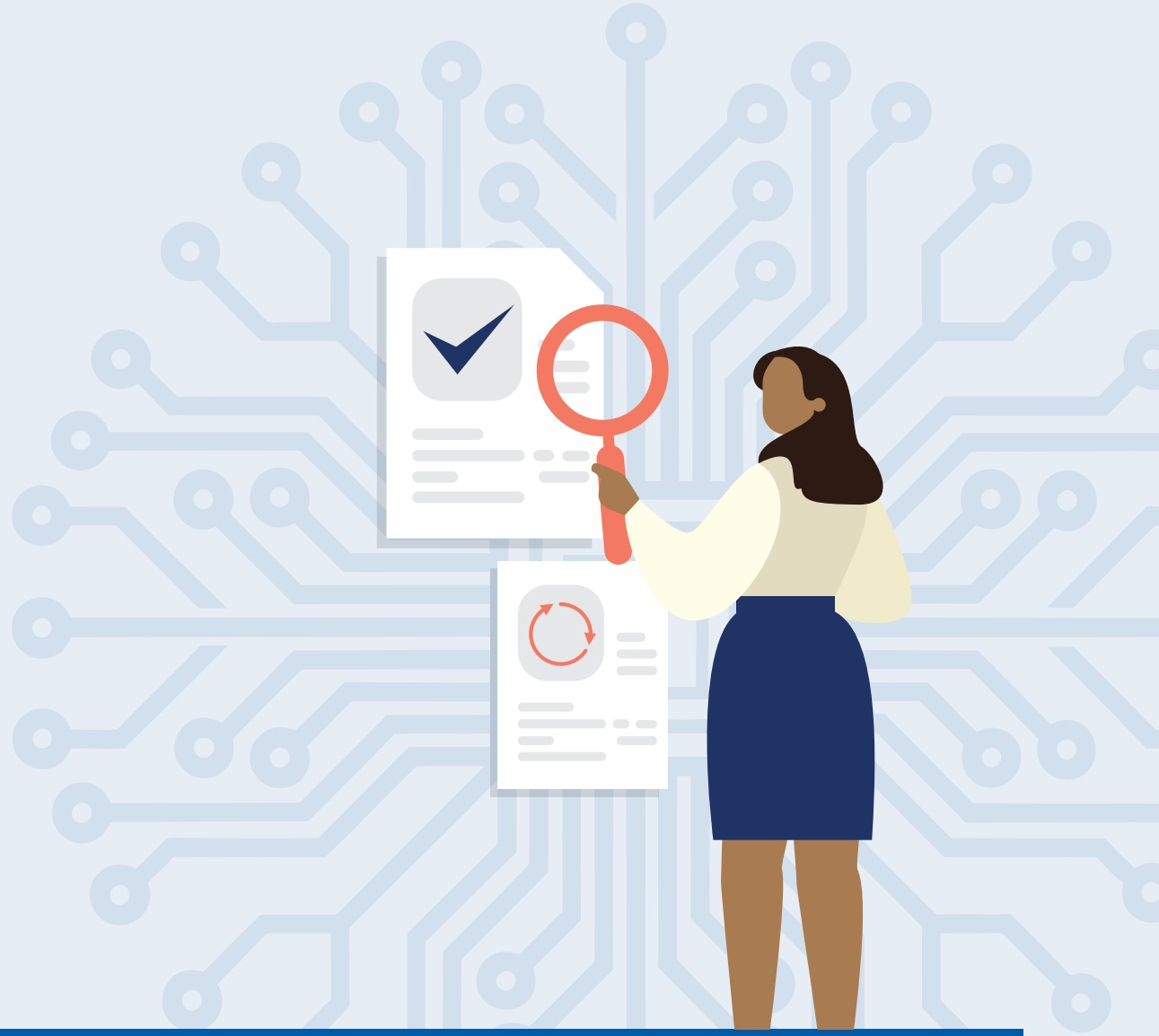
Algorithm Audit



Richtlijnen voor AI-verordening implementatie

Definitie van een AI-systeem

December 2024

1. Introductie

De AI-verordening stelt eisen aan de inzet van artificiële intelligentie (AI) in de Europese Unie (EU). De productveiligheidswetgeving ziet toe op verantwoorde ontwikkeling en gebruik van AI door publieke en private organisaties. Hiermee worden de veiligheid, gezondheid en grondrechten van burgers van de EU beschermd. Implementatie van de AI-verordening brengt echter lastige vraagstukken met zich mee, bijvoorbeeld welke algoritmische toepassingen binnen de reikwijdte van de verordening vallen.

Noch in het maatschappelijke debat rondom deze technologie, noch binnen de academische en technische wereld is in de afgelopen 50 jaar een vaste definitie gebruikt voor het begrip AI. Men maakte gebruik van intuïtieve, ongeschreven definities. Wat als AI wordt gezien groeit mee met wat geldt als de technologische cutting edge: zodra algemeen toegankelijke software complexe taken kan uitvoeren die voorheen waren voorbehouden aan 'AI', wordt het al snel niet meer als AI gezien.[1]

Met de komst van de AI-verordening wordt AI echter gevat in een juridisch bindende definitie. Het doel van de EU is om met deze definitie AI-systemen te onderscheiden van eenvoudigere traditionele software-systemen of programmeringsbenaderingen, en daarmee rechtszekerheid te bieden, brede acceptatie van de technologie te creëren, en toekomstbestendigheid te faciliteren.[2] De definitie zoals gebruikt door de Europese wetgever is niet nieuw: het volgt op de definitie van AI zoals ontwikkeld door de Organisatie voor Economische Samenwerking en Ontwikkeling (OESO).

