

ONAFHANKELIJKE

DESKUNDIGENGROEP OP HOOG NIVEAU

INZAKE KUNSTMATIGE INTELLIGENTIE

OPGERICHT DOOR DE EUROPESE COMMISSIE IN JUNI 2018



**ETHISCHE RICHTSNOEREN
VOOR BETROUWBARE KI**

ETHISCHE RICHTSNOEREN voor BETROUWBARE KI

Deskundigengroep op hoog niveau inzake kunstmatige intelligentie

Dit document is opgesteld door de deskundigengroep op hoog niveau inzake kunstmatige intelligentie (AI HLEG). De in dit document genoemde leden van de AI HLEG ondersteunen het algehele kader voor betrouwbare KI dat in deze richtsnoeren uiteen wordt gezet, maar zijn het niet noodzakelijkerwijs eens met iedere afzonderlijke bewering in het document. .

De controlelijst voor betrouwbare KI die in hoofdstuk III van dit document wordt gepresenteerd, wordt gedurende een testfase door belanghebbenden getest om praktische feedback te verzamelen. Begin 2020 wordt er een herziene versie van de controlelijst aan de Europese Commissie voorgelegd, waarin de tijdens de testfase verzamelde feedback wordt verwerkt.

De AI HLEG is een onafhankelijke deskundigengroep die in juni 2018 door de Europese Commissie is opgericht.

Contactpersoon Nathalie Smuha – Coördinator AI HLEG
E-mailadres CNECT-HLG-AI@ec.europa.eu

Europese Commissie
B-1049 Brussel

Document gepubliceerd op 8 april 2019.

Op 18 december 2018 is er een eerste ontwerp van dit document uitgebracht, waarop middels een openbare raadpleging feedback is gegeven door meer dan vijfhonderd belanghebbenden. Wij willen iedereen die feedback op het eerste ontwerp van het document heeft gegeven, uitdrukkelijk en hartelijk bedanken. Deze feedback is meegenomen bij het opstellen van deze herziene versie.

De Europese Commissie of personen die namens de Commissie optreden, zijn niet aansprakelijk voor het eventuele gebruik dat van de volgende informatie wordt gemaakt. De inhoud van dit werkdokument valt uitsluitend onder de verantwoordelijkheid van de deskundigengroep op hoog niveau inzake kunstmatige intelligentie (AI HLEG). Hoewel personeel van de Commissie heeft meegewerkt aan de totstandkoming van de richtsnoeren, geven de hierin geformuleerde meningen uitsluitend het standpunt van de AI HLEG weer en mogen zij in geen geval worden beschouwd als een officieel standpunt van de Europese Commissie.

Meer informatie over de deskundigengroep op hoog niveau inzake kunstmatige intelligentie is online beschikbaar (<https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>).

Het beleid ten aanzien van hergebruik van documenten van de Europese Commissie is vastgelegd in Besluit 2011/833/EU (PB L 330 van 14.12.2011, blz. 39). Voor gebruik of overname van foto's of andere materialen die niet onder het auteursrecht van de EU vallen, moet u rechtstreeks toestemming vragen aan de houders van het desbetreffende auteursrecht.

INHOUDSOPGAVE

SAMENVATTING	2
A. INLEIDING	5
B. EEN KADER VOOR BETROUWBARE KI	7
I. Hoofdstuk I: Grondslagen van betrouwbare KI	11
1. Grondrechten als morele en wettelijke rechten	11
2. Van grondrechten naar ethische beginselen	12
II. Hoofdstuk II: Betrouwbare KI verwezenlijken	17
1. Vereisten van betrouwbare KI	17
2. Technische en niet-technische methoden voor de verwezenlijking van betrouwbare KI	25
III. Hoofdstuk III: Betrouwbare KI controleren	30
C. VOORBEELDEN VAN DE MOGELIJKHEDEN EN PUNTEN VAN ZORG DIE KI MET ZICH MEEBRENGT	41
D. CONCLUSIE	46
WOORDENLIJST	47

