Ministerie van Infrastructuur
en Waterstaat

AI Impact Assessment

Het hulpmiddel voor een betrouwbaar AI-project

Versie 2.0 december 2024. In samenwerking door collega's van het ministerie Infrastructuur en Waterstaat (IenW) bij Concerndirectie Informatiebeleid (CDIB), Inspectie Leefomgeving en Transport (ILT IDlab en afdeling analyse) en Rijkswaterstaat (RWS Datalab)

Vragen en opmerkingen ontvangen we graag op teamai@minienw.nl

Inhoudsopgave

Het AI impact assessment helpt bij verantwoorde AI by design

Artificial Intelligence (**AI**) biedt kansen, maar het brengt ook risico's met zich mee. Het is belangrijk om de impact van een AI-systeem helder te hebben voor het ingezet wordt om te voorkomen dat er onbedoelde, negatieve gevolgen zijn. Op deze manier kunnen de kansen van AI optimaal benut worden. Voor een verantwoorde inzet van AI hebben het IDlab en afdeling analyse ILT, het RWS Datalab en Concerndirectie Informatiebeleid van IenW het AI Impact Assessment (hierna: AIIA) ontwikkeld.

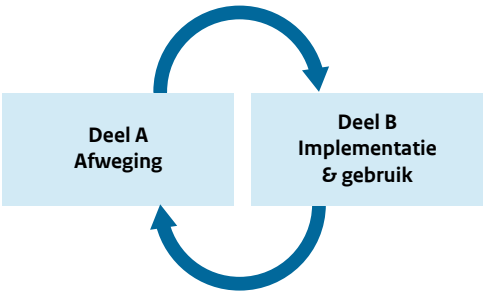
Het AIIA dient als hulpmiddel en begeleidt het denkproces, met als doel de verantwoording, kwaliteit en **reproduceerbaarheid** van AI-inzet te vergroten. Het AIIA kijkt naar obstakels in de dataverzameling, het AI-systeem, de algoritmiek en houdt rekening met geldende wet- en regelgeving. Een ingevuld AIIA maakt de gemaakte afwegingen om een **AI-systeem** wel of niet te gebruiken inzichtelijk.

Hierbij gaan we uit van de toepassing van AI binnen een specifieke context. Als het AI-systeem (of delen hiervan) bijvoorbeeld voor een ander doel wordt ingezet, moet er opnieuw een AIIA uitgevoerd worden. Denk bijvoorbeeld aan wanneer een beeldherkenningsmodel voor schepen ook ingezet wordt voor andere voertuigen.

In de **AI-verordening** is er een aantal risicogebieden voor AI-systemen, van onaanvaardbaar tot minimaal risico. Afhankelijk van het risico zijn er meer of minder maatregelen nodig. Hanteer daarom altijd de risico-categorisering van de AI-verordening (zie bijlage 1). Als het AI-systeem bijvoorbeeld van een laag- naar een hoog-risicotoepassing gaat, volgen er strengere eisen om aan de AI-verordening te kunnen voldoen.

Vul het AIIA in met een multidisciplinaire groep
Het AIIA moet worden ingevuld met een multidisciplinaire groep personen uit de organisatie. Voor de verschillende in te vullen onderdelen, is namelijk een ander soort kennis nodig. Het kopje 'Wie doet wat?' geeft een overzicht van rollen die relevant kunnen zijn per hoofdstuk. Sommige vragen moet een data scientist beantwoorden en andere vragen een jurist.

De verschillende delen van het AIIA
Het AIIA bestaat uit twee delen. Deel A gaat in op de afwegingen voor het gebruik van een AI-systeem: wat is het doel en de verwachte effecten? Met deze informatie wordt een afweging gemaakt voor de toepassing van het AI-systeem en eventuele maatregelen. Op deze manier is de ethische discussie rondom de wenselijkheid van de toepassing aantoonbaar. Deel B gaat over de inrichting, implementatie en het gebruik van het AI-systeem.



De delen zijn los van elkaar in te vullen, maar er zit een grote samenhang tussen. Soms is het nodig om bij het invullen van deel A alvast wat meer de diepte in te duiken op bepaalde onderwerpen uit deel B om een goede afweging te kunnen maken. Het type algoritme heeft bijvoorbeeld invloed op de duurzaamheid en daarmee de afweging. Keuzes tijdens de implementatie kunnen de afweging om het AI-systeem wel of niet te gebruiken beïnvloeden. Bijvoorbeeld wanneer een aanname in deel A bij implementatie in deel B niet blijkt te kloppen (bijvoorbeeld vanwege de beschikbare data of gebruik van andere IT-infrastructuur), waardoor de afweging van deel A niet meer klopt.

Gebruik het AIIA in alle fasen van je AI-project

Het AIIA kan toegepast worden in elke fase van ontwikkeling en inkoop van een AI-systeem. Het AIIA heeft de meeste toegevoegde waarde indien het vanaf de start van een AI-project wordt gebruikt. Het AIIA moet verplicht ingevuld zijn wanneer het systeem naar productie gaat (dit kan ook een pilot zijn). Onderstaande tabel geeft de mogelijkheden weer. Er is geen pre-AIIA omdat het AIIA verplicht is voor élk AI-systeem, onafhankelijk van het risico vanuit de AI-verordening. In onderstaande tabel staan de verschillende toepassingsmogelijkheden van het AIIA.

Quick-Scan AIIA	Met een Quick-Scan AIIA onderzoek je of een AI-idee haalbaar en wenselijk is. Gebruik hiervoor met name de blauwe vragen uit deel A, aangevuld met de blauwe vragen uit deel B als verdieping. Zoek hierbij de diepgang en detailniveau dat past bij het stadium van ontwikkeling en risico's van de applicatie. Bij een AI-idee hoeft bijvoorbeeld nog niet alles duidelijk te zijn, maar wel inzicht in waar mogelijke knelpunten liggen. De uitkomsten maken duidelijk of het een goed idee is om een AI-systeem in te kopen of te ontwikkelen en kan gebruikt worden als een go/no-go moment.
Opstellen projectplan	Stel je een projectplan op voor het gebruik van een AI-systeem, dan ben je verplicht om de AIIA in te vullen. Gebruik het AIIA om de afweging voor het AI-systeem transparant vast te leggen en de keuzes te verantwoorden. Het is een goed middel voor gesprek en als check of alle relevante aspecten zijn meegenomen. De handreiking AI voor opdrachtgevers bevat praktische tips voor de ontwikkeling, implementatie en beheer van een AI-systeem.
Tijdens de ontwikkeling	Implementatiekeuzes hebben invloed op de impact van een AI-systeem (bijvoorbeeld de gebruikte data of type model). Gebruik deel A om de impact te bepalen en de keuze om AI-systeem te gebruiken. Gebruik deel B en check of alle relevante aspecten zijn meegenomen in de implementatie.
AI-systeem in productie	Een ingevulde AIIA is verplicht voordat een AI-systeem naar productie gaat. Eenmaal in productie, evalueer je of het AI-project nog aan de eisen voldoet. Onderzoek of het toepassingsgebied is gewijzigd.

AIIA voor impactvolle algoritmes
Het AIIA is geschreven voor gebruik bij AI-systemen, machine-gebaseerde systemen met een lerend component. Er zijn ook eenvoudigere **ALGORITMES**, het gaat hierbij om een 'recept' met vooraf gedefinieerde regels. Het AIIA kan ook ingezet worden voor dit soort **ALGORITMES**, maar is niet verplicht. De vragen helpen bij het beoordelen van de impact van het algoritme en welke implementatie-keuzes relevant zijn voor een verantwoorde toepassing.

Hulp nodig?
Past het AIIA niet bij een AI-project of systeem waaraan je denkt? Loop je vast bij het invullen van een vraag of spelen er ethische dilemma's? Of heb je andere opmerkingen of vragen over het AIIA? Roept het invullen van het AIIA vragen op? Neem dan contact op met het Team AI van IenW.

Verantwoording versie 2.0
Het IDlab en afdeling analyse ILT, het RWS Datalab en Concerndirectie Informatiebeleid van IenW (CDIB) hebben de eerste versie van het AIIA ontwikkeld. Het AIIA is vastgesteld door de Bestuursraad van IenW op

4 juli 2022. In deze versie 2.0 (2024) is informatie over generatieve AI toegevoegd, zijn gebruikerservaringen verwerkt en is informatie over de definitieve teksten van de AI-verordening opgenomen. Daarnaast is de samenhang met het Impact Assessment Mens en Algoritme (IAMA) versterkt[1]. Het AIIA gaat hiermee niet alleen meer over hoe AI toegepast *kan* worden, maar ook of het *wenselijk* is om AI in te zetten.

Leeswijzer

Voor het invullen van het AIIA is het volgende van belang:

- Een ingevulde AIIA is verplicht op het moment dat het AI-systeem naar productie gaat.
- De mate waarin het AIIA wordt ingevuld ligt bij de expertise van de projectleider en proportioneel naar de impact van het AI-systeem. Enkel 'ja' of 'nee' volstaat echter niet als antwoord op de vragen, tenzij anders aangegeven.
- Blauwe vragen zijn verplicht. Vul deze altijd in.
- Groene vragen zijn hulpvragen. Vul deze in als ze relevant zijn.
- Dikgedrukte woorden zijn aanklikbare begrippen, gedefinieerd in bijlage Begrippenlijst.
- Er is een invultemplate beschikbaar.

Houd er rekening mee dat de Auditdienst Rijk en de Algemene Rekenkamer het AI-systeem kunnen controleren op correctheid en veiligheid. Een volledig ingevuld AIIA betekent niet per definitie dat het AI-systeem veilig is. Om te voldoen aan de AI-verordening (Regulation (EU) 2024/1689), moet Bijlage 1 ingevuld worden indien je AI-systeem met hoog risico is.

[1] Het is daarmee niet meer nodig om zowel een AIIA én IAMA in te vullen, behalve als er extra hulp nodig is bij de afweging voor grondrechten. Zie hoofdstuk 2.1.

Wie doet wat

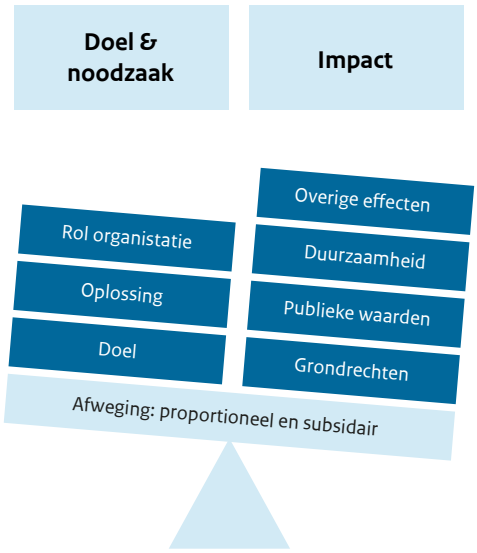
Het is belangrijk dat het AIIA wordt ingevuld door een multidisciplinair team omdat er voor de verschillende onderdelen specifieke kennis of deskundigheid nodig is. Onderstaande tabel geeft een *suggestie* voor de betrokkenheid van verschillende rollen voor een bepaald hoofdstuk van het AIIA. Uiteraard hangt het af van de scope en grootte van het project of deze rollen daadwerkelijk nodig zijn of juist meer betrokken moeten zijn dan de suggestie in onderstaande tabel.

	H1. Doel	H2. Impact	H3. Afweging	H4. Technische robuustheid	H5. Data governance	H6. Risicobeheer	H7. Verantwoordingsplicht
Belanghebbenden:		X					
CIO:	X	X	X	X	X	X	X
CISO				X	X	X	X
Communicatieadviseur:		X					X
Data scientists:	X	X	X	X	X	X	X
Databeheerder of bronhouder:					X		
Domeinexpert:	X	X	X	X	X	X	
Privacy professional:	X	X	X	X	X		
Jurist:		X	X		X		X
Opdrachtgever:	X	X	X	X	X	X	X
Overige leden projectteam:							
Projectleider:	X	X	X	X	X	X	X
Strategisch adviseur ethiek:	X	X	X				

Deel A: Afweging

Is de impact van het **AI-systeem** in verhouding tot de beoogde doelstellingen? Die vraag staat centraal in het eerste deel van het AIIA. Loop alle vragen door en besteed daarin aandacht aan de doelen, de beoogde oplossing en verwachte effecten. Met name de proportionaliteit van inzet van AI is relevant. Wellicht zijn er andere, minder ingrijpende manieren om het doel te bereiken?

