normen worden toegepast. Ook blijft onduidelijk hoe keuzes kunnen worden gemaakt tussen tegenstrijdige waarden. Daarom zijn de kaders meestal beter geschikt om achteraf te beoordelen welke impact AI-systemen hebben. Daarnaast vinden AI-ontwikkelaars normenkaders vaak abstract en lastig om toe te passen bij de ontwikkeling van systemen in de praktijk. Belangrijke vragen zijn dus voor welke toepassingen deze principes gelden, wie deze principes moet gaan toepassen, hoe om te gaan met conflicterende waarden en dilemma's en op welk moment in het ontwikkelingsproces van een AI-systeem deze principes getoetst moeten worden.

14 Europese Commissie High-Level Expert Group on AI, Shaping Europe's digital future (europa.eu).

Vijftien normenkaders die onderdeel van zijn van de analyse:

1. AI Guidelines – Deutsche Telekom
2. Everyday Ethics for Artificial Intelligence – IBM
3. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms – Fairness, Accountability and Transparency in Machine Learning (FATML)
4. Artificial Intelligence and Machine Learning: Policy Paper – Internet Society (ISOC)
5. Ethically aligned design – Institute of Electrical and Electronics Engineering (IEEE)
6. ITI AI Policy Principles – Information Technology Industry Council (ITI)
7. Top 10 Principles for Ethical Artificial Intelligence – UNI Global Union
8. Recommendation of the Council on Artificial Intelligence – OECD
9. Charlevoix Common Vision for the Future of Artificial Intelligence – Leaders of the G7
10. Montréal Declaration for Responsible Development of AI – Université de Montréal
11. AI in the UK: ready, willing and able? – UK House of Lords, Select Committee on AI
12. An Ethical Framework for a Good AI Society – Floridi et al.
13. Preparing for the future of Artificial Intelligence – US National Science and Technology Council, Committee on Technology
14. How can humans keep the upper hand? Report on ethical matters raised by AI algorithms – French Data Protection Authority (CNIL)
15. Automated and Connected Driving: Report – Federal Ministry of Transport and Digital Infrastructure, Ethics Commission

Deze vragen worden vaak beantwoord door normenkaders te combineren met specifieke methodieken gericht op het ontwerpen of toetsen van AI-systemen, zoals begeleidingsethiek, ‘by-design’-methodiek of impact assessments. Begeleidingsethiek gaat uit van de vraag hoe technologie te ontwikkelen en toe te passen, op basis van een dialoog met de betrokken actoren over de effecten en waarden die een rol spelen.[15] De ontwikkelaars verwerken vervolgens de uitkomsten van de dialoog in de technologie. Daarnaast bestaan verschillende ‘by-design’-methodieken die zich richten op de vraag hoe je waarden kunt meenemen of ‘inbouwen’ tijdens het ontwerpproces van een nieuwe digitale dienst of toepassing. Denk aan ‘privacy-by-design’,[16] waarbij de nadruk ligt op gegevensbescherming en cryptografie, en ‘value-sensitive-design’[17] dat de nadruk legt op mensenrechten en publieke waarden, en deze in het ontwerp- en ontwikkelproces toetst.

15 Verbeek, P-P. & Tijink, D. (2019). Aanpak begeleidingsethiek: een dialoog over technologie met handelingsperspectief. ECP | Platform voor de InformatieSamenleving.
16 TNO (2021), Privacy by design: data combineren voor betere overheidsdienstverlening | TNO.
17 Van de Poel, I. (2013). Translating values into design requirements. In Philosophy and engineering: Reflections on practice, principles and process (pp. 253-266). Springer, Dordrecht.

Impact assessments worden gebruikt om gevolgen te bepalen of een inschatting te maken van risico's. De werkelijke inschatting van de impact van het gebruik van technologie gebeurt doorgaans pas achteraf (ex-post), zoals via het toetsingskader van de Algemene Rekenkamer.[18] Maar er zijn ook impact assessments die voorafgaand aan de toepassing (ex-ante) ingezet kunnen worden via een risicoschatting. Sommige assessments zijn gericht op het bepalen van de risico's van de toepassing van een bepaalde technologie.[19] Andere zijn gericht op het bepalen van de mogelijke risico's ten aanzien van een grondrecht. Zo is als onderdeel van de AVG het uitvoeren van een dataprotectie-impact assessment (DPIA) verplicht bij verwerking van grote hoeveelheden data en mogelijke hoge risico's voor individuen.[20]