Aan de hand van deze juridische definitie moeten organisaties aan de slag met implementatie van de AI-verordening. Dit blijkt ingewikkeld. Zo hebben juristen vaak weinig praktijkervaring met de technologieën onderliggend aan AI en zijn technici juist onervaren met juridische definities. Daarnaast bevat de definitie van een AI-systeem termen die niet allemaal evenveel gewicht hebben. Voor implementatie van de verordening is het daarom nodig om bruggen te slaan tussen deze verschillende werelden, waarbij niet alleen volledigheid maar ook pragmatiek in ogenschouw moet worden genomen. Deze white paper doet hiertoe een voorzet. We analyseren de kernelementen uit de definitie van een AI-systeem zowel vanuit juridisch als statistisch perspectief, waarmee de reikwijdte van de AI-verordening wordt verkend. Daartoe volgt een analyse van de zeven kenmerken van de definitie van een AI-systeem (sec. 2). Waarbij speciale aandacht wordt besteed aan het begrip inferentievermogen (sec. 3) en autonomie (sec. 4).

[1] Gezichtsherkenning en schaakcomputers werden lang als het ultieme voorbeeld van AI gezien, terwijl de toepassingen nu zijn geïntegreerd in het alledaagse leven en er niet meer als zodanig naar wordt verwezen. Dit fenomeen is door Pamela McCorduck beschreven als "het AI effect".

[2] Zie Appendix.

Box 1

Voorbehouden naleving AI-verordening

Dit document is een interpretatie van de wetstekst van de AI-verordening door stichting Algorithm Audit. Aan deze analyse kunnen geen rechten worden ontleed. Lezers worden geattendeerd op aanvullende richtlijnen voor interpretatie van de AI-verordening, die in de loop van 2025 worden gepubliceerd door de Europese Commissie.

2. AI-systeem definitie

De definitie van een AI-systeem wordt geïntroduceerd in artikel 3(1) van de AI-verordening. Deze definitie bepaalt de reikwijdte van de verordening. Alleen systemen die aan deze definitie voldoen vallen onder de wet. De definitie kan worden gevonden in Box 2.

We analyseren en interpreteren de bovenstaande zeven gekleurde begrippen. De belangrijkste bron voor deze interpretatie is overweging 12 uit de preambule van de verordening. Deze overweging bestaat uit 13 zinsdelen die in de Appendix worden vermeld. De overwegingen (Engels: *recitals*) geven inzicht in de intenties van de Europese wetgever bij het opstellen van de wetstekst en geven daarmee duiding hoe de begrippen geïnterpreteerd moeten worden. Voor de interpretatie zijn zowel de Engelse als Nederlandse wetsteksten geraadpleegd.

Bij analyse van deze begrippen wordt verwezen naar het memorandum van de OESO[3] (hierna: 'OESO-memorandum') over de definitie van een AI-systeem. Tijdens onderhandelingen over de AI-verordening is dit memorandum, inclusief eerdere conceptversies hiervan, gebruikt om tot een definitie te komen van een AI-systeem in definitieve wetstekst. In dit licht benoemt overweging 12 expliciet dat de EU de wens heeft *"nauw te aansluiten op het werk van internationale organisaties die zich bezighouden met AI, om rechtszekerheid te waarborgen, internationale convergentie en brede acceptatie te faciliteren"*.

[3] Explanatory Memorandum on the Updated OECD definition of an AI system (2024) https://www.oecd-ilibrary.org/science-and-technology/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en

Box 2

Artikel 3(1) van de AI-verordening definieert een AI-systeem als volgt:

"een op een machine gebaseerd systeem dat is ontworpen om met verschillende niveaus van autonomie te werken en dat na het inzetten ervan aanpassingsvermogen kan vertonen, en dat, voor expliciete of impliciete doelstellingen, uit de ontvangen input afleidt hoe output te genereren zoals voorspellingen, content, aanbevelingen of beslissingen die van invloed kunnen zijn op fysieke of virtuele omgevingen."

Box 3

Inconsistente vertalingen: Beslissingen of besluit? Content of inhoud?]

Merk op: Verschillende begrippen in de AI-verordening zijn niet consequent vertaald van het Engels naar het Nederlands.

In artikel 3 van de Nederlandse wetstekst wordt het woord *"afleidt"* gebruikt, waar de Engelse wettekst spreekt over *"infers"*. In overweging 12 wordt gesproken over *"inferentievermogen"* en in het Engels over *"capability to infer"*. Op een vergelijkbare manier wordt in artikel 3 gesproken over *"beslissingen"*, terwijl in overweging 12 wordt gesproken over *"besluiten"*. In de Engelse wetstekst wordt consequent "decisions" gebruikt. Ook "*content*" wordt niet consequent vertaald: in de Nederlandse wetstekst wordt in artikel 3 gesproken over *"inhoud"*, en in overweging 12 over *"content"*.

Deze verschillen in de vertaling benadrukken dat het van waarde is om bij interpretatie van internationale wetgeving ook altijd te kijken naar de Engelse wetstekst.

Iedere analyse van bovenstaande begrippen sluiten we af met een oordeel in hoeverre het als eenduidig criterium kan dienen om AI-systemen te onderscheiden van reguliere algoritmen.[4] Een uitgebreide analyse van het begrip inferentie en autonomie wordt behandeld in 3. Inferentie en 4. Autonomie.

2.1 Interpretatie van de definitie van een AI-systeem aan de hand van overweging 12

Overweging 12 bevat een aantal zinsneden die helpen om interpretatie van de definitie van een AI-systeem te kaderen:

i) *"de [AI-systeem] definitie moet gebaseerd zijn op de belangrijkste kenmerken van AI-systemen die het onderscheiden van eenvoudigere traditionele softwaresystemen of programmeringsbenaderingen"*;

ii) *"[de AI-systeem definitie] mag geen betrekking hebben op systemen die gebaseerd zijn op regels die uitsluitend door natuurlijke personen zijn vastgesteld om automatisch handelingen uit te voeren."* – Zie overweging 12 zin 2.

Uit zinsnede i) volgt de lens waardoorheen we de definitie van een AI-systeem interpreteren: de kenmerken in de definitie moeten het onderscheid tussen AI-systemen en andere softwaresystemen mogelijk maken. De zin is ook een ondergrens waarmee de wetgever aangeeft dat de reikwijdte van de definitie van een AI-systeem niet alle programmeringsbenaderingen betreft. Onder 'eenvoudige traditionele software-systemen' zou simpele dataverwerking in Excel of SQL kunnen worden verstaan. AI kan in deze programmeringsbenaderingen ook meer geavanceerde dataverwerking worden uitgevoerd wat mogelijk wél een AI-systeem betreft.