Als het een lastige afweging is, bijvoorbeeld omdat er verschillende belangen botsen of omdat het AI-systeem inbreuk maakt op grondrechten, zijn er verschillende tools om een gestructureerd gesprek over ethische aspecten te voeren. Denk bijvoorbeeld aan de datadialoog van IBDS, DEDA, Moreel beraad of een van vele andere methoden. Het AI-team IenW biedt workshops aan binnen het IenW-concern om dit gesprek te ondersteunen.



1. Doel en noodzaak van het systeem

Deze vragen gaan over het doel van het **AI-systeem**, en de rol die het systeem binnen de organisatie gaat hebben.

1.1 Doel van het systeem

Het 'toepassen van AI' is geen doel op zich. Een AI-systeem wordt ingezet om een bepaald doel te bereiken binnen de organisatie, bijvoorbeeld om het werk efficiënter of beter uit te voeren. Welk probleem moet opgelost worden? Kijk hierbij naar het gehele proces waarin AI een rol speelt. Het antwoord op deze vragen kan gebruikt worden bij het invullen van de rest van het AIIA

1. Geef een korte beschrijving van het beoogde doel en beoogde resultaat van het AI-systeem. (titel, algemene omschrijving, probleemstelling, verwachte tijdsperiode, plaats, doelgroepen, het domein en het werkproces)

2. In welk risicogebied van de AI-verordening past jouw AI-systeem: **Onaanvaardbaar,** hoog of minimaal risico?

3. Waar in de organisatie (in welke processen?) is beoogd het AI-systeem te gebruiken?

Om er achter te komen in welk risicogebied van de AI-verordening jouw systeem valt, raden we je aan de toets in bijlage 1 te doen. Zo krijg je meer duidelijkheid over de juridische verplichtingen achter jouw systeem.

Valt het systeem in de categorie '**onaanvaardbaar**', dan is het AI-systeem niet toegestaan. Dit AIIA verder invullen is dan niet nodig.

Indien je AI-systeem in de hoog-risicocategorie valt, moet je de resterende vragen in bijlage 1 ook invullen. Zodoende krijg je een overzicht van de extra verplichtingen waaraan hoog-risicosystemen moeten voldoen.

Voor alle overige systemen, uit de categorie laag of minimaal risico, zijn de vragen van het AIIA afdoende.

1.2 Beoogde oplossing

In dit deel kijken we naar de beoogde oplossing voor het eerder beschreven probleem, zoals de AI-technieken die toegepast worden en de data die gebruikt wordt. Maak, waar nodig, gebruik van de vragen in deel B om onderstaande vragen goed te kunnen beantwoorden

1. Geef een korte beschrijving van het beoogde **AI-systeem** (techniek, data en type algoritme)
2. Waarom is er voor deze vorm van AI gekozen? (Denk bv. aan generatieve AI, lineaire regressie of neuraal netwerk)

3. Welke alternatieven zijn overwogen? (Denk bv. aan geen AI, minder complexe AI, ander type algoritme)

1.3 Rol binnen de organisatie

Een AI-systeem heeft, net als elk ander IT-systeem een opdrachtgever en eindverantwoordelijke. Eigenaarschap is essentieel. Bij deze vragen stel je ook de rolverdeling binnen het ontwikkelen en gebruiken van je systeem vast. Deze rollen staan gedefinieerd in de begrippenlijst. Houd deze definities aan.

1. Beschrijf de rolverdeling binnen het opzetten van het AI-systeem (zoals de **ontwikkelaar**, **opdrachtgever, projectleider, beheerorganisaties en eindverantwoordelijke**). Indien deze is ontwikkeld door een externe partij: wat zijn de contractuele afspraken?
2. Wie is de **gebruiker** van het AI-systeem, wie zijn de **eindgebruikers** die met het systeem werken en welke **betrokkenen** ondervinden impact van het AI-systeem?

3. Met welke stakeholders, mensen en/of groepen is er afgestemd bij het ontwikkelen van het AI-systeem?
4. Welke feedback is er verzameld van teams of groepen die verschillende achtergronden en ervaringen representeren? En wat is hier vervolgens mee gedaan?

1.4 Onderhoud en beheer

Een AI-systeem heeft, net als elk ander IT-systeem onderhoud en beheer nodig. Ook bij een *proof of concept* of pilot is het goed om te onderzoeken waar en op welke wijze het AI-systeem in productie gaat. Zo kunnen vooraf al de juiste keuzes gemaakt worden en dit voorkomt bijvoorbeeld een niet-compatible IT-infrastructuur of AI-technieken waardoor een AI-systeem niet in beheer kan worden genomen.

1. Beschrijf de rolverdeling voor het beheer en onderhoud van het AI-systeem (zoals de **ONTWIKKELAAR, OPDRACHTGEVER, PROJECTLEIDER, BEHEERORGANISATIES en EINDVERANTWOORDELIJKE**). Indien het systeem is ontwikkeld door een externe partij: wat zijn de contractuele afspraken?

2. Hoe wordt rekening gehouden met het ingaan en updaten van wet- en regelgeving tijdens de levensduur van dit AI-systeem?

3. Is de benodigde deskundigheid voor het beheer van AI-systeem gedocumenteerd?

4. Hoe wordt rekening gehouden met verandering van de context van het AI-systeem?

2. Impact

De vragen in dit onderdeel brengen de impact van de toepassing van het AI-systeem in een bepaalde context in kaart. We nemen hier expliciet grondrechten en duurzaamheid mee, maar er zijn ook andere effecten, bijvoorbeeld positief, negatief of op brede welvaart. Het gaat om impact in de breedste zin van het woord, bijvoorbeeld de impact op bepaalde doelgroepen of op de brede welvaart.

Loop het hoofdstuk stapsgewijs door, er zit namelijk een volgordelijkheid in. Publieke waarden gaan bijvoorbeeld ook over mensenrechten en deze zijn al afgedekt in het hoofdstuk 'Grondrechten'. Ook in deel B is een groot deel van de impact al afgedekt.

2.1 Grondrechten

Iedereen die belang heeft bij de werking van het **AI-SYSTEEM** moet goed behandeld worden. Dat betekent dat de (fundamentele) rechten van alle **BETROKKENEN** gewaarborgd moeten worden. Bij het beantwoorden van de vragen zijn de grondrechten van de mens van toepassing, deze staan vastgelegd in de Grondwet en het Europese Verdrag voor de Rechten van de Mens. In een bijlage van het Impact Assessment Mensenrechten en Algoritmes (IAMA) is een lijst van grondrechtenclusters opgenomen, gebruik deze eventueel als hulpmiddel bij het beantwoorden van onderstaande vragen.

Het AI-systeem kan een lichte, middelzware of zware inbreuk op de grondrechten maken. In het geval van een middelzware of zware inbreuk moet de afweging extra zorgvuldig gemaakt worden en kan het zijn dat er extra maatregelen nodig zijn.

In deel B van het AIIA komen maatregelen voor een aantal grondrechten terug, zoals bijvoorbeeld persoonsgegevensbescherming, recht op toegang tot informatie of een eerlijk proces. Er kunnen ook andere grondrechten zijn die mogelijk impact hebben door het gebruik van het AI-systeem.

Hulp nodig? Deel 4 van het IAMA bevat een stappenplan grondrechten dat extra toelichting geeft op dit onderwerp[2].

1. Wat is de mogelijke impact op de grondrechten van burgers door het gebruik van het AI-systeem?

2. Wat is de (wettelijke) grondslag van de inzet van het AI-systeem en van de beoogde besluiten die genomen worden op basis van het AI-systeem?

3. Welke grondrechtelijke bepalingen zijn mogelijk van toepassing?

4. Op welk van deze grondrechtelijke bepalingen kan mogelijk een inbreuk worden gemaakt bij verkeerde uitvoering van het AI-systeem?

[2] Vraag 1, 2 en 3 vanuit hoofdstuk 4 van het IAMA corresponderen met de vragen in hoofdstuk 2.1, vraag 4 uit het IAMA met hoofdstuk 1.1 en vraag 5, 6 en 7 kunnen gebruikt worden voor het maken van de afweging in hoofdstuk 3.

2.2 Duurzaamheid

Tijdens de gehele levensduur bouwt een **AI-systeem** een ecologische voetafdruk op, zowel door het gebruik van energie, water en grondstoffen . Zo is er continu water nodig om het **AI-systeem** te draaien. Een **AI-systeem** verbruikt tevens energie. Het is belangrijk om bewust de vraag te stellen of het energie gebruik proportioneel is met de probleemstelling en hoe deze energie is opgewekt.

Het is op dit moment lastig om de milieu-impact van een AI-systeem te meten: bedrijven zijn bijvoorbeeld niet transparant over het energieverbruik en onderzoek staat nog in de kinderschoenen. Gebruik deze vraag, zolang er geen goede cijfers zijn over de impact, vooral om na te denken over duurzame keuzes. Soms kan de milieu-impact verkleind worden door gebruik te maken van een andere techniek, infrastructuur of model, met als bijkomend voordeel vaak dat het model sneller wordt.

Over het algemeen kan gesteld worden dat een 'eenvoudiger' model minder energie-intensief is dan een uitgebreider model zoals een **large language model.** Daarnaast is de grootte en soort in- en output van grote invloed is op het energiegebruik (tekst verbruikt bijvoorbeeld minder rekenkracht dan beeld).

Aan de andere kant kan een **AI-systeem** juist worden ingezet om milieuwinst te behalen. Die impact moet afgewogen tegen de milieukosten van bijvoorbeeld het laten draaien van het systeem.

1. Wat is de impact op het milieu door het invoeren van het **AI-systeem** (ontwikkeling, installatie en gebruik), en hoe wordt dit gemeten?

2. Wat voor maatregelen zijn er genomen om de (negatieve) milieu-impact van het AI-systeem te minimaliseren?

2.3 Overige effecten

In dit onderdeel gaat het om de effecten of gevolgen die nog niet eerder genoemd zijn. In deel B van het AIIA komen veel effecten langs, bekijk deze om er zeker van te zijn dat alles meegenomen wordt. In dit hoofdstuk kunnen ook zaken genoemd worden die voor dit project extra van belang zijn, zodat deze voldoende aandacht krijgen.

Hieronder staat een lijst met mogelijke effecten. Het hangt van de scope van het project af in hoeverre onderstaande punten relevant zijn.

Overige effecten om te overwegen:
- Publieke waarden: veranderende, context gebonden ideeën over wat we als samenleving als waardevol ervaren. Meer uitleg en inzicht over publieke waarden is te vinden in de toolbox van BZK voor ethisch verantwoorde innovatie [3]
- Missie en visie van de organisatie: Voor IenW is dit bijvoorbeeld werken aan een mooier, schoon, veilig, duurzaam en bereikbaar Nederland
- Korte en lange termijn
- Effecten voor de individu, organisatie en samenleving
- Positieve en negatieve aspecten van het gebruik van het AI-systeem

Denk hierbij niet alleen aan risico's, maar ook kansen en positieve effecten van het toepassen van het AI-systeem.

1. Hoe draagt het AI-systeem bij aan de missie van de organisatie?

2. Zijn er, naast de vragen hierboven, nog andere effecten (positief, negatief, risico's, voor bepaalde doelgroepen, op verschillende niveau's, brede welvaart) van het AI-systeem die relevant zijn voor de afweging?

[3] https://www.digitaleoverheid.nl/overzicht-van-alle-onderwerpen/nieuwe-technologieen-data-en-ethiek/publieke-waarden/toolbox-voor-ethisch-verantwoorde-innovatie/publieke-waarden-centraal/

3. Afweging voor het wel of niet inzetten van het AI-systeem

In deze stap bepaal je of het zinvol is een AI-systeem in te zetten of niet, waarbij je kijkt naar het doel, de oplossing en de impact van het AI-systeem. Is het proportioneel om een AI-systeem toe te passen? Lost de beoogde toepassing daadwerkelijk het probleem op? Of zijn andere technieken ook passend? Beschrijf duidelijk hoe deze afweging gemaakt is.