De vraag is of het toepassen van de normenkaders via dergelijke methodieken de geïdentificeerde nadelige gevolgen voorkomt. Hoewel bovenstaande instrumenten de verantwoordelijkheid van nieuwe technologie toetsen, zijn er nog drie uitdagingen ten aanzien van de specifieke toepasbaarheid op AI voor algoritmische besluitvorming:

1. Veel normenkaders zijn gericht op AI-ontwikkelaars in plaats van op degenen die de technologie toepassen. AI-ontwikkelaars zijn vaak commerciële partijen die slechts op onderdelen betrokken zijn bij de uitvoering van de besluitvorming. De eindverantwoordelijkheid voor het systeem ligt in veel gevallen echter bij partijen die de AI-systemen toepassen voor geautomatiseerde besluitvorming. Organisatorische en procesmatige veranderingen zijn nodig om de uitkomsten van algoritmische besluitvorming inzichtelijk te maken. Hiervoor moet het gehele systeem waarin AI wordt toegepast worden meegewogen.

2. De uitlegbaarheid van algoritmen en het betrekken van partijen bij AI-toepassing. Zoals beschreven kunnen de nadelige gevolgen van AI voor burger en maatschappij ingrijpend zijn. De AVG verplicht dataverwerkende partijen om burgers te informeren over de verwerking van hun gegevens. Dit is des te meer van belang bij zelflerende systemen die zelf patronen aanleren of inschattingen maken, en lastig uit te leggen zijn. In het geval van AI zal regelgeving dan ook afdwingen om inzichtelijk te maken wie de data gebruikt voor het trainen van algoritmische modellen en binnen welke context de uitkomsten van invloed zijn op de zowel gebruiker als burger.

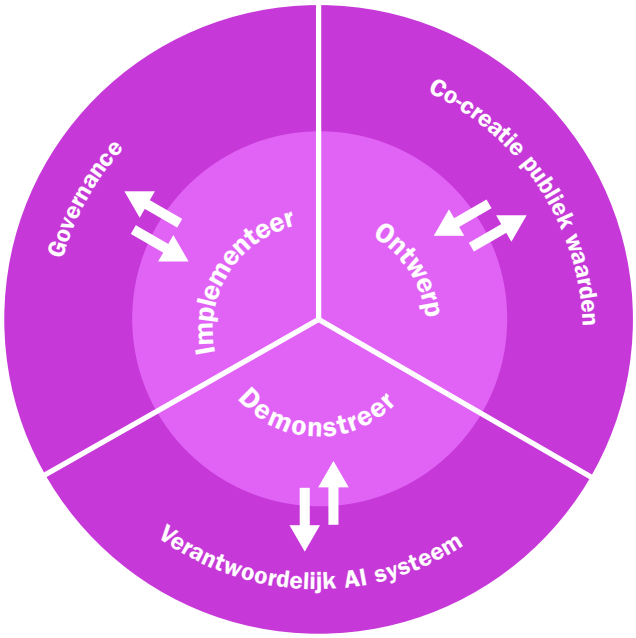
3. Er is nog veel onduidelijk over de langetermijngevolgen van algoritmische besluitvorming. Zo is het de vraag welke secundaire effecten optreden. Zoals het voorbeeld van de toepassing van een algoritme voor eindexamenscores in het Verenigd Koninkrijk liet zien, leidt dit mogelijk tot grotere ongelijkheid tussen leerlingen. Daarbij is het de vraag of deze effecten via een ex-ante impact assessment inzichtelijk worden. Daarom is een langduriger en dynamische controle en monitoring van AI-systemen gewenst.

18 Algemene Rekenkamer (2021), Aandacht voor algoritmes | Rapport | Algemene Rekenkamer.
19 ECP (2018), Artificial Intelligence Impact Assessment
20 Autoriteit Persoonsgegevens, Data protection impact assessment (DPIA) | Autoriteit Persoonsgegevens.