Zinsnede ii) refereert naar regelgebaseerde algoritmen waarbij de regels door natuurlijke personen zijn opgesteld. Een voorbeeld van een regel is als leeftijd <65 jaar, dan geen recht op seniorenkorting. Als de variabele leeftijd en de drempelwaarde 65 jaar uitsluitend door natuurlijke personen zijn vastgesteld om de automatische handelingen van het bepalen van een korting uit te voeren, is het regelgebaseerde algoritme niet een AI-systeem. Dit is ook het geval wanneer dit algoritme wordt ingezet voor impactvolle doeleinden, zoals risicoprofilering. Zinsnede ii) heeft een sterk vermogen om AI-systemen te onderscheiden van algoritmen.

2.2 Machine gebaseerd systeem

Overweging 12 vermeldt dat *"op een machine gebaseerd systeem"* uit de AI-systeem definitie de volgende betekenis kent:

"De term 'op een machine gebaseerd' verwijst naar het feit dat AI-systemen op machines draaien." – zie overweging 12 zin 7.

Aangezien vrijwel alle moderne softwaresystemen of programmeringsbenaderingen gebruikmaken van een machine, zijnde een computer, server of virtual machine (VM)[5], voldoen vrijwel alle softwaresystemen en algoritmen aan deze vereiste.

We stellen daarom vast dat de 'machine-gebaseerd systeem'-vereiste geen onderscheidend vermogen heeft om AI-systemen van andere algoritmen te scheiden, omdat alle moderne softwaresystemen of programmeringsbenaderingen machine-gebaseerd zijn.

---

[4] 'Algoritme' zoals gedefinieerd door de Algemene Rekenkamer (2021): 'Een set van regels en instructies die een computer geautomatiseerd volgt bij het maken van berekeningen om een probleem op te lossen of een vraag te beantwoorden'.

[5] Met een VM wordt verwezen naar een microprocessor die op een PC, laptop, of in een cloud-omgeving, algoritmen uitvoert. Zie ook 3.32 uit ISO/IEC 13522-6:1998 Information technology — Coding of multimedia and hypermedia information

2.3 Verschillende niveaus van autonomie

Overweging 12 vermeldt dat "*verschillende niveaus van autonomie*" uit de AI-systeem definitie de volgende betekenis kent:

"AI-systemen worden zodanig ontworpen dat zij in verschillende mate autonoom kunnen functioneren, wat betekent dat zij een zekere mate van onafhankelijkheid van menselijke betrokkenheid bezitten en zonder menselijke tussenkomst kunnen functioneren." – zie overweging 12 zin 12.

In de Engelse tekst wordt gesproken van *"some degree of independence"*.

Er moet dus sprake zijn van *enige mate* van autonomie. Daarom zien we autonomie als een factor waarmee AI-systemen onderscheiden kunnen worden van algoritmen. In 4. Autonomie wordt uitgebreider ingegaan op de betekenis en interpretatie van autonomie.

2.4 Aanpassingsvermogen kan vertonen

Overweging 12 vermeldt dat *"na het inzetten ervan aanpassingsvermogen kan vertonen"* uit de AI-systeem definitie de volgende betekenis kent:

"Het aanpassingsvermogen dat een AI-systeem na het inzetten ervan kan vertonen, heeft betrekking op zelflerende capaciteiten, waardoor het systeem tijdens het gebruik kan veranderen."

Door gebruik van het werkwoord kan is het op basis van de Nederlandse wetstekst onduidelijk of aanpassingsvermogen een vereiste eigenschap is van een AI-systeem. De Engelse versie van de overweging luidt:

"The adaptiveness that an AI system could exhibit after deployment, refers to self-learning capabilities, allowing the system to change while in use." – zie overweging 12 zin 12.

Door gebruik van het werkwoord *could* (zou kunnen) wordt opgemaakt dat aanpassingsvermogen van een AI-systeem geen vereiste is. Ook OESO ziet aanpassingsvermogen na het inzetten ervan als optioneel, in het memorandum benoemt het expliciet ook een systeem dat eenmalig is geleerd uit data als een AI-systeem.[6] Veel AI-systemen die momenteel in gebruik zijn vertonen geen aanpassingsvermogen na het inzetten ervan. Gezichtherkenningssoftware, waar de AI-verordening op verschillende plekken naar verwijst, zijn een voorbeeld waarbij modelparameters over het algemeen niet tijdens gebruik maar enkel voorafgaand aan een softwarerelease worden geüpdatet. Kortom, ook AI-systemen die tijdens gebruik geen aanpassingsvermogen vertonen kunnen nog steeds een AI-systeem zijn, als aan de andere voorwaarden wordt voldaan.

We concluderen dat 'aanpassingsvermogen' geen vereiste is voor de AI-systeem definitie. Daarmee is het geen onderscheidende factor om AI-systemen te scheiden van andere algoritmen.

2.5 Expliciete of impliciete doelstellingen

Overweging 12 vermeldt dat *"voor expliciete of impliciete doelstellingen"* uit de AI-systeem definitie de volgende betekenis kent:

"De verwijzing naar expliciete of impliciete doelstellingen onderstreept dat AI-systemen kunnen functioneren volgens expliciete, gedefinieerde doelstellingen, of volgens impliciete doelstellingen. De doelstellingen van een AI-systeem kunnen verschillen van het beoogde doel van het AI-systeem in een specifieke context." – zie overweging 12 zin 8.

[6] Supra noot 3

Een toepassing streeft altijd een doel na, dat ofwel expliciet ofwel impliciet gedefinieerd kan zijn. De reden dat dit element is opgenomen in de definitie is om uit te drukken dat een expliciete doelstelling geen vereiste is voor een AI-systeem.[7] Bijvoorbeeld middels reinforcement learning kunnen AI-systemen zelf doelstellingen afleiden, die niet expliciet zijn geformuleerd maar wel impliciet vervat zijn in het AI-systeem. Dit is ook het geval bij Large Language Models (LLMs) zoals ChatGPT en andere toepassingen van generatieve AI.

De 'doelstelling'-vereiste heeft geen onderscheidend vermogen om AI-systemen van andere algoritmen te scheiden.