Zorg voor een heldere onderbouwing. Probeer dit goed te nuanceren. Houdt hierbij rekening met verschillende groepen betrokkenen, er mogen geen groepen mensen onterecht benadeeld worden. Uiteraard kan de impact soms zo groot zijn dat een afweging niet meer nodig is, bijvoorbeeld als er een schending van grondrechten is of het niet mag volgens wet- en regelgeving.

In deel B staan er maatregelen voor het verantwoord toepassen van het AI-systeem. Het kan zijn dat er extra of specifieke maatregelen nodig zijn, bijvoorbeeld in het proces. Benoem deze hier.

1. Is de impact in verhouding met de beoogde doelen en zijn er geen andere minder ingrijpende manieren om deze doelen te behalen? Oftewel: is het **PROPORTIONEEL** en **SUBSIDIAIR** om dit systeem in te zetten om de gestelde doelen te realiseren?

2. Zijn er aanvullende maatregelen (bv. in processen) die je kunt nemen om het AI-systeem op een verantwoorde manier te gebruiken?

Deel B: Implementatie en gebruik AI-systeem

Deel B van de AIIA gaat over de inrichting, implementatie en het gebruik van het AI-systeem. Het gaat meer de details in en hierbij is vaak meer kennis over AI en techniek nodig. Betrek deze expertise, bijvoorbeeld een data-analist, bij het invullen en gebruik van dit hoofdstuk.

4. Technische robuustheid

Of een **AI-systeem** werkt waarvoor het bedoeld is wordt afgevangen met technische **robuustheid**.

4.1 Bias

Bias betekent het doen van aannames over dingen, mensen of groepen. Dit heeft twee kanten. Enerzijds is het noodzakelijk om conclusies over data op een nieuwe situatie te projecteren. We maken in generalisaties namelijk altijd aannames. Tegelijkertijd is het van belang dat er geen onrechtmatige vertekening ontstaat met vormen van onterechte en onwenselijke bias die in strijd kunnen zijn met de rechten van de mens of regelgeving.

Bias kan overal in het systeem voorkomen. Denk bijvoorbeeld aan: **bias in de input**, **bias in het model** en **bias in de output**. Tijdens het ontwikkelen en inzetten van AI zijn **data bias** en **design bias** belangrijk om rekening mee te houden. Deze soorten bias worden vaak veroorzaakt door socio-economische aannames in de data of van een ontwikkelaar. Het model kan deze aannames versterken. Dat kan betekenen dat een AI-systeem niet voor alle **betrokkenen** goed werkt, indien je niet bewust van de bias bent en deze niet corrigeert.

Het kernelement van dit thema is bewustzijn en integriteit. Het is onmogelijk om volledig zonder bias te werken. Vaak bestaat bias al decennialang en herkennen we deze (onterecht) niet eens. Denk hierbij aan de taalmodellen (ChatGPT, Google Bing, Microsoft Co-pilot): deze zijn waarschijnlijk getraind op informatie van internet en alle aannames die daar al jaren lang in verwerkt zitten. Dus in plaats van ons richten op bias-loze AI, moeten we ernaar streven om ons zoveel mogelijk bewust te zijn van mogelijke discriminatie.

Bias hangt nauw samen met menselijke **diversiteit**, **gelijkheid** en **eerlijkheid**, maar het is van belang om bewust te zijn dat aannames ook over niet-menselijke aspecten kunnen gaan, zoals de natuur of leefomgeving. Daarnaast kan het voor de wijze waarop je bias wilt mitigeren relevant zijn om onderscheid te maken tussen **negatieve impact**, **geen positieve impact** en een **positieve impact** die de bias kan hebben.

1. Hoe wordt rekening gehouden met mogelijk onwenselijke **bias**, bijvoorbeeld **bias in de input**, **bias in het model** en **bias in de output** van het **AI-systeem**?

Bias in de input(data)

2. Is de input(data) relevant en representatief, rekening houdende met het beoogde doel (vraag i1) van het AI-systeem?
3. Worden indien nodig subpopulaties beschermd bij het trekken van steekproeven?
4. Is de keuze voor de inputvariabelen onderbouwd en afgestemd met de **betrokkenen**?

Bias in het model

5. Op welke manier wordt er rekening gehouden met de eis dat er geen onterechte of onrechtvaardige **bias** in een AI-systeem wordt gecreëerd of versterkt?

6. Is het AI-systeem te gebruiken door de beoogde **EINDGEBRUIKERS** (dus ongeacht diens kenmerken zoals leeftijd, geslacht of capaciteit)?

Bias in de output(data)

7. Zijn er stop-, toezicht- of controle- mechanisme ingesteld om te voorkomen dat groepen in de maatschappij disproportioneel getroffen kunnen worden door de negatieve implicaties van het AI-systeem? Specifiek voor ILT: maak hier onderscheid tussen ondertoezichtstaanden (OTS) en de rest van de maatschappij.

4.2 Accuraatheid

Een accuraat **AI-SYSTEEM** presteert goed en geeft een goede beoordeling. Het is belangrijk om doorlopend de prestaties van een AI-systeem te meten (tijdens zowel de ontwikkel- als productiefase). Ook de kwaliteit van de gebruikte trainingsdata is van belang. Een AI-systeem is nooit af; het blijft noodzakelijk AI-systemen regelmatig te testen en hertrainen. Het is wenselijk een kwantificering te hebben van de kans dat er een verkeerde beoordeling wordt gemaakt.

ACCURAATHEID van het systeem kan je bepalen door vooraf **ACCEPTATIECRITERIA** op te stellen voor zowel de (trainings)data als het systeem. Acceptatiecriteria kunnen bijvoorbeeld een minimale hoeveelheid data zijn of bepaalde drempelwaardes van het meetsysteem. Er zijn veel verschillende soorten meetsystemen (vaak 'performance metrics' genoemd door data scientists) beschikbaar om de kwaliteit van **MODELLEN** te kwantificeren, denk bijvoorbeeld aan een accuratesse, precision en recall of F1-score.

Het is van belang dat het meetsysteem en de acceptatiecriteria goed worden afgestemd op de data en het beoogde doel van het AI-systeem[4]. Dit moet samenhangen met onder andere de bevindingen uit de risicoanalyse (zie 'Risicobeheer'), omdat na verloop van tijd nieuwe risico's kunnen ontstaan bij de inzet van een AI-systeem. Ook is het van belang om de kwaliteit van het systeem doorlopend te monitoren. Evalueer tijdens het hertrainen of doorontwikkelen de acceptatiecriteria en keuze voor meetsystemen opnieuw.

[4] Het gekozen meetsysteem moet geschikt zijn voor het model en de data die gebruikt wordt om de kwaliteit te meten. Neem bijvoorbeeld één dat in een bepaalde tekst 5 woorden per 100 woorden zou moeten labelen. Als het systeem in deze tekst 0 woorden labelt, dan heeft het model een accuratesse van 95%. Op het moment dat je de kwaliteit van het systeem bepaalt met accuratesse lijkt het model het dus heel erg goed te doen, terwijl de recall 0 is en het dus helemaal niet zo heel goed doet. Daarom is accuratesse niet geschikt om te bepalen hoe goed dit model werkt.

1. Hoe wordt de doorlopende accuraatheid van het systeem gemeten en gewaarborgd?

2. Wat zijn de opgezette **ACCEPTATIECRITERIA** om de kwaliteit van de **INPUT(DATA)** en **OUTPUT(DATA)** van het **MODEL** aan te toetsen?

3. Passen de **ACCEPTATIECRITERIA** bij de data en het doel van het AI-systeem ?

4. Hoe wordt de **OUTPUT(DATA)** (periodiek) steekproefsgewijs en doorlopend gemonitord op juistheid?

5. Hoe worden afwijkingen in de output(data) ten opzichte van acceptatiecriteria tijdig geanalyseerd en gecorrigeerd?

6. Wat zijn de resultaten als er alternatieve **MODELLEN** zouden worden ingezet?

4.3 Betrouwbaarheid

Een **BETROUWBAAR** AI-systeem geeft in vergelijkbare gevallen dezelfde resultaten. De vraag die centraal staat bij betrouwbaarheid is of de individuele **OUTPUT(DATA)** nogmaals te verkrijgen is met behulp van hetzelfde **MODEL** en dezelfde **INPUT(DATA)**, dezelfde instellingen en dezelfde **PARAMETERS**. Ook is het van belang dat het systeem een betrouwbare indicatie geeft van hoe goed het model gaat presteren in nieuwe situaties.

1. Is het AI-systeem betrouwbaar?

2. Wat zijn de belangrijkste factoren die de prestaties van het AI-systeem beïnvloeden?

3. Wordt een deel van de (sub)dataset uitgesloten voor het leren van het model en alleen gebruikt voor het bepalen van de betrouwbaarheid of wordt de betrouwbaarheid van het model berekend met behulp van cross-validatie?

4. Hoe is de (hyper)parameter-tuning onderbouwd en getoetst?

4.4 Technische implementatie

De **technische implementatie** beschrijft hoe het AI-systeem technisch binnen het ICT-landschap van de organisatie is geïntegreerd. De specifieke eisen van het AI-systeem aan hard- en software zijn gedocumenteerd zodat hier rekening mee gehouden kan worden bij de uitrol en beheer van het systeem. Daarnaast wordt uit de systeemarchitectuur duidelijk hoe de verschillende softwarecomponenten zich tot elkaar verhouden. Een goed doordachte architectuur vermindert de bedrijfsrisico's die gepaard gaan met het bouwen van een technische oplossing en slaat een brug tussen bedrijfs- en technische vereisten. Vaak is er al bestaande documentatie, verwijs daar dan vooral naar. Bekijk ook de Infrastructuur en Waterstaat Enterprise Architectuur (IWEA).

1. Hoe is het AI-systeem technisch geïmplementeerd?

2. Is er nagedacht hoe het AI-systeem past in de al bestaande technische- en systeeminfrastructuur en zijn hier passende maatregelen voor genomen om deze uit te rollen (indien van toepassing)?
3. Hoe ziet de systeemarchitectuur eruit (hoe verhouden de softwarecomponenten zicht tot elkaar)?
4. Zijn eventuele specifieke hardware- en software-eisen gedocumenteerd?
5. Indien de applicatie extern wordt gehost, onder welke voorwaarden gebeurt dit?
6. Hoe is de toegang tot het AI-systeem en diens componenten ingericht? (Denk aan de Generieke IT-beheersmaatregelen).
7. Hoe kan het AI-systeem interageren met andere hardware of software (indien van toepassing)?
8. Hoe is de logging en monitoring ingericht?

4.5 Reproduceerbaarheid

Reproduceerbaarheid gaat over trainen, valideren en testen. Bij reproduceerbaarheid gaat bijvoorbeeld over het vastleggen van de gebruikte data, totstandkoming van het model, bijhouden van wijzigingen in de data, of uit dezelfde **input(data)** dezelfde resultaten voortvloeien, en of er bepaalde situaties of condities zijn waarin de **output(data)** beïnvloed kunnen worden.

Reproduceerbaarheid hangt nauw samen met **traceerbaarheid**. Bij traceerbaarheid gaat het er voornamelijk om dat de datasets en processen goed worden gedocumenteerd. Versiebeheer op de data, het model en de training van het model speelt daarin een belangrijke rol.

1. Is het **AI-systeem reproduceerbaar**? Is er een proces ingesteld om dit te meten?

2. Kan verkregen **output(data)** nu of in de toekomst gereconstrueerd worden (dus bijvoorbeeld zijn oude versies van het **model**, datasets en omstandigheden opgeslagen middels versiebeheer)?
3. Is het mogelijk om gegeven de **parameters** en een vaste **seed** het model te reconstrueren?
4. Is het **AI-systeem** aan de hand van documentatie op hoofdlijnen te reproduceren?
5. Hoe worden de wijzigingen tijdens de levensduur van het systeem gedocumenteerd?

4.6 Uitlegbaarheid

Technische **UITLEGBAARHEID** heeft te maken met het vermogen om zowel technische processen als daaraan gerelateerde menselijke beslissingen te kunnen begrijpen. Verder moet helder zijn welke verschillende ontwerpkeuzes zijn gemaakt en wat de rationale is voor het inzetten van het **AI-SYSTEEM**. Zie ook 'Verantwoordingsplicht' voor uitlegbaarheid, transparantie en communicatie richting gebruikers en andere **BETROKKENEN**.