Normenkaders, zeker wanneer ze gekoppeld zijn aan begeleidingsethiek, waardengebaseerde ontwerpmethodiek of impact assessments, kunnen dus een bijdrage leveren aan de ontwikkeling van verantwoordelijke AI. Daarbij is het echter nodig dat het gehele systeem waarin AI wordt toegepast wordt meegewogen, dat de verschillende stakeholders, waaronder burgers, beter worden betrokken om hun belangen te vertegenwoordigen en dat ze dynamischer worden toegepast.[21]

4. PROEFTUIN VOOR AI-SYSTEMEN

Normenkaders gericht op het beschermen van publieke waarden zouden dus niet alleen moeten sturen op de ontwikkeling van een AI-systeem, maar ook gedurende de toepassing van een AI-systeem en op basis van belangen van alle betrokkenen van kracht moeten zijn. Om hiervoor te zorgen zou de toepassing van AI voor algoritmische besluitvorming – gebruikmakend van normenkaders en impact assessments – eerst moeten plaatsvinden in een experimentele omgeving, een *proeftuin* voor verantwoordelijke AI. In een dergelijke omgeving kan tijdig in kaart worden gebracht welke risico's zich voordoen in de verschillende stappen – van het genereren van gegevens tot verzameling en bewerking; van het gekozen AI-model tot aan de uitkomst van het systeem. Bovendien biedt zo'n proeftuin de gelegenheid om verschillende betrokkenen, inclusief burgers, mee te laten denken. Een voorbeeld van een methodiek die wordt toegepast om in samenwerking met partners verantwoordelijke AI-systemen te toetsen is de *dynamische impact assessment* van TNO (zie kader). Deze methodiek bestaat uit drie fasen met *leidende vragen* die aan bod komen in drie verschillende stadia van AI-toepassing bij algoritmische besluitvorming.



Dynamische impact assessment methodiek voor verantwoordelijke algoritmische besluitvorming

De *dynamische impact assessment* methodiek kent drie pijlers: (1) het betrekken van stakeholders en burgers om verschillende belangen mee te wegen in het ontwerp *(ontwerpen)*, (2) het toetsen van verantwoordelijkheid van de systemen *(demonstreren)*, en (3) het toepassen van AI-systemen voor algoritmische besluitvorming *(implementeren)*.

21 TNO werkt aan bovengenoemde uitdagingen in o.a. het AI Oversight Lab: https://appl-ai-tno.nl/service-labs/ai-oversight-lab/ AI Lab - TNO (appl-ai-tno.nl).

Ontwerpen:

1. Systeembeschouwing: **verkennen** van het systeem waarin de AI zou worden toegepast en dit samen met betrokken stakeholders vatten in een gemeenschappelijk model dat leidt tot specifieke deelvragen die met data worden beantwoord. *Leidende vragen* die aan de orde komen zijn: op welke maatschappelijke vragen is de AI-toepassing gericht? Wat is het gerechtvaardigd belang om data en AI in te zetten?
2. Experimenteeromgeving: indien AI een geschikte oplossing biedt, dan volgt het **inrichten** van een afgeschermde omgeving voor experimenten. Dit biedt de gelegenheid om ongewenste nadelige gevolgen in een vroeg stadium te identificeren en hier actie op te nemen. *Leidende vragen* die aan de orde komen zijn: welke stakeholders zijn nodig om een AI-toepassing vorm te geven? Welke wetgeving is van toepassing? Welk normenkader is passend?

Demonstreren:

3. Co-creatie en uitlegbaarheid: het **meewegen van belangen** van burgers en andere betrokkenen in de toepassing van AI-algoritmen in een systeem. *Leidende vragen* die aan de orde komen zijn: begrijpen burgers voldoende hoe het systeem werkt en wat voor impact dit op hun leven kan hebben? Wie is de maker, eigenaar en/of beheerder van het AI-systeem?
4. Op verantwoordelijke wijze **aanpassen en toepassen** van AI-systemen voor algoritmische besluitvorming. *Leidende vragen* die aan de orde komen, zijn: zijn datasets correct of vooringenomen of zorgen algoritmen voor discriminatie of uitsluiting? Wat betekenen transparantie en uitlegbaarheid in de context van dit beleidsonderwerp of besluitvorming? Hoe wordt menselijke controle ingericht?

Implementeren:

5. Toetsing van verantwoordelijkheid: het **toetsen** van de verantwoordelijkheid, waaronder de uitlegbaarheid van het systeem. *Leidende vragen* die aan de orde komen, zijn: welke methodiek is passend voor de toetsing van het systeem? Hoe kunnen we de burgers die de algoritmische besluitvorming betreft uitleg geven over de werking van de AI-systemen?
6. (Lange-termijn)impact: het **in kaart brengen en monitoren** van systemische effecten die kunnen optreden als gevolg van de toepassing van AI-systemen. *Leidende vragen* die aan de orde komen zijn: hoe en door wie kan controle worden uitgeoefend om te voorkomen dat dergelijke systemische effecten optreden? Wat zijn bewezen technologische en/of organisatorische maatregelen die we daarvoor kunnen inzetten?