2.6 Uit de ontvangen input afleidt hoe output te genereren

Overweging 12 vermeldt dat "*uit de ontvangen input afleidt hoe output te genereren*" uit de AI-systeem definitie de volgende betekenis kent:

"Een belangrijk kenmerk van AI-systemen is hun inferentievermogen. Dit inferentievermogen slaat op het proces waarbij output, zoals voorspellingen, content, aanbevelingen of besluiten, wordt verkregen waarmee fysieke en virtuele omgevingen kunnen worden beïnvloed, en op het vermogen van AI-systemen om modellen of algoritmen, of beide, af te leiden uit input of data." – zie overweging 12 zin 3-4.

Het begrip 'afleiden' uit de definitie wordt aan de hand van het begrip 'inferentie' in overweging 12 toegelicht. We concluderen dat inferentievermogen het belangrijkste element is van de definitie om AI-systemen te onderscheiden van andere algoritmes. In 3. Inferentie wordt ingegaan op de betekenis en interpretaties van inferentie.

2.7 Voorspellingen, inhoud, aanbevelingen of beslissingen

Overweging 12 vermeldt dat "*voorspellingen, content, aanbevelingen of besluiten*" uit de AI-systeem definitie de volgende betekenis kent:

"… de output die door het AI-systeem wordt gegenereerd is een uiting van de verschillende functies van AI-systemen en kan de vorm aannemen van voorspellingen, content, aanbevelingen of besluiten." – zie overweging 12 zin 10.

Deze passage houdt verband met inferentie, het afleiden van output uit input, waarvan een analyse volgt in 3. Inferentie. Met betrekking tot *"voorspellingen, inhoud, aanbevelingen of beslissingen"* gaat het hier om verschillende vormen van output die worden afgeleid:

1. **Voorspellingen:** Hieronder vallen ingeschatte scores, rangschikkingen, kansen, labels en classificaties. Dit hoeft niet per se een voorspelling over de toekomst te zijn, aangezien een voorspelling ook betrekking kan hebben op een niet eerder geobserveerd datapunt. Het statistische begrip 'schatter' wordt in dit geval ook een voorspelling genoemd (Engels: estimation, prediction).
2. **Inhoud/content:** Hieronder valt gegenereerde tekst, beeld en spraak, bijvoorbeeld gecreëerd middels generatieve AI.
3. **Aanbevelingen:** Hieronder vallen aanbevelingssystemen, zoals gepersonaliseerde tijdlijnen op social media platforms, zoekmachineresultaten en online advertenties. Tot deze categorie behoren ook aanbevolen handelingen, zoals een aanbeveling voor extra controle die volgt op een toegekende risicoscore voor onrechtmatig gebruik van een sociale voorziening, of een auto die aanbeveelt om naar een andere versnelling te schakelen.[8] Scores of classificaties waaraan

[7] Supra noot 3

[8] Supra noot 3

beleidsmatig een vaste actie of handeling verbonden is, kunnen ook worden gezien als aanbevelingen. Denk aan: een toegekende risicoscore in transactiemonitoring binnen banken, aan de hand waarvan een werkinstructie voorschrijft dat aanvullend onderzoek moet worden uitgevoerd.

4. **Beslissingen/besluiten:** Hieronder lijken beslissingen (*"decisions"*) in de breedste zin des woords te vallen, zoals de beslissing om een actie of handeling uit te voeren, bijvoorbeeld een auto die geautomatiseerd remt voor een voetganger[9], de keuze om een controle uit te voeren, het vaststellen van iemands identiteit (verificatie) of een formeel besluit zoals gedefinieerd in de Algemene wet bestuursrecht (Awb art.1:3).[10] Voor de publieke sector is het belangrijk op te merken dat algoritmische output die wordt gebruikt in de voorbereidende fase van een besluit ook als onderdeel van het gehele besluitvormingsproces beschouwd dient te worden en daarmee ook dient te voldoen aan de algemene beginselen van behoorlijk bestuur (abbb), zoals het motiveringsbeginsel, het zorgvuldigheidsbeginsel en het beginsel van *fair play*.[11] Wanneer de output een aanbeveling of besluit is, is het begrip 'geautomatiseerde besluitvorming' uit de Algemene Verordening Gegevensbescherming (AVG) relevant.[12]

De voorbeelden (voorspellingen, inhoud/content, aanbevelingen of besluiten/beslissingen) zijn een belangrijk signaal wat de wetgever ziet als output van AI-systemen. Aan de hand van deze opsomming kunnen een aantal type algoritmen uitgesloten worden die niet kwalificeren als AI-systemen. Zo stellen we vast dat algoritmen die beschrijvende (populatie)statistieken berekenen, zoals gemiddelden en standaardafwijkingen, geen AI-systeem zijn. Bij het berekenen van het gemiddelde inkomen van een groep natuurlijke personen is de output geen *"voorspelling, content, aanbeveling of beslissing/besluit"*. Wanneer een statistisch model gebruikt wordt om een score te schatten voor een nieuw datapunt, dan is er wel sprake van een voorspelling. Volgens deze redeneerwijze gelden eenvoudige dataverwerking- en visualisatiesystemen, zoals dashboards die populatiestatistieken weergeven, niet als AI-systeem.

We zien kenmerken van de output van een AI-systeem daarom als een belangrijke factor om AI-systemen te onderscheiden van andere algoritmen, zeker in combinatie met en in relatie tot de begrippen autonomie en inferentie.

Voor de vraag of een algoritme met een *"voorspelling, content, aanbeveling of beslissing/besluit"* als output ook daadwerkelijk een AI-systeem betreft, is het belangrijk om na te gaan hoe de output tot stand komt. Manieren om het proces van de verkregen output, vanuit het licht van de AI-systeem definitie, nader te onderzoeken wordt toegelicht in 3. Inferentie.