1. Is het **AI-SYSTEEM** voldoende **UITLEGBAAR** en te interpreteren voor de **ONTWIKKELAARS**?

1. Hoe is bij het ontwikkelen van het AI-systeem rekening gehouden met de uitlegbaarheid van het model, bijvoorbeeld voor de gebruikers?

2. Welke technieken zijn gebruikt om het AI-systeem uitlegbaar te maken en waarom is voor deze techniek gekozen?

5. Data governance

Data **GOVERNANCE** gaat over een werkwijze rondom data met betrekking tot toegang, eigenaarschap, bruikbaarheid, integriteit en veiligheid. Daarnaast is er aandacht voor de kwaliteit van de data die wordt gebruikt.

Onder data governance valt ook privacy. Privacy is een van de fundamentele rechten van de mens, die mogelijk door AI kan worden aangetast. Het is daarom belangrijk dat er adequate data governance en bescherming van persoonsgegevens is, minimaal conform de Algemene Verordening Gegevensbescherming (**AVG**) en het privacybeleid IenW.

5.1 Kwaliteit en integriteit van data

Datakwaliteit is essentieel voor de goede werking van een **AI-SYSTEEM**. De term 'garbage in = garbage out' bestaat niet voor niets. Daarnaast kunnen verzamelde gegevens bijvoorbeeld sociaal geconstrueerde **bias**, onjuistheden, fouten en vergissingen bevatten (zie ook 'Bias'). Dit moet worden geadresseerd voordat er verder met deze data wordt gewerkt.

Gebruik hiervoor de FAIR-principes (Vindbaar, Toegankelijk, Interoperabel en Herbruikbaar) en het nauw samenhangende FACT (Eerlijk, Nauwkeurig, Vertrouwelijk en Transparant) als inspiratie. Deze principes worden binnen de AVG gebruikt. De datasets en de werkwijze moeten worden getest en gedocumenteerd bij iedere stap: training, testen, uitrolfase en operationele fase. Dit geldt ook voor AI-systemen die niet intern gebouwd zijn, maar zijn verworven.

Datakwaliteit is des te meer van belang wanneer gebruik wordt gemaakt van persoonsgegevens. Volgens de AVG is het essentieel dat de persoonsgegevens juist zijn en zo nodig worden geactualiseerd.[5] Ook dienen de gegevens noodzakelijk te zijn voor het doel van de analyse. Beschrijf daarom ook het doel van de analyse en hoe je omgaat met de data en fouten in de datasets en uitkomsten.[6]

De 'Richtlijnen voor het toepassen van algoritmen' van J&V[7] maakt data-kwaliteit expliciet op een heel operationeel niveau.

1. Welke trainingsdata wordt gebruikt als input voor het algoritme en uit welke bronnen is de data afkomstig?

2. Hoe wordt de kwaliteit van de data gewaarborgd?

[5] Artikel 5, eerste lid, onder d, AVG. Zie ook Afdeling 3 Rectificatie en wissing van gegevens, AVG.
[6] Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses, JNV, p.20.
[7] https://open.overheid.nl/documenten/ronl-1411e45f-b822-49fa-9895-2d76e663787b/pdf

Overkoepelend

3. Is de gebruikte data noodzakelijk voor het **AI-SYSTEEM**?

4. Hoe voorkom je onbedoelde verdubbelingen van data?

5. Is het mogelijk om de trainings- en testgegevens te actualiseren als de situatie daar om vraagt? Wanneer besluit je het AI-systeem te her-trainen, tijdelijk stop te zetten, of door te ontwikkelen?[8]

6. Voldoet de data aan de aannames van het **MODEL**?

7. Op welke manier is de **INPUT(DATA)** die wordt gebruikt in het AI-systeem verzameld en samengevoegd?

8. Hoe wordt de data gelabeld?

9. Welke factoren (denk aan beperkingen in de verzamelmethode, de opslag) hebben invloed op de kwaliteit van de input(data)? En wat kan je daaraan doen? En wat kan je daaraan doen?

10. Is de input(data) getoetst op veranderingen die zich voordoen tijdens trainen, testen en evalueren? Ook door de tijd heen tijdens het gebruik van het algoritme?

Output(data)

11. Indien output(data) wordt gebruikt als nieuwe input, hoe wordt de output(data) opgeslagen en gecontroleerd op juistheid en volledigheid?

12. Hoe zorg je ervoor dat de output(data) tijdig beschikbaar is?

[8] Artikel 1 lid 4.

5.2 Privacy en vertrouwelijkheid

Bij het ontwerp van een AI-systeem moet rekening worden gehouden met de privacywetgeving. Het is immers makkelijker om het vooraf op de juiste manier in te richten dan op een later moment te repareren. Uiteraard moet privacy gewaarborgd zijn gedurende de hele levenscyclus van het **AI-systeem**.

Bij de verwerking van persoonsgegevens moet er een pre-scan **Data Protection Impact Assessment** (DPIA) worden ingevuld om te bepalen of de volledige DPIA ingevuld moet worden.

Naast persoonsgegevens kunnen er ook andere vertrouwelijke gegevens gebruikt worden, die niet zomaar openbaar gemaakt mogen worden. Dit geldt bijvoorbeeld voor het gebruik van vertrouwelijke informatie zoals gerubriceerde informatie of bedrijfsgeheimen. Ook deze data moet goed beschermd zijn. De **AI- verordening** biedt, bovenop de **AVG**, aanvullende regels voor het gebruik van (persoons-) gegevens in AI-systemen. Het is van belang dat de vertrouwelijke gegevens voldoende beveiligd worden (zie Risicobeheer).

1. Hoe wordt er omgegaan met persoonsgegevens of vertrouwelijke gegevens?

Met betrekking tot persoonsgegevens

2. Werkt het **AI-systeem** met persoonsgegevens (is de AVG van toepassing)? Zo ja, vul de volgende vragen ook in. Zo nee, ga verder bij ‘met betrekking tot vertrouwelijke gegevens’.
3. Is de verwerking van de persoonsgegevens proportioneel en subsidiair? (Gebruik hiervoor de afweging in hoofdstuk 3 als basis)
4. Is de output van het AI-systeem tot personen direct of indirect te herleiden (is de AVG van toepassing)?
5. Zijn functionarissen betrokken, zoals de Chief Privacy Officer, informatiebeveiliger, Chief Information Officer, privacy officer etc.?
6. Hoe vaak wordt de kwaliteit en de noodzakelijkheid van de verwerking van persoonsgegevens geëvalueerd?

Met betrekking tot vertrouwelijke gegevens (niet zijnde persoonsgegevens)

7. Worden vertrouwelijke gegevens gebruikt of opgeslagen?
8. Hoe wordt de veiligheid van deze informatie gewaarborgd?

6. Risicobeheer

Het is van belang dat risico's in de gaten worden gehouden. Wanneer risico's niet zijn voorzien, kan een **AI-systeem** tot onbetrouwbare resultaten komen. Dit kan schade veroorzaken, bijvoorbeeld door slecht functioneren van het AI-systeem, of bijvoorbeeld door **hackaanvallen**.

6.1 Risicobeheersing

Bij het ontwikkelen en **in gebruik nemen** van een **AI-systeem** komen gevaren kijken, die in deze AIIA zoveel mogelijk ingekaderd worden. Toch kunnen zich alsnog onvoorziene problemen voordoen. Het is belangrijk om vast te stellen hoe je met deze potentiële gevaren omgaat. Er moeten mechanismes zijn om risico's te beheersen, en deze mechanismen moeten zijn toetst. Denk aan het voorkomen van datavergiftiging, de mate van beheersmaatregelen en de beveiliging van de bewaarplaats van uitkomsten. Daarnaast kunnen er risico's ontstaan na invoering van het AI-systeem. Controleer de risicobeheersmaatregelen periodiek, maar minimaal elke 3 jaar of bij grote wijzigingen.

Voor hoog-risico-AI-systemen is een risicobeheerssysteem verplicht. Daarvoor vindt je in bijlage 1 additionele vragen. Voor andere systemen is een eenmalige risicoanalyse voldoende.

1. Hoe is het AI-systeem getest op de passende en gerichte risicobeheersmaatregelen?

6.2 Alternatieve werkwijze

Het is wenselijk om een plan te hebben voor wanneer er problemen optreden met het **AI-systeem**. Dit betekent dat er een alternatieve werkwijze beschikbaar moet zijn. Denk aan de mogelijkheid om van een machine learning model naar een beperkter rule-based **model** terug te schakelen. Of om zelfs nog een stap terug te doen door het proces handmatig uit te voeren. Als het acceptabel is dat het systeem tijdelijk niet beschikbaar is, is geen plan uiteraard ook een optie.

Door het gebruik van een AI-systeem kan het gebeuren dat bepaalde menselijke vaardigheden minder worden. Denk bijvoorbeeld aan het effect van de rekenmachine op onze vaardigheid hoofdrekenen. Een alternatieve werkwijze moet daarom niet ineens gebruik maken van deze vaardigheid.

1. Wat is het plan als er problemen met de werking van het **AI-systeem** zijn?

2. Wat is de impact als het systeem uitvalt?

3. Zie hierboven het voorbeeld over de rekenmachine. Wat is een equivalent effect wat kan optreden als het **AI-systeem** in gebruik wordt genomen, en is dit wenselijk?

6.3 Informatiebeveiligingsrisico's

Een AI-systeem moet zoveel mogelijk *secure* **BY DESIGN** zijn gebouwd, waarbij in de ontwerpfase al over security is nagedacht. Informatiebeveiligingsrisico's zoals manipulatie van het model, ongewenste toegang of **HACKAANVALLEN** worden zo beheerst. Maak een overzicht van de voorziene risico's via het risicomanagementproces van de organisatie. Daaronder vallen onder meer: BIV- classificatie, rubriceringsniveau van informatie, implementatie van BIO-maatregelen, beveiligingstesten en, indien het beveiligingsniveau van de BIO niet voldoet, het eventueel uitvoeren van een aanvullende (technische) risicoanalyse.

Goed ingerichte autorisaties en een solide wijzigingenproces zijn hierbij essentieel. Daarnaast is het van belang om te kijken of fouten en onregelmatigheden te detecteren en technisch af te vangen zijn. Meer praktische handvatten zijn te vinden in het Toetsingskader Algoritmes van de Rekenkamer[9] of onderzoekskader algoritmes ADR.[10] De AIVD heeft een handleiding over het veilig ontwikkelen van AI-systemen[11]. Daarnaast heeft Open Worldwide Application Security Project (OWASP) veel resources over het veilig toepassen van AI.

1. Op welke manier worden informatiebeveiligingsrisico's inzichtelijk gemaakt, teruggebracht naar een acceptabel niveau en (technisch) getest?

2. Hoe wordt er voorkomen dat ongeautoriseerde derden gebruik, maken van de kwetsbaarheden van het AI-systeem?

3. Wat is de impact als derden ongewenst toegang hebben tot de broncode, data of uitkomsten van het AI- systeem?

4. Kunnen mensen misbruik maken van het feit dat er een AI-systeem wordt ingezet in plaats van een menselijke beslissing?

5. Hoe wordt er geregistreerd wie er gebruik maakt van het AI-systeem en hoe lang?

6. Zijn er buiten de standaard beveiligingsmaatregelingen van IenW extra maatregelen genomen om het AI-systeem te beveiligen?

[9] https://www.rekenkamer.nl/onderwerpen/algoritmes/toetsingskader
[10] https://www.rijksoverheid.nl/documenten/rapporten/2023/07/11/onderzoekskader-algoritmes-adr-2023
[11] https://www.aivd.nl/documenten/publicaties/2023/02/15/ai-systemen-ontwikkel-ze-veilig

7. Verantwoordingsplicht

De Rijksoverheid moet verantwoording afleggen binnen de organisatie, aan de Tweede Kamer en de samenleving. De techniek wordt steeds vaker toegepast, maar er zijn ook zorgen over de inzet van AI. Om de inzet en resultaten van AI-systemen te kunnen verantwoorden, is het cruciaal om hier een proces voor in te richten.

7.1 Transparantie richting gebruikers

Eindgebruikers moeten inzicht krijgen in de werking van een AI-systeem (naast de uitlegbaarheid voor de ontwikkelaars, zie 'Technische Robuustheid'). Dit geldt met name voor medewerkers die in hun proces gebruik maken van AI. Het is niet nodig dat ze het AI-systeem compleet begrijpen. Op hoofdlijnen moet de werking van een AI-systeem en beperkingen ervan duidelijk zijn.