Deze methodiek zou toegepast moeten worden op een transdisciplinaire manier, die verder gaat dan het betrekken van AI-ontwikkelaars en juristen, en ook het perspectief van de beleidsmedewerkers, toezichthouders en burgers meeweegt. Daarnaast is het van belang dat publieke organisaties actief aan de slag gaan met het stellen van de bovengenoemde vragen bij de toepassing van AI voor algoritmische besluitvorming. Dit geldt in het bijzonder voor de zogenaamde hoog-risicotoepassingen die zijn genoemd in de door de EU voorgestelde regulering van AI. Een voorbeeld wordt weergegeven in het kader CJIB dienstverlening schuldhulpverlening (zie kader).[22]

22 Steen, M., Timan, T. & van de Poel, I. (2021) Responsible innovation, anticipation and responsiveness: case studies of algorithms in decision support in justice and security, and an exploration of potential, unintended, undesirable, higher-order effects. AI Ethics.

CJIB dienstverlening schuldhulpverlening – secundaire en langetermijneffecten van AI
Een recente evaluatie van hoe ethische normen uit het normenkader van de AI HLEG zich in de praktijk manifesteren betreft een systeem waarmee het CJIB kijkt of ze mensen uit problematische schulden kan houden met behulp van AI. Dit gebeurt door te op basis van historisch betaalgedrag te voorspellen wanneer een bepaalde groep mensen met openstaande verkeersboetes in de problemen zou komen. De gedachte is om de uitvoeringsambtenaar in staat te stellen tijdig en gericht in te grijpen, door de groep van mensen die wel wíllen maar net niet kúnnen betalen eruit te lichten om een regeling te treffen. Er was in de ontwerpfase gekozen voor een minder goed werkend, maar gemakkelijker uitlegbaar algoritme om deze voorspelling te doen. Terwijl het AI-systeem volgens 'ethics by design'-principes werd ontwikkeld en goed leek te werken, kwam men in de toepassing achter 'secondaire' effecten, zoals de oneerlijkheid van het uitlichten van deze groep ten opzichte van anderen. Daarnaast ontstond er niet-voorzien extra werk voor de ambtenaar om de aanbevelingen van het AI-systeem te volgen dat ook ten koste ging van anderen. Dit maakt duidelijk dat het toetsen van de langetermijneffecten van AI-systemen onderdeel moet zijn van het controlesysteem.

De kern van de aanpak is dat de verschillende belangen gedurende het experiment meedoen met de toepassing van het AI-systeem, dat er op dynamische wijze controle wordt uitgeoefend op de toepassing van AI-systemen, dat het AI systeem desgewenst aangepast wordt en dat de resultaten op begrijpelijke wijze worden gecommuniceerd. Gedurende de toepassing wordt gerapporteerd wat de (verwachte) risico's en impact zijn. Hiervoor wordt een gestandaardiseerde 'bijsluiter' ontwikkeld die begrijpelijk is voor burgers.

5. CONCLUSIE

De toepassing van AI voor algoritmische besluitvorming is in opkomst, maar kan ook ongewenste nadelige gevolgen hebben. Om te zorgen dat AI zich in algoritmische besluitvorming verantwoordelijk gedraagt, wordt er onder andere naar bestaande en nieuwe wetgeving gekeken voor bijvoorbeeld dataprivacy en productaansprakelijkheid. Daarnaast wordt ook gekeken naar normenkaders en waardengedreven methodieken gebaseerd op mensenrechten, specifiek in relatie tot hoog-risicotoepassingen van AI. Het opstellen van normenkaders vooraf en toetsing van de systemen achteraf is echter onvoldoende om ongewenste negatieve effecten te voorkomen. Belangrijk is om dergelijke systemen in samenspraak met verschillende partijen – inclusief burgers – in de praktijk te beproeven en te monitoren. Ook vragen AI-algoritmen om dynamische controle en toezicht – ook op langetermijneffecten van AI.
De belangrijkste aanbeveling in dit paper is dan ook om experimentele omgevingen in te richten zodat AI-systemen voor algoritmische besluitvorming stapsgewijs worden beproefd.

Hoofd auteurs
Anne Fleur van Veenstra
Tjerk Timan

Overige auteurs
Gabriela Bodea, Cass Chideock, Ilina Georgieva,
Claudio Lazo, Mathilde Theelen

Contact
Babette Bakker, Strategie en Beleid
Locatie Den Haag – New Babylon
babette.bakker@tno.nl
+31 621 137 231

TNO innovation for life

TNO.NL