---

[9] Supra noot 3

[10] Zie ook Advies geautomatiseerde besluitvorming, Autoriteit Persoonsgegevens https://www.autoriteitpersoonsgegevens.nl/documenten/advies-geautomatiseerde-besluitvorming

[11] Hoe 'algoprudentie' kan bijdragen aan een verantwoorde inzet van machine learning-algoritmen, A. Meuwese, J. Parie, A. Voogt, 2024, Nederlands Juristenblad (NJB) https://algorithmaudit.eu/nl/knowledge-platform/knowledge-base/white_paper_algopruden-ce/

[12] Art. 22 AVG. Zie ook supra noot 10.

2.8 Fysieke en virtuele omgeving

Overweging 12 vermeldt dat *"invloed kunnen zijn op fysieke of virtuele omgevingen"* uit de AI-systeem definitie de volgende betekenis kent:

"Voor de toepassing van deze verordening moeten onder omgevingen de contexten worden verstaan waarin de AI-systemen werken, terwijl de output die door het AI-systeem wordt gegenereerd een uiting is van de verschillende functies van AI-systemen en de vorm kan aannemen van voorspellingen, content, aanbevelingen of besluiten." – zie overweging 12 zin 10.

De fysieke en virtuele omgeving zijn complementair. De combinatie van de twee omgevingen is uitputtend. Het gaat hier dus om systemen die überhaupt invloed uitoefenen, op welke omgeving dan ook. Dit sluit systemen uit die helemaal geen invloed uitoefenen, bijvoorbeeld omdat ze nog niet in gebruik genomen zijn. Verder biedt noch overweging 12 noch het OESO-memorandum behulpzame duiding voor het begrip invloed. Er lijkt bijna geen systeem denkbaar dat niet invloed uitoefent op een omgeving.

In ieder geval is de vereiste van 'invloed op de fysieke of virtuele omgeving' geen criterium waarmee AI-systemen van algoritmen onderscheiden kunnen worden. Het begrip invloed wordt indirect ook besproken in de begrippen in 3. Inferentie en 4. Autonomie.

3. Inferentie

Inferentievermogen is het belangrijkste element van de definitie om AI-systemen te onderscheiden van reguliere algoritmen. In deze sectie worden verschillende passages uit overweging 12 geanalyseerd en gerelateerd aan de AI-systeem definitie.

Overweging 12 vermeldt dat inferentievermogen de volgende betekenis kent:

"Een belangrijk kenmerk van AI-systemen is hun inferentievermogen. Dit inferentievermogen slaat op het proces waarbij output, zoals voorspellingen, content, aanbevelingen of besluiten, wordt verkregen waarmee fysieke en virtuele omgevingen kunnen worden beïnvloed, en op het vermogen van AI-systemen om modellen of algoritmen, of beide, af te leiden uit input of data." – Zie overweging 12 zin 3-4.

"De technieken die inferentie mogelijk maken bij de opbouw van een AI-systeem, omvatten benaderingen op basis van machinaal leren waarbij aan de hand van data wordt geleerd hoe bepaalde doelstellingen kunnen worden bereikt, alsook op logica en kennis gebaseerde benaderingen waarbij iets wordt geïnfereerd uit gecodeerde kennis of uit een symbolische weergave van de op te lossen taak." – Zie overweging 12 zin 5.

"Het inferentievermogen van een AI-systeem overstijgt de elementaire verwerking van data door leren, redeneren of modelleren mogelijk te maken." – Zie overweging 12 zin 6.

De eerste en laatste zin kaderen de interpretatie: het inferentievermogen is een belangrijk kenmerk waaraan AI-systemen geïdentificeerd kunnen worden en het is specifiek dit kenmerk dat AI-systemen onderscheidt van andere dataverwerking door *"leren, redeneren of modelleren"*. Merk op dat enkel sprake hoeft te zijn van een van deze drie kenmerken: leren, redeneren óf modelleren.

Aan de hand van deze drie kernbegrippen worden bovenstaande zinnen uit overweging 12 geanalyseerd.

3.1 Leren en modelleren

Overweging 12 benoemt dat inferentievermogen betrekking heeft op:

"het vermogen van AI-systemen om modellen of algoritmen, of beiden, af te leiden uit input of data." – zie overweging 12 zin 4.

Wanneer modellen of algoritmen zijn afgeleid uit data, is er sprake van modelleren of leren. Voorbeelden hiervan zijn het leren van de gewichten van een neuraal netwerk gebruikt voor spraakherkenning of een algoritme dat kenmerken selecteert voor profilering. Verschillende experts gebruiken hiervoor verschillende termen zoals leren, modeleren, trainen of fitten. Ongeacht de gebruikte terminologie, volgt uit deze passage van overweging 12 dat er sprake is van inferentie wanneer een model of algoritme wordt afgeleid uit input of data. Uit deze passage blijkt dat AI-systemen het vermogen moeten hebben om af te leiden. Hieronder verstaan we dat er een mate van automatisering moet zijn bij het afleiden van modellen of algoritmen uit data. Wanneer eerst een data-analyse wordt uitgevoerd, bijvoorbeeld om de gemiddelde leeftijd van een populatie vast te stellen, wat als input dient voor domeinexperts die handmatig een algoritme opstellen, dan is er geen sprake van een situatie waarin een AI-systeem een algoritme afleidt uit data.

Overweging 12 benoemt verder:
"De technieken die inferentie mogelijk maken bij de opbouw van een AI-systeem, omvatten benaderingen op basis van machinaal leren …" – zie overweging 12 zin 6.

Bij machinaal leren wordt een model 'geleerd' uit een dataset, vaak training data genoemd. In veel gevallen wordt statistiek gebruikt om modelparameters te berekenen die het beste passen bij de beschikbare dataset. Voor datawetenschappers laat het berekenen van parameters op basis van inputdata zich het beste uitdrukken als de .fit()-functie, zoals gebruikt in scikit-learn en statsmodels Python-software. Het berekenen van een gemiddelde, aan de hand van een simpele formule, is een voorbeeld van een parameter. Zo ook het berekenen van lineaire regressie-coëfficiënten, aan de hand van een meer uitgebreide formule, of de gewichten van een neuraal netwerk aan de hand van een zeer complexe formule.

Machinaal leren omvat ook het leren van de variabelen en drempelwaarden van een beslisboom voor regressie en classificatie. Dit betreft het leren van een simpele beslisboom, maar ook het leren van groepen beslisbomen, zoals ensemble-based tree learning. Denk aan: random forest, xgboost, explainable boosting etc. Dit zijn allen voorbeelden van machinaal leren.