Indien je een beslissing moet nemen over AI-gebruik is het, naast begrip van de werking en beperkingen, essentieel dat je een betekenisvolle mate van invloed hebt op de beslissing.

1. Op welke manier geef je eindgebruikers inzicht in de werking en beperkingen van het AI-systeem? En blijven deze voldoende onder de aandacht zolang ze bestaan?
2. Welke rol spelen mensen bij het nemen van beslissingen op basis van input van het AI-systeem ('human in the loop') en hoe worden zij in staat gesteld om die rol te spelen?
3. Hoe is het systeem voor iedereen te monitoren en begrijpen (menselijk toezicht)?

7.2 Communicatie naar betrokkenen

Deze sectie gaat over twee vormen van communicatie naar de **EINDGEBRUIKERS**. Ten eerste, eindgebruikers moeten weten dat ze met de resultaten van een **AI-SYSTEEM** te maken hebben (en geen mens bijvoorbeeld). Ten tweede, eindgebruikers hebben te allen tijde het recht om te weten hoe een **ALGORITME** de uitkomsten van een AI-systeem bepaalt. Dat betekent ook dat het doel en beperkingen van het AI-systeem duidelijk, eerlijk en transparant moeten worden gecommuniceerd. Zowel technische processen als daaraan gerelateerde menselijke beslissingen moeten begrijpelijk zijn, opgevraagd kunnen worden en indien nodig gecorrigeerd. Wijs bijvoorbeeld een contactpersoon aan met inhoudelijke kennis over het AI-systeem. Gezien het zelflerende karakter van AI kan de werking van het systeem niet altijd 100% te herleiden zijn. Wel moet in ieder geval mogelijk zijn om gepaste uitleg te geven over het proces aan eindgebruikers.

Daarnaast moeten burgers informatie kunnen opvragen over het AI-systeem of zich beroepen op de rechten vanuit de AVG/Wpg. Men moet in staat zijn om resultaten van het AI-systeem te kunnen betwisten. Dat betekent ook dat de data, en de omstandigheden waarin de data ter beschikking is gesteld, bewaard moeten worden (zie Archivering).

1. In hoeverre ben je transparant richting verschillende groepen betrokkenen over het AI-systeem en op welke wijze?
2. Worden er mechanismes ingesteld waarin eindgebruikers opmerkingen over het systeem (data, techniek, doelgroep, etc.) kunnen maken? En hoe of wanneer worden deze meldingen gewaarborgd (geanalyseerd en gevolgd)?
3. Moet het systeem onder invloed van de AI Act in het algoritmeregister en/of (voor hoog-risico-toepassingen) in de EU-databank?

4. Wordt er aan de **EINDGEBRUIKER** en **BETROKKENEN** van het AI-systeem gecommuniceerd dat de resultaten gegenereerd worden door een AI-systeem en wat dat voor hen betekent?

5. Is er een handleiding opgesteld?

6. Wat zijn de potentiële (psychologische) bijwerkingen, zoals het risico op verwarring, voorkeur of cognitieve vermoeidheid van de **EINDGEBRUIKER** bij het gebruik maken van het AI-systeem?

7. Op welke manier krijgen verschillende groepen **BETROKKENEN** (burgers, collega's, managers, etc.) inzicht in verschillende aspecten van het **AI-SYSTEEM**? Denk hierbij bijvoorbeeld aan de gebruikte data, model of uitkomsten.

8. Hoe heb je invulling gegeven aan de uitlegbaarheid specifiek richting de **EINDGEBRUIKER**?

9. Is het systeem voldoende **TRANSPARANT** om **gebruiksverantwoordelijken** in staat te stellen de output(data) van het systeem te interpreteren en op passende wijze te gebruiken?

10. Is er iets ingericht om eindgebruikers eventueel bij te scholen?

11. Hoe wordt ervoor gezorgd dat commentaar van betrokkenen en eindgebruikers intern goed wordt behandeld?

12. Als een betrokkene bezwaar wil aantekenen, of een klacht wil indienen tegen een besluit van het AI-systeem, is het dan duidelijk welke stappen hij/zij kan nemen? Hetzelfde geldt voor beroep instellen.

7.3 Controleerbaarheid

Met controleerbaarheid kijken we naar de manier waarop de data, het **MODEL** en de resultaten worden geëvalueerd. Deze controle kan in de vorm van interne of externe audits plaatsvinden. Bij toepassing van het AI-systeem in risicovollere gebieden zijn er strengere eisen.

Het is van belang dat er inzicht in de bronnen, het systeem en de uitkomst is. Deze **VERANTWOORDELIJKHEID** ligt bij de **eigenaar van het systeem**.

1. Hoe en door wie wordt het **AI-systeem** gecontroleerd?

2. Op welke manier wordt verantwoording afgelegd over het AI-systeem?

3. Wie verzorgt de onafhankelijke controle van het AI-systeem? En op welke wijze?

7.4 Archivering

Archivering is het bewaren van informatie om deze in de toekomst te kunnen hergebruiken. Denk bijvoorbeeld aan het herconstrueren van het model (zie 'Reproduceerbaarheid'), of een nieuwe medewerker uitleggen hoe het systeem werkt (zie 'Uitlegbaarheid'), of om verantwoording af te leggen aan een **betrokkene** (zie 'Verantwoordingsplicht'). Archivering is ook belangrijk om te voldoen aan wet- en regelgeving. Zo bestaan er minimale bewaartermijnen voor hoog-risicotoepassingen van logs in een AI-systeem en deze staan ook beschreven in de **selectielijsten** van IenW.

Input(data)

1. Hoe wordt de **INPUT(DATA)** opgeslagen?
2. Wat is de bewaartermijn van de input(data)?

Model

3. Hoe wordt het **MODEL** opgeslagen?
4. Hoe is het versiebeheer geregeld?

Output(data)

5. Wat is de bewaartermijn van de output(data)?

Begrippenlijst

In dit document worden begrippen gebruikt die in de literatuur vaak verschillend gedefinieerd worden. Hieronder volgen eenduidige definities die gebruikt worden in dit document.

ACCEPTATIECRITERIA	Op het beoogde doel en data afgestemde voorwaarden, waaraan het **AI-SYSTEEM** moet voldoen. Dit kan bijvoorbeeld de hoeveelheid data zijn, een accuratesse maatstaaf voor de **OUTPUT(DATA)** of het inrichten van een onafhankelijke controle van output. Acceptatiecriteria moeten waar mogelijk meetbaar gemaakt worden zodat deze gemonitord kunnen worden met een geschikt meetsysteem. Goede acceptatiecriteria zijn SMART en voldoende verschillend zodat alle relevante aspecten van het AI-systeem goed gemonitord worden.
ACCURAATHEID	Zeer nauwgezet, precies of zorgvuldig; als een systeem in staat is om juiste én accuratesse beoordelingen te maken. In een formule: TP+TN/(TP+TN+FP+FN). TP= werkelijke positief, TN=Werkelijk negatief, FP=Verkeerde positief, FN= Verkeerd negatief. Hoe meer werkelijke resultaten t.o.v. verkeerde resultaten hoe hoger de accuraatheid.
AI MET BEPERKT RISICO	De AI-verordening stelt vast wat beperkt risico AI is. AI ingericht op interactie met mensen, emoties herkennen, of gemanipuleerde beelden produceren. Denk aan spamfilters, het samenvatten van teksten, het classificeren van onderwerpen van luchtvaartvoorvallen, of bijvoorbeeld AI-systemen die kantoorverlichting regelen.
AI MET HOOG RISICO	De AI-verordening stelt vast wat hoog-risico-AI is. Dit zijn vaak producten die nauw te maken hebben met fundamentele rechten en/of productveiligheid. Denk hierbij bijvoorbeeld aan AI in vliegtuigen, vaartuigen, voertuigen, rails, wegverkeer, vliegnavigatie en drinkwatertoevoer. Altijd: een AI-systeem dat individuen profileert d.m.v. het verwerken van persoonlijke data. Daarnaast alle AI-systemen op het gebied van: Niet-verboden biometrische gegevens; kritieke infrastructuur; onderwijs en beroepsopleiding; werkgelegenheid; toegang tot en gebruik van essentiële openbare en particuliere diensten; rechtshandhaving; migratie, asiel en grenscontroles; Rechtsbedeling en democratische processen.
AI MET MINIMAAL RISICO	Alle AI-systemen die niet verboden zijn of onder **AI MET HOOG RISICO** of **AI MET BEPERKT RISICO** vallen.
AI-ACT	Zie AI-verordening.
AI-SYSTEEM	'AI-systeem' is 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, inhoud, aanbevelingen of beslissingen die van invloed kunnen zijn op fysieke of virtuele omgevingen.
AI-VERORDENING	Een Europese wet die regels stelt aan de ontwikkeling en het gebruik van AI-systemen.

ALGORITME	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.
ARTIFICIAL INTELLIGENCE	AI kent geen eenduidige definitie. Wij hanteren de omschrijving van AI door de Algemene Rekenkamer: 'het vermogen [..] om externe gegevens correct te interpreteren, om te leren van deze gegevens en om deze lessen te gebruiken om specifieke doelen en taken te verwezenlijken via flexibele aanpassing'. Ook belichten wij graag al die van de Europese Commissie alhoewel deze nog niet gehanteerd wordt in dit document: AI omvat systemen die intelligent gedrag vertonen door hun omgeving te analyseren en – met een zekere mate van zelfstandigheid – actie ondernemen om specifieke doelen te bereiken.
AVG	Algemene verordening gegevensbescherming. Deze privacywet regelt dat bedrijven en organisaties persoonsgegevens zorgvuldig beschermen.
BEHEERORGANISATIE	Een organisatie die applicatiebeheer van het **AI-SYSTEEM** inricht en optimaliseert.
BELANGENGROEP	Samenstelling van **STAKEHOLDERS** om **DIVERSITEIT** te meten. Dit kan zowel een groep van **EINDGEBRUIKERS** zijn als een groep van mensen die impact ervaren door het systeem.
BETROKKENE	Natuurlijk persoon of organisatie die bij het gebruik of de uitkomsten van het systeem belang heeft, of belang denkt te hebben. Hier wordt bewust niet het woord 'belanghebbende' gebruikt, omdat het meer omvat dan het in het bestuursrecht gedefinieerde 'belanghebbende'. Denk aan burgers, een onder toezicht staande, maar ook de **EINDGEBRUIKER** zelf.
BETROUWBAAR BIAS	De eigenschap beschikken van consistent gedrag en consistente resultaten. Vooringenomenheid. Het doen van aannames over dingen, mensen of groepen die vaak niet gebaseerd zijn op werkelijke metingen.
BIAS IN DE INPUT	Kwaliteit, consistentie en integriteit van data is een belangrijke voorwaarde voor een unbiased analyse.
BIAS IN DE OUTPUT	De manier waarop de **OUTPUT**(**DATA**) wordt gebruikt kan invloed hebben op de levens van mensen. Het is belangrijk dat hierbij geen onterechte correlatie gaat leiden tot causaliteit.
BIAS IN HET MODEL	Hoe correct zijn de **MODELLEN**; in hoeverre corrigeren ze voor bekende gebreken in representativiteit van de data? Dit kan bijvoorbeeld ook gaan over wat het **AI-SYSTEEM** leert en wat ongewenste leereffecten zijn.
BIO	Baseline Informatiebeveiliging Overheid, het basisnormenkader voor informatiebeveiliging binnen alle overheidslagen.
BY DESIGN	Bij het ontwerp al rekening houden met de geldende AI, privacy of security-wetgeving. Denk aan AI, privacy en security by design.
CIO	Chief Information Officer.
CISO	Chief Information Security Officer.