Of een datagedreven toepassing machinaal leren (Engels: 'machine learning') genoemd wordt verschilt per domeinexpertise. Een econometrist of statisticus zal het ontwikkelen van een lineair model zoals een regressievergelijking of general linear model (GLM) waarschijnlijk geen machine learning noemen. Toch wordt in dit geval een model afgeleid van een beschikbare dataset. We zien op basis van de tekst van overweging 12 geen onderscheid tussen welke techniek er wordt gebruikt. We concluderen dat alle gevallen wanneer een model gefit, getraind of geleerd wordt uit data vallen onder inferentievermogen.

Toch maakt enkel het afleiden van modelparameters of regels uit input data, bijvoorbeeld het leren van regressiecoëfficiënten, een model of algoritme nog geen AI-systeem. Overweging 12 vermeldt dat inferentievermogen slaat op:

a) *"het proces waarbij output voorspellingen, content, aanbevelingen of besluiten wordt verkregen […] waarmee omgevingen kunnen worden beïnvloed";*
b) *"het vermogen van AI-systemen om modellen of algoritmen, of beide, af te leiden uit input of data"* – zie overweging 12 zin 4.

Bij het leren van regressiecoëfficiënten wordt voldaan aan b) – namelijk: .fit() – maar niet aan a). Bij het leren van regressiecoëfficiënten worden immers geen voorspellingen gemaakt voor nieuwe datapunten. Bij a) gaat het over het toepassen van het geleerde model of algoritme op nieuwe data. Naar dit proces wordt door datawetenschappers verwezen als .predict(), zoals gebruikt in scikit-learn en statsmodels Python-software. Dit relateert ook aan de door de wetgever gespecificeerde output van een AI-systeem, namelijk: "voorspellingen, content/inhoud, aanbevelingen of beslissingen/ besluiten". Enkel na het toepassen van deze .predict()-functie wordt dus output gegenereerd die volgens de definitie vereist is. In het geval van aanbevelingen en beslissingen wordt vaak eerst een score voorspeld, aan de hand van een geleerd model, waarna aan de hand van deze score een aanbeveling wordt gedaan of beslissing wordt genomen. Een model – gebaseerd op statistiek of machine learning – is een AI-systeem als modelparameters of regels worden berekend én daarna een voorspelling of soortgelijk volgt. Zie ook 2.7 Voorspellingen, inhoud, aanbevelingen of beslissingen en 2.8 Fysieke en virtuele omgeving.

Het 'genereren van output'-aspect is een belangrijke factor om AI-systemen van algoritmen te onderscheiden.

3.2 Redeneren: op logica en kennis gebaseerde benaderingen

Inferentievermogen kan ook betrekking hebben op het vermogen van een AI-systeem om te *redeneren* – zie overweging 12 zin 6. Hieruit blijkt dat er een type systemen is waarbij geen sprake is van leren of modelleren, maar wel van inferentie.

Dit roept de vraag op: bij welk type algoritmen is er sprake van redeneren? Overweging 12 noemt een aantal voorbeelden van systemen die hier niet onder vallen: *"regels die uitsluitend door natuurlijke personen zijn vastgesteld om automatisch handelingen uit te voeren"* en *"elementaire verwerking van data"* – zie overweging 12 zin 2 en 6.

Verder biedt overweging 12 weinig aanvullende duiding voor het begrip *"redeneren"*. In overweging 12 wordt wel het volgende benoemd:
"De technieken die inferentie mogelijk maken bij de opbouw van een AI-systeem, omvatten … op logica en kennis gebaseerde benaderingen waarbij iets wordt geïnfereerd uit gecodeerde kennis of uit een symbolische weergave van de op te lossen taak." – zie overweging 12 zin 5.

Bij op logische en kennis gebaseerde benaderingen van AI is geen sprake van machinaal leren, er is hier sprake van inferentievermogen omdat er sprake is van redeneren.

Op logica en kennis gebaseerde benaderingen van AI worden in de academische wereld ook wel *symbolische AI* genoemd, zo ook in het OESO-memorandum.[13] Symbolische AI is sinds de jaren '80 en '90 gebruikt in bijvoorbeeld schaakcomputers of medische beslissingsondersteuningssystemen. Met de grote vooruitgang op het gebied van machine learning, deep learning en generatieve AI, is er echter steeds minder aandacht uitgegaan naar deze vorm van AI.

[13] Supra noot 3

Overweging 12 bevat geen aanvullende informatie over de definitie en interpretatie van op logica en kennis gebaseerde benaderingen van AI-systemen. In het originele voorstel van de AI-verordening is wel aanvullende duiding opgenomen: *"Op logica en op kennis gebaseerde benaderingen, waaronder kennisrepresentatie, inductief (logisch) programmeren, kennisbanken, inferentie- en deductiemachines, (symbolisch) redeneren en expertsystemen"*.[14] Deze voorbeelden zijn in lijn met interpretaties van symbolische AI in de academische wereld.

Om op logica en kennis gebaseerde AI-systemen te onderscheiden van algoritmen moeten we onderscheiden wat deze technieken anders maakt dan *"regels die uitsluitend door natuurlijke personen zijn vastgesteld om automatisch handelingen uit te voeren"* en *"de elementaire verwerking van data"*. We duiden op logica en kennis gebaseerde benaderingen aan de hand van twee academische standaardwerken in AI: Artificial Intelligence van Russel and Norvig en Artificial Intelligence van Poole and Mackworth.[15] Samengevat bestaan op logica en kennis gebaseerde benaderingen van AI uit:

i) **Knowledge-base:** Een expliciete representatie van (domein)kennis. Hiervoor wordt vaak logica gebruikt, waarbij kennis wordt uitgedrukt in proposities en connectieven, zoals ¬A, A∧B, A∨B, waarbij een propositie (bijv. A) enkel waar of onwaar kan zijn. Andere bekende vormen van knowledge bases zijn knowledge graphs.

ii) **Redeneer component:** Deze component definieert hoe het systeem kan redeneren over de kennis in de knowledge-base en input data, bijvoorbeeld door middel van formele logica. Deze component wordt ook wel een inference engine genoemd. Door middel van de redeneer component kunnen nieuwe kennis én nieuwe regels worden afgeleid.

Beide componenten worden zorgvuldig opgebouwd en vereisen veel domeinkennis. Vaak worden deze benaderingen gebruikt wanneer er sprake is van een grote hoeveelheid vaste kennis en regels in een domein, waarover vervolgens geredeneerd kan worden. Denk aan medisch beslissingsondersteuningssysteem waarbij de kennisbasis medische feiten bevat over symptomen, diagnoses en mogelijke behandelingen, het redeneersysteem kan dan op basis van input data van symptomen een mogelijke behandeling voorstellen.

Op logica en kennis gebaseerde benaderingen van AI vormen tegenwoordig een minderheid. Meestal worden deze technieken vandaag de dag gebruikt in combinatie met vormen van machine learning. In dat geval zou het systeem vanwege het gebruik van machine learning een AI-systeem zijn, zie 3.1 Leren en modelleren. Ontwikkelaars die dit type technologie gebruiken zijn hier waarschijnlijk op de hoogte dat zij dit type AI-systeem gebruiken. We zien het 'op logica en kennis gebaseerde benaderingen'-aspect alleen in die zeldzame gevallen dat geen ML gebruikt wordt, als een belangrijk vereiste om AI-systemen van algoritmen te onderscheiden.

3.2.1 Redeneren, gecodeerde kennis en rule-based systemen

Er zijn, anders dan op logica en op kennis gebaseerde benaderingen, geen andere benaderingen waar met betrekking tot redeneren naar wordt verwezen in de AI-verordening. In de context van de AI-

[14] Zie Annex I van Voorstel voor een verordening van het Europees Parlement en de Raad tot vaststelling van geharmoniseerde regels betreffende artificiële intelligentie (Wet op de artificiële intelligentie) en tot wijziging van bepaalde wetgevingshandelingen van de Unie. https://eur-lex.europa.eu/legal-content/NL/TXT/HTML/?uri=CELEX:52021PC0206

[15] Artificial Intelligence: foundations of computational agents. Poole, D.L. and Mackworth, A.K., 2010. Cambridge University Press. Artificial intelligence: a modern approach. Russell, Stuart J., and Peter Norvig. Pearson, 2016. Zie voor een begrijpelijke uitleg ook: https://en.wikipedia.org/wiki/Knowledge-based_systems

verordening houdt redeneren dus alleen verband met op logica en kennis gebaseerde benaderingen.

Er kan worden beargumenteerd dat in het geval van eenvoudige handmatig opgestelde regelgebaseerd algoritme sprake is van redeneren. Dit is echter onverenigbaar met de toelichting die op de definitie van een AI-systeem wordt gegeven: *"de definitie [moet] gebaseerd zijn op de belangrijkste kenmerken van AI-systemen die het onderscheiden van eenvoudigere traditionele softwaresystemen of programmeringsbenaderingen..."*. Als regel-gebaseerde algoritmen redeneren, redeneren alle soorten softwaresystemen en dat gaat tegen de strekking van voorgaande zin in. Ongeacht of er sprake is van redeneren geldt dat *"regels die uitsluitend door natuurlijke personen zijn vastgesteld om automatisch handelingen uit te voeren"* geen AI-systeem zijn – overweging 12 zin 2.

De passage over *"gecodeerde kennis"* – overweging 12 zin 6 – moet ook in licht van op logica en kennis gebaseerde benaderingen worden bezien. Gecodeerde kennis relateert in deze context aan de vorm waarin kennis wordt gecodeerd in een knowledge-base, zoals in 3.2 Redeneren: op logica en kennis gebaseerde benaderingen beschreven. Regelgebaseerde algoritmen, waarin menselijke kennis is gecodeerd, worden in de praktijk niet toegepast middels een knowledge-base (ook wel: een 'op kennis gebaseerde benadering'). De passage *"gecodeerde kennis"* heeft dus geen betrekking op regelgebaseerde algoritmen die we kennen uit de praktijk.

4. Autonomie

Overweging 12 vermeldt dat "*verschillende niveaus van autonomie*" uit de AI-systeem definitie de volgende betekenis kent:

"AI-systemen worden zodanig ontworpen dat zij in verschillende mate autonoom kunnen functioneren, wat betekent dat zij een zekere mate van onafhankelijkheid van menselijke betrokkenheid bezitten en zonder menselijke tussenkomst kunnen functioneren." – zie overweging 12 zin 11.

Om te voldoen aan de 'autonomie'-vereiste moet er sprake zijn van enige mate van autonomie, zoals ook besproken in 2.3 Verschillende niveaus van autonomie.

'Enige mate' is een zwakke vereiste: een systeem hoeft niet geheel autonoom te zijn om aan deze vereiste te voldoen. De AI-verordening geeft echter geen nadere duiding wat onder het begrip autonomie, en de verschillende gradaties hiervan, moet worden verstaan.

Het OESO-memorandum vermeldt dat: *"de autonomie van een AI-systeem betrekking heeft op de mate waarin een systeem kan leren of handelen zonder menselijke betrokkenheid"*. Dit impliceert dat ieder lerend algoritme in een zekere mate autonoom is. Oftewel, als aan de inferentie-vereiste wordt voldaan, wordt ook aan de autonomie-vereiste voldaan. Verder wordt in het OESO-memorandum autonomie gekoppeld aan de verschillende types gegenereerde output, waarbij beslissingen het meest autonoom en voorspellingen het minst autonoom zijn. Uit deze formulering maken we op dat OESO ook voorspellingen als in 'enige mate' van autonoom beschouwd. Met een beschouwing van het type output van een algoritme (2.7 Voorspellingen, inhoud, aanbevelingen of beslissingen) en het inferentievermogen (3. Inferentie) kan daarmee ook de autonomie-vereiste worden ingevuld.

Al met al concluderen we dat de 'autonomie'-vereiste geen aanvullend onderscheidend vermogen heeft ten opzichte van de andere vereisten om AI-systemen van algoritmen te scheiden.