CORRUPTIE	Het misbruiken of uitbuiten van fouten van het systeem, of het uitbuiten van ogenschijnlijke neutrale eigenschappen van het systeem.25 We maken onderscheid met **ONBEDOELDE CORRUPTIE**.
DATA BIAS	Wanneer de steekproef niet representatief is voor de gehele populatie. data pipeline.
DATA PIPELINE	Hoe de data vanuit het veld naar het model komt; het proces dat de data doorloopt.
DESIGN BIAS	Problemen in het technisch ontwerp, inclusief beperkingen van computerhulpmiddelen zoals hardware en software.
DIVERSITEIT	Hieronder verstaan we het herkennen van verschillende typen 'subjecten' in onze analyses. Wij proberen hierbij te voorkomen dat groepen van relevante subjecten onterecht niet worden meegenomen in ontwikkelen van een **AI-SYSTEEM**, waardoor het systeem niet op hen aansluit.
DOMEINEXPERT	Iemand die veel kennis heeft over het probleemgebied waarin het **AI-SYSTEEM** gebouwd wordt.
DPIA	Data Protection Impact Assessment. Met deze tool breng je de privacyrisico's van een gegevensverwerking in kaart.
EERLIJKHEID	Als niet elk subject een gelijke behandeling krijgt, moet dat verklaard kunnen worden. Hierbij is het van belang dat we zoveel mogelijk onderscheidende subjectkenmerken in beeld hebben. Zowel om aan te kunnen tonen welke kenmerken daadwerkelijk een rol spelen (en partij A een lager risico toebedelen dan partij B) en welke kenmerken dit juist niet zijn (waardoor partij A en B een onderbouwd gelijkwaardig risico hebben).
EINDGEBRUIKER	Eindgebruikers zijn de personen die het **AI-SYSTEEM** in de praktijk toepassen binnen de organisatie. Het gaat hierbij om een natuurlijk persoon. Wie zitten er met de handen aan de knoppen? Wie vergaart binnen de organisatie informatie uit het AI-systeem? Denk aan een inspecteur of een wegverkeersleider.
EINDVERANTWOORDELIJKE	Een rol binnen de organisatie die de **VERANTWOORDELIJKHEID** over het **AI-SYSTEEM** draagt. Dat betekent bijvoorbeeld de verantwoordelijkheid over dat de juiste resultaten van het AI-systeem bereikt worden. Dit is meestal de proceseigenaar.
ENTITEIT	Een functie binnen een afdeling van een organisatie.
GEBRUIKER	Volgens de AI-verordening "een (…) overheidsinstantie, agentschap of ander orgaan die/dat een **AI-SYSTEEM** onder eigen verantwoordelijkheid gebruikt (…)". De gebruiker zet het systeem in. Dit is nooit een natuurlijk persoon.Bijvoorbeeld de ILT of RWS.
GEEN POSITIEVE IMPACT	**BETROKKENEN** die niet per definitie negatieve impact ondervinden van de inzet van het AI-systeem, maar bijvoorbeeld in dezelfde situatie blijven als daarvoor. Daarbij kan een gevaar zijn dat deze betrokkenen niet dezelfde 'positieve impact' van het AI-systeem ervaren dat andere betrokkenen wel krijgen.

GEEN RISICO AI-SYSTEEM	Alle Ai die niet in de andere categorieën valt. Deze systemen zijn niet gereguleerd door de AI-verordening.
GELIJKHEID	Hieronder verstaan we de gedachte dat elk gelijksoortig subject een gelijke behandeling krijgt.
GENERATIEVE AI	Een specifieke vorm van AI waarbij algoritmes worden ingezet om content te genereren. Met een eenvoudige opdracht, een 'prompt', kunnen gebruikers in een handomdraai tekst, beeld, geluid of computercode genereren. Het bekendste voorbeeld hiervan is ChatGPT.
GOVERNANCE	De handeling of de wijze van besturen, de gedragscode en het toezicht op organisaties. Het betreft beslissingen die verwachtingen bepalen, macht verlenen of prestaties verifiëren. Het bestaat ofwel uit een afzonderlijk proces ofwel uit een specifiek deel van management- of leiderschapsprocessen.
HACKAANVAL	Inbreken in het **AI-SYSTEEM**. Met als gevolg bijvoorbeeld vervuiling van data, het ongewenst uitlekken van (de werking van) een AI-systeem, of aantasting van software of hardware.
IN GEBRUIK NEMEN	Het moment dat een AI-systeem 'in gebruik wordt genomen' betekent het moment waarop deze voor het eerst wordt ingezet in een proces. Hier gaat een externe test of pilot aan vooraf. Op het moment van in gebruik name moet de AIIA af zijn.
INPUT(DATA)	Die gegevens welke worden verwerkt met een vooropgesteld doel. In de context van een **AI-SYSTEEM** kunnen dit ruwe data zijn, bijvoorbeeld de waarnemingen uit de werkelijkheid. In de context van het **MODEL** zijn dit normaal gesproken voorbewerkte data.
LAAG RISICO AI-SYSTEEM	AI-systemen met een risico op manipulatie of bedrog. Deze AI- praktijk systemen moeten transparant zijn en gebruikers moeten worden geïnformeerd over hun interactie met de AI.
LARGE LANGUAGE MODEL	Een LLM is een generatieve AI die tekst kan generen. Het is getraind op zeer grote datasets en bevat veel parameters.
METADATA	Gegevens die de eigenschappen van andere gegevens beschrijven. Bijvoorbeeld van wie de gegevens zijn, of wie ze verstuurd heeft, of wanneer ze voor het laatst gewijzigd zijn.
MODEL	Een (versimpelde) wiskundige vertegenwoordiging van de werkelijkheid, welke wordt gebruikt om informatie te verwerken. In een **AI-SYSTEEM** wordt de wiskundige vertegenwoordiging vaak deels of in zijn geheel volgens een **ALGORITME** 'geleerd', waardoor zelfs door de **ONTWIKKELAARS** niet volledig uit te leggen is hoe het model aan diens uitkomsten komt.
MOREEL BERAAD	Overlegorgaan Fysieke Leefomgeving (mei 2021), Moreel Beraad.
NEGATIEVE IMPACT	**BETROKKENEN** die nadelige gevolgen ondervinden door de toepassing van het **AI-SYSTEEM**, bijvoorbeeld omdat ze gediscrimineerd worden op basis van een **bias** in het AI-systeem.

ONAANVAARDBAAR RISICO AI-SYSTEEM	AI-systemen die: onderbewuste, manipulatieve of bedrieglijke technieken hanteren; misbruik maken van kwetsbaarheden, biometrisch categoriseren; social scoring hanteren; criminialiteitpotentie van individuen voorspellen; databases voor gezichtsherkenning samenstellen; emoties afleiden; of live op afstand emoties herkennen.
ONBEDOELDE CORRUPTIE	Zonder kwaadwillende bedoelingen de werking van het **AI-SYSTEEM** beïnvloeden door bijvoorbeeld verkeerde input te voeden, of de verkeerde knoppen in te drukken. Onbedoelde corruptie valt onder **betrouwbaarheid**. We maken onderscheid met (bedoelde) corruptie.
ONTWIKKELAAR	Een organisatie of een persoon die een **AI-SYSTEEM** ontwerpt, ontwikkelt en/of traint.
OPDRACHTGEVER	Een persoon of organisatieonderdeel die een opdracht verstrekt aaneen opdrachtnemer. Deze is ook (samen met de projectleider) eindverantwoordelijk voor het maken van een AIIA.
OUTPUT(DATA)	De gegevens die een **AI-SYSTEEM** oplevert. Dit zijn de resultaten van het model.
PARAMETER	Een variabele binnen het **MODEL**. Wanneer deze variabele gewijzigd wordt, wordt ook de resulterende grootheid van het model of van de berekening gewijzigd.
POSITIEVE IMPACT	**BETROKKENEN** die gunstige gevolgen ervaren van de inzet van het **AI-SYSTEEM**. Denk aan een minderheidsgroep die wordt bevoordeeld. Daarbij kan het gevaar zijn dat deze positieve bias te optimistisch is, en dus niet waarheidsgetrouw. Ook kan de keerzijde hiervan een 'negatieve impact' voor andere betrokkenen zijn.
PROJECTLEIDER	De eindverantwoordelijke voor het project waarbinnen het **AI-SYSTEEM** valt. Deze is ook (samen met de **OPDRACHTGEVER**) **EINDVERANTWOORDELIJK** voor het maken van een AIIA.
PROPORTIONEEL	Proportionaliteit gaat over een redelijke verhouding tussen het doel en de ingezette oplossing. Staat het gebruik van AI in verhouding tot het probleem wat er met het **ALGORITME** opgelost gaat worden? Het verwachte voordeel moet groter zijn dan het risico dat AI met zich meebrengt.
REPRODUCEERBAAR	Het steeds opnieuw kunnen bereiken van een vergelijkbaar resultaat wanneer een beschreven procedure wordt uitgevoerd.
ROBUUSTHEID	Met een preventieve benadering ontwikkeld zijn; zich gedragen zoals voorzien en van tevoren beschreven. Onaanvaardbare schade vermijden.

SEED	Een 'seed' is het uitgangspunt van een willekeurig getal generator. Deze generator maakt altijd vanuit dit uitgangspunt volgens dezelfde 'route' nieuwe (pseudo) willekeurige getallen. Door de 'seed' te documenteren kan de 'route' van (pseudo) willekeurige getallen worden herhaald. Dit betekent dat deze seed nodig is om reconstructie van een **MODEL** te controleren wanneer het model ergens gebruik maakt van willekeurige getallen.
DE SEED ZELF IS OOK EEN GETAL	Er zijn geen specifieke eisen aan dit getal, dus vaak wordt er voor iets 'herkenbaars' gekozen (bijvoorbeeld '123456', of '0, 42, 1234' of de geboortedatum van een **ONTWIKKELAAR**).
SELECTIELIJST	Een lijst die beschrijft hoe lang archiefstukken bewaard moeten worden, bijvoorbeeld op basis van de archiefwet.
STAKEHOLDER	Persoon of organisatie die een beslissing of activiteit kan beïnvloeden, erdoor kan worden beïnvloed, of zichzelf als beïnvloed beschouwd. Een stakeholder kan bijvoorbeeld ook de eigenaar van gebruikte data zijn.
SUBSIDIAIR	Subsidiariteit gaat over het inzetten van het minst ingrijpende middel dat het probleem oplost. Kan het probleem ook met minder vergaande middelen opgelost worden?
TOEPASSINGSGEBIED	Term uit de AI-wetgeving die gebruikt wordt om de context waarin een AI-systeem gebruikt wordt aan te duiden. Denk bijvoorbeeld aan infrastructuur.
TRACEERBAARHEID	Wanneer processen en resultaten te controleren zijn.
TRANSPARANT	Wanneer de werking en doelen van het **AI-systeem** duidelijk worden gecommuniceerd en resultaten van het AI-systeem **UITLEGBAAR** zijn.
TYPE ALGORITMES	Verschillende technieken kunnen gebruikt worden om AI te maken, zoals neurale netwerken, random forests of andere vormen van machine learning. Maar ook minder complexe **ALGORITMES** zoals business-rules of beslisbomen kunnen gebruikt worden.
UITLEGBAAR	Een verklaring van hoe input variabelen bijdragen aan een output van het algoritme die uitgelegd moet worden.
VERANTWOORDELIJK	Handelingen van een **ENTITEIT** kunnen op unieke wijze worden herleid tot die entiteit, en deze entiteit is voor deze handelingen aansprakelijk. Wie is wie Vul in welke personen een rol hebben gespeeld bij het beantwoorden van deze AIIA.

Bijlage 1: Toets risicoclassificatie

De AI-verordening beschrijft een aantal risicotoepassingsgebieden voor AI-systemen. Hoe hoger her risico, hoe meer maatregelen er moeten worden getroffen. De risicoclassificatie wordt bepaald door het toepassingsgebied waar het AI-systeem wordt ingezet. De vragen hieronder helpen de risicoclassificatie van jouw AI-systeem te bepalen.

De risicoclassificatie in de AI-verordening is absoluut. Dat betekent dat andere risico's, zoals voortkomende uit privacywetgeving, niet meetellen. Een AI-systeem kan bijvoorbeeld grote impact hebben rond privacy, maar volgens de AI-verordening een toepassingsgebied hebben met minimaal risico. Het is dan géén hoog-risico-AI-systeem. De reguliere vragen in het AIIA kunnen je wel helpen om zulke risico's te mitigeren.

Definitie hoog-risico-AI-systeem (AI-verordening)

Een AI-systeem dat: individuen profileert d.m.v. het verwerken van persoonsgegevens. Daarnaast alle AI-systemen op het gebied van: (door de AI verordening) niet-verboden biometrische gegevens; Kritieke infrastructuur; onderwijs en beroepsopleiding; werkgelegenheid; toegang tot en gebruik van essentiële openbare en particuliere diensten; rechtshandhaving; migratie, asiel en grenscontroles; Rechtsbedeling en democratische processen.