Appendix

Overweging 12 uit de preambule van de AI-verordening.

Zin 1 – geanalyseerd in 1. Introductie

Het begrip "AI-systeem" in deze verordening moet duidelijk worden gedefinieerd en moet nauw aansluiten op het werk van internationale organisaties die zich bezighouden met AI, om rechtszekerheid te waarborgen, internationale convergentie en brede acceptatie te faciliteren, en tegelijkertijd de nodige flexibiliteit te bieden om op de snelle technologische ontwikkelingen op dit gebied te kunnen inspelen.

Zin 2 – geanalyseerd in 2.1 Interpretatie van de definitie van een AI-systeem aan de hand van overweging 12

Bovendien moet de definitie gebaseerd zijn op de belangrijkste kenmerken van AI-systemen die het onderscheiden van eenvoudigere traditionele softwaresystemen of programmeringsbenaderingen, en mag het geen betrekking hebben op systemen die gebaseerd zijn op regels die uitsluitend door natuurlijke personen zijn vastgesteld om automatisch handelingen uit te voeren.

Zin 3-4 – geanalyseerd in 3.1 Leren en modelleren

Een belangrijk kenmerk van AI-systemen is hun inferentievermogen. Dit inferentievermogen slaat op het proces waarbij output, zoals voorspellingen, content, aanbevelingen of besluiten, wordt verkregen waarmee fysieke en virtuele omgevingen kunnen worden beïnvloed, en op het vermogen van AI-systemen om modellen of algoritmen, of beide, af te leiden uit input of data.

Zin 5-6 – geanalyseerd in 3.2 Redeneren: op logica en kennis gebaseerde benaderingen

De technieken die inferentie mogelijk maken bij de opbouw van een AI-systeem, omvatten benaderingen op basis van machinaal leren waarbij aan de hand van data wordt geleerd hoe bepaalde doelstellingen kunnen worden bereikt, alsook op logica en kennis gebaseerde benaderingen waarbij iets wordt geïnfereerd uit gecodeerde kennis of uit een symbolische weergave van de op te lossen taak. Het inferentievermogen van een AI-systeem overstijgt de elementaire verwerking van data door leren, redeneren of modelleren mogelijk te maken.

Zin 7 – geanalyseerd in 2.2 Machine gebaseerd systeem

De term "op een machine gebaseerd" verwijst naar het feit dat AI-systemen op machines draaien.

Zin 8-9 – geanalyseerd in 2.5 Expliciete of impliciete doelstellingen

De verwijzing naar expliciete of impliciete doelstellingen onderstreept dat AI-systemen kunnen functioneren volgens expliciete, gedefinieerde doelstellingen, of volgens impliciete doelstellingen. De doelstellingen van een AI-systeem kunnen verschillen van het beoogde doel van het AI-systeem in een specifieke context.

Zin 10 – geanalyseerd in 2.7 Voorspellingen, inhoud, aanbevelingen of beslissingen

Voor de toepassing van deze verordening moeten onder omgevingen de contexten worden verstaan waarin de AI-systemen werken, terwijl de output die door het AI-systeem wordt gegenereerd een uiting is van de verschillende functies van AI-systemen en de vorm kan aannemen van voorspellingen, content, aanbevelingen of besluiten.

Zin 11 – geanalyseerd in 2.3 Verschillende niveaus van autonomie en 4. Autonomie

AI-systemen worden zodanig ontworpen dat zij in verschillende mate autonoom kunnen functioneren, wat betekent dat zij een zekere mate van onafhankelijkheid van menselijke betrokkenheid bezitten en zonder menselijke tussenkomst kunnen functioneren.

Zin 12 – geanalyseerd in 2.4 Aanpassingsvermogen kan vertonen

Het aanpassingsvermogen dat een AI-systeem na het inzetten ervan kan vertonen, heeft betrekking op zelflerende capaciteiten, waardoor het systeem tijdens het gebruik kan veranderen.

Zin 13 – niet geanalyseerd, want niet van specifieke toegevoegde waarde

AI-systemen kunnen op standalonebasis of als component van een product worden gebruikt, ongeacht of het systeem fysiek in het product is geïntegreerd (ingebed) dan wel ten dienste staat van de functionaliteit van het product zonder daarin te zijn geïntegreerd (niet-ingebed).

Gemeente Amsterdam

Deze white paper is in samenwerking met de gemeente Amsterdam tot stand gekomen.

Over Algorithm Audit

Algorithm Audit is een Europees kennisplatform voor AI bias testing en normatieve AI-standaarden.
De doelen van de stichting zijn drieledig:

Normatieve adviescommissies	Adviseren over ethische kwesties in concrete algoritmische toepassingen door het samenbrengen van deliberatieve, diverse adviesommissies, met algoprudentie als resultaat
Technische hulpmiddelen	Implementeren en testen van technische methoden voor bias-detectie en -mitigatie, zoals onze bias detection tool
Kennisplatform	Samenbrengen van kennis en experts voor collectief leerproces over verantwoorde inzet van algoritmes, bijvoorbeeld ons AI Policy Observatory en position papers

Structurele partners van Algorithm Audit

SIDN Fonds

Het SIDN Fonds staat voor een sterk internet voor iedereen. Het Fonds investeert in projecten met lef en maatschappelijke meerwaarde, met als doel het borgen van publieke waarden online en in de digitale democratie.

European AI&Society Fund

Het European AI&Society Fund ondersteunt organisaties uit heel Europa die AI beleid vormgeven waarin mens en maatschappij centraal staan. Het fonds is een samenwerkingsverband van 14 Europese en Amerikaase filantropische organisaties.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Het ministerie van BZK maakt zich sterk voor een democratische rechtsstaat, met een slagvaardig bestuur. Ze borgt de kernwaarden van de democratie. BZK staat voor een goed en digitaalvaardig openbaar bestuur en een overheid waar burgers op kunnen vertrouwen.

Opbouwen van *publieke kennis*

over verantwoorde AI *zonder winstoogmerk*

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Algorithm Audit

Stichting Algorithm Audit is geregistreerd bij de
Kamer van Koophandel onder nummer 83979212