Nog niet zeker binnen welke risicoclassificatie het AI-systeem valt? Beantwoord dan de volgende vragen.

A) Toetsing onaanvaardbaar risico

1. Maakt het AI-systeem gebruik van onbewuste boodschappen? Of worden er doelbewust manipulatieve of misleidende technieken gebruikt?
2. Maakt het AI-systeem gebruik van manipulatie/misbruik van een kwetsbare groep, wat tot fysieke of psychologische schade kan leiden?
3. Maakt het AI-systeem gebruik van ***social scoring***?[12]
4. Heeft het AI-systeem als doel emoties van een persoon op de werkplek af te leiden, buiten medische of veiligheidsoverwegingen?
5. Maakt het AI-systeem gebruik van biometrische identificatie in een openbare ruimte? Of biometrische categorisering van personen op basis van bijzondere persoonsgegevens?

Alles met nee beantwoord? Dan is het AI-systeem waarschijnlijk géén onaanvaardbaar risico toepassing.

B) Toetsing Hoog risico

1. Profileert het AI-systeem individuen?
2. Is het AI-systeem een product of veiligheidscomponent van zo'n product binnen een van de volgende velden:

- **Machines** (Richtlijn 2006/42/EG)
- **Speelgoed** (Richtlijn 2009/48/EG)
- **Pleziervaart** (Richtlijn 2013/53/EU)
- **Liften** (Richtlijn 2014/33/EU)
- **Apparaten en beveiligingssystemen voor gebruik op plaatsen met ontploffingsgevaar** (Richtlijn 2014/34/EU)
- **Radioapparatuur** (Richtlijn 2014/53/EU)
- **Drukapparatuur** (Richtlijn 2014/68/EU)
- **Kabelbaaninstallaties** (Verordening (EU) 2016/424)
- **Persoonlijke beschermingsmiddelen** (Verordening (EU) 2016/425)

[12] Social scoring: AI-systemen voor de evaluatie of classificatie van natuurlijke personen of groepen personen gedurende een bepaalde periode op basis van hun sociale gedrag of bekende, afgeleide of voorspelde persoonlijke of persoonlijkheidskenmerken (p. 51 AI Act)

- **Gasverbrandingstoestellen** (Verordening (EU) 2016/425)
- **Medische hulpmiddelen** (Verordening (EU) 2017/745)
- **Medische hulpmiddelen voor in-vitrodiagnostiek** (Verordening (EU) 2017/746)

Daarnaast noemt de verordening nog een lijst producten die ook als hoog-risicotoepassing AI wordt gezien. Op artikel 6.1, 102 tot 109 en 122 na, gelden de hoog-risicoverplichtingen vanuit de AI-verordening voor deze producten nog niet. Wél worden de eisen uit de AI-verordening op een later moment gebruikt om invulling te geven aan de specifieke productwetgeving die voor deze producten geldt. Wanneer dit gebeurt, is nog niet bekend en zal per product verschillen. Het gaat om de volgende producten en wetgeving:

- **(Beveiliging van) burgerluchtvaart** (Verordening (EG) 300/2008 en Verordening (EU) 2018/1139)
- **Twee- of driewielige voertuigen en vierwielers** (Verordening (EU) 168/2013)
- **Landbouw- en bosbouwvoertuigen** (Verordening (EU) 167/2013)
- **Uitrusting van zeeschepen** (Richtlijn 2014/90/EU)
- **Interoperabiliteit van het spoorwegsystem in de EU** (Richtlijn (EU) 2016/797)
- **Motorvoertuigen en aanhangwagens** (Verordening (EU) 2018/858 en Verordening (EU) 2019/2144)

3. Maakt het AI-systeem gebruik van biometrische identificatie op afstand van mensen?
4. Wordt het AI-systeem gebruikt als een veiligheidscomponenten bij beheer/exploitatie van: kritieke digitale infrastructuur[13], wegverkeer, de levering van water, gas, verwarming of elektriciteit?
5. Heeft het AI-systeem invloed op het werven en toegang tot arbeid van personen?
6. Wordt het AI-systeem ingezet rondom de toegang en gebruik van essentiële particuliere en publieke diensten en uitkeringen?
7. Is het AI-systeem actief op het gebied van rechtshandhaving?
8. Wordt het AI-systeem ingezet rondom migratie-, asiel- en grenstoezichtsbeheer?

Alles met nee beantwoord? Dan is het AI-systeem waarschijnlijk géén Hoog-risicotoepassing.

C) Toetsing transparantieverplichting

1. Is er sprake van een AI-systeem dat interactie heeft met personen?
2. Kan het AI-systeem kunstmatig inhoud genereren of manipuleren? (Generatieve AI GPAI)
3. Kan het AI-systeem emotie herkennen of biometrisch categoriseren?
4. Maakt het AI-systeem gebruik van deepfake technologieën?[14]

Als je een van de bovenstaande vragen met 'ja' antwoord, neem dan de nodige transparantiemaatregelen.[15] Dat wil over het algemeen zeggen: zorg dat gebruikers weten dat ze met een AI-systeem te maken hebben.

[13] Kritieke digitale infrastructuur: internetknooppunten, DSN-dienstverleners, registers voor topleveldomeinnamen, cloud computing, datacenters.

[14] Een door AI gemaakt of gemanipuleerd beeld-, audio-, of videomateriaal of tekst dat door personen ten onrechte voor authenthiek en waarheid kan worden gezien.

[15] AI Act, artikel 50.

Uitzonderingen

Als het systeem uitsluitend van toepassing is op een van de volgende terreinen zijn de reguliere vragen in de AIIA afdoende.

Let op, dit geldt alleen wanneer het AI systeem individuen niet profileert. Anders is het systeem **altijd** een hoog-risicotoepassing.[16]

Uitzonderingen, AI voor:
1. Militaire doeleinden
2. Rechtshandhaving en justitie
3. Onderzoek naar AI
4. Open source AI in de ontwikkelingsfase (vóór markttoetreding)
5. Non professionele activiteiten
6. Het AI-systeem beïnvloedt de uitkomst van de besluitvorming niet wezenlijk. Hierbij moet gedacht worden aan een systeem dat louter procedurele taken vervult of dient als controle en verbetering van een eerdere menselijke activiteit (bijvoorbeeld door taalverbetering). Ook systemen die functioneren ter opsporing van besluitvormingspatronen vallen hieronder.[17]

Nota bene: nog veel definities rondom AI-toepassingen worden uitgewerkt. Daarom is het mogelijk dat na het invullen van deze vragen nog niet helemaal duidelijk is in welke categorie het AI-systeem valt. Je kan overwegen om deze vraag en andere voor te leggen aan de Regulatory Sandbox AI van de Rijksoverheid (meer info).

[16] Ai Act, artikel 6, lid 3.
[17] AI Act, artikel 6, lid 3.

Bijlage 2: Hoog-risicosystemen

Is het AI-systeem een hoog-risicotoepassing (zie bijlage 1)? Gebruik dan de volgende additionele vragen als hulpmiddel. Op deze manier neem je de juiste waarborgen om het AI-systeem op een verantwoorde manier te gebruiken.

De AI-verordening beschrijft verschillende partijen die ieder haar eigen rol en verantwoordelijkheid hebben rondom een AI-systeem. In deze bijlage staan de vragen voor de gebruiker (gebruiksverantwoordelijke) en ontwikkelaar (aanbieder) van een AI-systeem. Een aanbieder schaft bijvoorbeeld een AI-systeem aan en verbindt daar een eigen naam aan. Of brengt substantiële wijzigingen aan het systeem aan.[18]

Vragen indien je gebruik wil maken van een hoog-risico-AI-systeem

Volgens de AI-verordening moeten hoog-risico-systemen aan een aantal eisen voldoen. De onderstaande vragen helpen daarbij.

Effecten
- Vormt het AI-systeem een significant risico voor de gezondheid, veiligheid of fundamentele rechten van personen?[19] Beredeneer hieronder waarom wel/niet.

Onderhoud & beheer
- Hoe is monitoring en werking van het AI-systeem geborgd?[20]
- Hoe worden de logs die het AI-systeem produceert ten minste 6 maanden bewaard?[21]
- Zijn er maatregelen getroffen om menselijk toezicht te regelen door personen met de benodigde bekwaamheid, opleiding en autoriteit?

Technische robuustheid
- Is de data(input) relevant en representatief, rekening houdende met het beoogde doel (vraag i1) van het AI-systeem?[22]
- Is het AI-systeem voorzien van de technische documentatie?[23]
- Is het voor het menselijk toezicht mogelijk de **OUTPUT(DATA)** op de juiste wijze te controleren, interpreteren, of eventueel te negeren?[24]

Betrouwbaarheid
- Zijn de niveaus van maatstaven voor nauwkeurigheid vermeld in de gebruikersaanwijzingen?

Data governance
- Voldoet de trainingsdata van het AI-systeem aan de volgende kwaliteitseisen?
- Relevante ontwerpkeuzes voor datasets.
- Inzichtelijk maken wat de herkomst van data is.
- Relevante verwerkingsactiviteten zoals annotatie, labelen, opschoning, actualisatie, verrijking en aggregatie.
- Het opstellen van aannames die de data moet meten en vertegenwoordigen.
- Een beoordeling van de beschikbaarheid, kwantiteit en geschiktheid van de benodigde datasets.

[18] Ai Act, artikel 25.
[19] AI Act, artikel 6, punt 2a.
[20] AI Act, artikel 26, lid 5.
[21] AI Act, artikel 26, lid 6.
[22] Ai Act, artikel 26, lid 4.
[23] AI Act, bijlage IV.
[24] Ai Act, artikel 14, lid 4.

- Een beoordeling van mogelijke vooringenomenheid van de data met negatieve gevolgen voor gezondheid, veiligheid, grondrechten, en discriminatie.
- Maatregelen om vooringenomenheid op te sporen, te voorkomen en te beperken.
- Identificeren en aanpakken van tekortkomingen die naleving van de regelgeving belemmeren.

Risicobeheer
- Hoe is het AI-systeem getest op het beoogde doel en is dat in overeenstemming met de risicobeheerseisen voordat het in gebruik wordt genomen?
- Is er een risicobeheersysteem vastgesteld en gedocumenteerd? Deze bestaat uit volgende stappen:
  - Een risicoanalyse van het AI-systeem voor de gezondheid, veiligheid of grondrechten
  - Een inschatting en evaluatie van de risico's die zich kunnen voordoen
  - Een evaluatie van nieuwe risico's na markttoetreding, op basis van het monitoringsysteem van het AI-systeem[25]
  - Het opstellen van risicobeheersmaatregelen
- Hoe waarborg je eventuele samenwerking met de toezichthouders en andere bevoegde autoriteiten?[26] Denk aan contactpersonen, bereikbaarheid etc.
- Is het waarschijnlijk dat kwetsbare groepen (zoals kinderen) toegang zullen hebben tot het AI-systeem? In dat geval moeten de risicobeheersmaatregelen extra worden aangescherpt

Nota bene: Het AI-bureau ontwikkelt een sjabloon voor een vragenlijst, onder meer via een geautomatiseerd instrument, om gebruiksverantwoordelijken te helpen deze verplichtingen op vereenvoudigde wijze na te komen.

Communicatie
- Hoe communiceer je over de in gebruik neming van het hoog-risico-AI-systeem?[27]
- Is het AI-systeem geregistreerd in de EU-databank voor hoog-risicosystemen of in het Nederlandse algoritmeregister? Hoog Risico AI-systemen met een kritieke- infrastructuur-toepassing dienen altijd in het algoritmeregister gepubliceerd te worden.[28]
- Zijn er gebruiksinstructies opgesteld? Deze moeten minstens het volgende bevatten[29]:
  - De identiteit en contactgegevens van de aanbieder;
  - Kenmerken, capaciteiten en beperkingen (doel);
  - Mogelijke toekomstige wijzigingen;
  - Maatregelen omtrent menselijk toezicht;
  - De benodigde rekenkracht en hardware, verwachte levensduur, de noodzakelijke maatregelen (en de frequentie daarvan) voor onderhoud en verzorging
  - Een beschrijving van de mechanismen in het AI-systeem
  - De niveaus van nauwkeurigheid en de relevante maatstaven voor de nauwkeurigheid

Generatieve AI
- Is er sprake van informatie en documentatie waarmee de mogelijkheden en limitatie van het AI-systeem duidelijk worden?
- Is er beleid hoe auteursrechten worden gewaarborgd door het AI-systeem?
- Is er een gedetailleerde samenvatting over de content waarmee het AI-systeem is getraind?

Overig
- Indien er gebruik wordt gemaakt van biometrische identificatie op afstand, is er toestemming van een gerechtelijke instantie?[30]

[25] The post-market monitoring system shall actively and systematically collect, document and analyse relevant data which may be provided by deployers or which may be collected through other sources on the performance of high-risk AI systems throughout their lifetime, and which allow the provider to evaluate the continuous compliance of AI systems with the requirements set out in Chapter III, Section 2.
[26] AI Act, artikel 26, lid 12.
[27] AI Act, artikel 26, lid 7.
[28] AI Act, artikel 26, lid 8, en artikel 49 & 71.
[29] AI Act, artikel 13, lid 2 & 3.
[30] Ibidem, lid 10.

Vragen voor ontwikkelaars (aanbieders) van hoog-risico-AI-systemen

Ontwikkelaars (of zoals in de AI-verordening: **aanbieder**[31]) van een hoog-risico-AI-systeem moeten ook onderstaande vragen te beantwoorden.

- Is er documentatie opgesteld waarin de naleving van vereisten gedocumenteerd is? Zie bijlage IV van de AI-verordening voor de vereisten.
- Op welke wijze is technische documentatie vindbaar voor gebruikers? Ook als er geen betrokkenheid is bij de uitrol van je systeem?
- Is vermeld, en zo ja op welke wijze, dat het een hoog-risicosysteem is?[32]
- Is er een kwaliteitsbeheersysteem opgezet?[33] Ook dit dien je te documenteren. Daarvoor is ten minste nodig:
  - Een nalevingsstrategie van de regelgeving
  - Ontwerpprocedures
  - Kwaliteitscontroleprocedures
  - Een inspectieprocedure
  - Een overzicht van de technische eisen en wat ervoor nodig is om die te handhaven
  - Databeheer procedures
  - Een risicomanagement systeem
  - Een monitoringsprocedure
  - Een procedure voor het melden van ernstige incidenten
  - Een communicatiestrategie met de bevoegde autoriteiten
  - Systemen en procedures voor de registratie van relevante documentatie
  - Het beheer van hulpmiddelen en voorzieningszekerheid
  - Een verantwoordingskader
- Hoe worden de logs van het AI-systeem bewaard?[34]
- Is er een conformiteitsbeoordelingsprocedure en is deze vastgesteld in een EU-conformiteitsverklaring?[35] Hierin wordt aangetoond of er voldaan is aan de vastgestelde eisen voor het AI-systeem. Deze moet worden voorgelegd aan een aangewezen instantie en ten minste 10 jaar te bewaren.
- Is de relevante toezichthouder akkoord met de conformiteitsverklaring?[36] Dit hoeft alleen als het hoog-risico-AI-systeem betrekking heeft op een relevant hoog-risicoproduct (bijlage I).
- Is er, indien het AI-systeem aan de AI-verordening meent te voldoen, een CE-markering aangebracht, bijvoorbeeld in de documentatie?[37]
- Hoe worden, indien het AI-systeem niet meer in overeenstemming is met de AI-verordening, corrigerende maatregelen genomen, zoals het systeem uit handel nemen, deactiveren of terugroepen? Op welke wijze worden gebruikers hiervan op de hoogte gesteld?[38]
- Voldoet het systeem aan de Europese toegankelijkheidseisen?[39]

[31] **"Aanbieder"**: een natuurlijke of rechtspersoon, overheidsinstantie, agentschap of ander orgaan die/dat een AI-systeem of een AI-model voor algemene doeleinden ontwikkelt of laat ontwikkelen en dat systeem of model in de handel brengt of het AI-systeem in gebruik stelt onder de eigen naam of merk, al dan niet tegen betaling.

[32] Ai Act, artikel 16.

[33] Ai Act, artikel 17.

[34] AI Act, artikel 19.

[35] AI Act, artikel 43 & 47.

[36] AI Act, artikel16, lid k.

[37] Ai Act, artikel 48.

[38] Ai Act, artikel 20.

[39] Richtlijnen (EU) 2016/2102 en (EU) 2019/882.

Bijlage 3: Aandachtspunten generatieve AI

Voor het gebruik van generatieve AI, zoals large language models (LLM's), zijn er een aantal aandachtspunten die belangrijk zijn bij het invullen van het AIIA. Deze bijlage focust op de belangrijkste afwijkingen van generatieve AI ten opzichte van andere AI-systemen.

Wat is generatieve AI?
Generatieve AI is een vorm van AI waarbij algoritmes worden ingezet om content te genereren. Het kabinet schrijft in haar Overheidsbrede visie op generatieve AI[40], dat generatieve AI in dienst moet staan van het vergroten van het menselijk welzijn en autonomie, duurzaamheid, welvaart, rechtvaardigheid en veiligheid. Door in te zetten op verantwoorde toepassingen van generatieve AI, grijpen we de kansen die deze technologie biedt, aldus de visie.

Onder de motorkap maakt generatieve AI gebruik van een neuraal netwerk dat bestaat uit biljoenen parameters. Welke output gekozen wordt is op basis van statistiek, er zit geen logica of kennis over de werkelijkheid achter. De output wisselt en is moeilijk te reproduceren of te verantwoorden. Dit is gelijk het belangrijkste aandachtspunt voor het invullen van de AIIA: gebruik generatieve AI alleen als het acceptabel is dat de uitkomst niet uitlegbaar of controleerbaar is.

Voorlopig standpunt voor Rijksorganisaties: in beginsel niet toegestaan
Het voorlopig standpunt over gebruik van generatieve AI bij Rijksorganisaties, stelt vooralsnog hoge eisen voor het gebruik van LLM's bij het Rijk: "Niet-gecontracteerde generatieve AI-toepassingen zoals ChatGPT, Bard en Midjourney, voldoen over het algemeen niet aantoonbaar aan de geldende privacy- en auteursrechtelijke wetgeving. Zodoende is het gebruik hiervan door Rijksorganisaties (of in opdracht daarvan) in beginsel niet toegestaan, in die gevallen waarin het risico bestaat dat wetgeving wordt overtreden, tenzij de aanbieder en de gebruiker aantoonbaar voldoen aan de geldende wet- en regelgeving."[41]

Aandachtspunten bij het invullen van het AIIA
Hieronder staan de belangrijkste afwijkingen van generatieve AI ten opzichte van andere AI-systemen. Dit helpt bij het invullen van het AIIA. Als een punt uit het AIIA niet genoemd is dan betekent het niet dat deze niet relevant is, maar dat er geen specifieke aandachtspunten zijn omdat het generatieve AI is.

Doel en Noodzaak
- **Publieke waarden:** er zijn geen specifieke aandachtspunten voor generatieve AI. Vaak komen transparantie, uitlegbaarheid en duurzaamheid in gedrang bij de keuze voor generatieve AI.
- **Grondrechten**: Ga na of het model op ethische wijze is getraind en 'gefinetuned'. In deze laatste stap wordt menselijke feedback gebruikt om de antwoorden te verbeteren. Deze werkomstandigheden van deze mensen ('click workers') zijn, net als voor veel andere producten, niet altijd even goed.[42]
- **Duurzaamheid**: Ga na of met een minder complex systeem of tool, ook de doelstellingen behaald kunnen worden. Generatieve AI gebruikt enorm veel energie voor zowel het trainen als het gebruiken van het model. Onderzoek of er energiebesparende technieken toegepast kunnen worden, zoals een kleiner model, optimalisatie of een andere software-architectuur.
- **Afweging**: Neem het voorlopig standpunt voor het gebruik van generatieve AI bij Rijksorganisaties (zie inleiding) mee in de beslissing om generatieve AI wel of niet in te zetten.

[40] https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2024Z00480&did=2024D01191
[41] https://www.rijksoverheid.nl/actueel/nieuws/2023/12/11/voorlopig-standpunt-generatieve-ai-kabinet
[42] AI draait op werk van miljoenen onzichtbare, slechtbetaalde mensen. Wie komt er voor ze op? - De Correspondent

Technische robuustheid

- **Bias**: Onderzoek de uitvoer van het systeem altijd op bias en breng in kaart wat de risico's van de eventuele bias zijn. De data waar een generatief AI-model op getraind is bevat vaak al een bias, en door de aard (statistisch taalmodel) van het model wordt deze bias versterkt. Softwareleveranciers halen via menselijke feedback en 'guardrails' de scherpe randen eraf, maar de bias zit nog steeds in het model. Er zijn daarnaast ook andere methodes om bias te verminderen en/of te detecteren.[43]
- **Accuraatheid, betrouwbaarheid en reproduceerbaarheid:** Gebruik een generatief AI-model alleen voor doelen waarbij geen hoge **accuraatheid**, betrouwbaarheid en reproduceerbaarheid vereist is en/of neem maatregelen in de opzet van het systeem om de kans op hallucinaties te verkleinen. Breng in kaart wat de risico's van een onbetrouwbare uitkomst zijn. Een LLM is een statistisch taalmodel zonder kennis van de werkelijkheid. De output is altijd anders en niet persé gebaseerd op waarheid. Ook kan een generatief AI-model gaan 'hallucineren', waarbij antwoorden worden gegenereerd die logisch en overtuigend klinken, maar niet waar zijn. Een andere uitdaging bij generatieve AI is dat het lastig is om de kwaliteit van de output te meten.
- **Technische implementatie**: Maak voor een extern gehost generatief AI-model, contractuele afspraken over het gebruik van de data(input) van de organisatie. Vaak wil men dit gebruiken als data om het model verder te trainen. Het is mogelijk dat anderen (onbevoegden) deze data (bij gebruik) kunnen inzien bij het gebruik van de juiste prompts. Bij de inkoop van een systeem moet de leverancier aantonen dat ze aan alle gestelde eisen voldoen. Het kabinetsstandpunt voor generatieve AI spreekt een voorkeur uit voor *open source* generatieve AI.
- **Uitlegbaarheid**: Gebruik generatieve AI-modellen alleen wanneer het de uitlegbaarheid van resultaten geen vereiste is. Breng in kaart wat de risico's van het gebrek aan uitlegbaarheid zijn. Leg uit aan medewerkers die met het systeem moeten werken wat de beperkingen zijn. Een generatieve AI is een zeer complex systeem met biljoenen parameters waarbij de werking niet uitlegbaar is: een 'black box'. Het is vrijwel onmogelijk om aan te tonen hoe het model tot een antwoord komt, alhoewel er wel initiatieven zijn om dit inzichtelijk te maken.[44]

Data governance

- **Kwaliteit en integriteit van data**: Het is waarschijnlijk dat de trainingsdata van generatieve AI persoonsgegevens en auteursrechtelijk beschermd materiaal bevat. Voor de meeste AI-modellen is niet bekend welke trainingsdata (en wat de kwaliteit daarvan is) gebruikt is in het model. Vanuit de AI-verordening is transparantie vereist. Vanuit de AVG mogen persoonsgegevens niet zomaar verwerkt worden.
- **Privacy**: maak afspraken met de leverancier over de verwerkingsverantwoordelijkheid van de ingevoerde data (bijvoorbeeld of het gebruikt wordt als trainingsdata). Stel een goede verwerkersovereenkomst op. Door goede afspraken wordt de kans op datalekken verkleind.
- **Informatiebeveiliging**: De OWASP heeft een overzicht van de 10 meest kritieke kwetsbaarheden, en daarmee informatiebeveiligingsrisico's, van generatieve AI-systemen.[45]

Verantwoordingsplicht

- **Controleerbaarheid:** zie 'uitlegbaarheid'.

[43] https://www.datacamp.com/blog/understanding-and-mitigating-bias-in-large-language-models-llms
[44] https://arxiv.org/abs/2309.01029
[45] OWASP Top 10: LLM & Generative AI Security Risks

Dit is een publicatie van:

Ministerie van Infrastructuur en Waterstaat

Postbus 20901

2500 EX Den Haag

December 2024 | 73263