SAMENVATTING

- (1) Het doel van deze richtsnoeren is het bevorderen van betrouwbare KI. Betrouwbare KI bestaat uit **drie componenten**, waaraan gedurende de volledige levenscyclus van het systeem moet worden voldaan: de KI moet a) **wettig** zijn, door te voldoen aan alle toepasselijke wet- en regelgeving, b) **ethisch** zijn, door naleving van ethische beginselen en waarden te waarborgen, en c) **robuust** zijn uit zowel technisch als sociaal oogpunt, aangezien KI-systemen ongewild schade kunnen aanrichten, zelfs al zijn de bedoelingen goed. Iedere component is op zichzelf nodig, maar niet voldoende om betrouwbare KI te bewerkstelligen. In het ideale scenario sluiten alle drie de componenten op elkaar aan en valt de werking ervan gedeeltelijk samen. Als er in de praktijk spanningen ontstaan tussen deze componenten, moet de maatschappij zich inspannen om ze weer op één lijn te brengen.
- (2) In deze richtsnoeren wordt een **kader voor het bewerkstelligen van betrouwbare KI** geschetst. In het kader wordt niet expliciet ingegaan op de eerste component van betrouwbare KI (wettige KI)¹. In plaats daarvan is het bedoeld als richtsnoer voor het bevorderen en waarborgen van ethische en robuuste KI (de tweede en derde component). Deze richtsnoeren, die gericht zijn op alle belanghebbenden, vormen een poging verder te gaan dan een lijst van ethische beginselen, doordat wordt toegelicht hoe dergelijke beginselen in de praktijk in sociaal-technische systemen kunnen worden verwerkt. Deze toelichting wordt in drie lagen met verschillend abstractieniveau gegeven: van het meest abstract in hoofdstuk I tot het meest concreet in hoofdstuk III, dat eindigt met voorbeelden van de mogelijkheden en punten van zorg die KI-systemen met zich meebrengen.
- I. Op basis van een op grondrechten gefundeerde benadering worden in hoofdstuk I de **ethische beginselen** en de bijbehorende waarden vastgesteld die bij de ontwikkeling, de installatie en het gebruik van KI-systemen moeten worden gerespecteerd.

Essentiële richtsnoeren uit hoofdstuk I:

- ✓ Ontwikkel, installeer en gebruik KI-systemen op dusdanige wijze dat de volgende ethische beginselen worden nageleefd: *respect voor menselijke autonomie, preventie van schade, rechtvaardigheid en verantwoording*. Erken de potentiële spanningen tussen deze beginselen en pak deze aan.
- ✓ Schenk bijzondere aandacht aan situaties waarbij kwetsbaardere groepen betrokken zijn, zoals kinderen, mensen met een beperking en anderen die historisch gezien kansarm zijn of het risico lopen te worden uitgesloten, en aan situaties die worden gekenmerkt door ongelijkheid wat betreft macht of beschikking over informatie, bijvoorbeeld tussen werkgevers en werknemers of tussen bedrijven en consumenten².
- ✓ Erken en wees u ervan bewust dat KI-systemen de mens en de samenleving grote voordelen kunnen opleveren, maar ook bepaalde risico's en negatieve gevolgen met zich mee kunnen brengen, die mogelijk lastig te voorspellen, vast te stellen of te meten zijn (bijv. gevolgen voor de democratie, de rechtsstaat en de verdelende rechtvaardigheid of voor de menselijke geest). Neem waar nodig passende maatregelen, evenredig aan de omvang van het risico, om deze risico's te beperken.

- II. Voortbouwend op hoofdstuk I worden er in hoofdstuk II richtsnoeren geboden voor de bewerkstelling van betrouwbare KI, in de vorm van **zeven vereisten** voor KI-systemen. Voor de verwezenlijking hiervan kunnen zowel technische als niet-technische methoden worden gebruikt.

¹ Alle normatieve beweringen in dit document zijn bedoeld als richtsnoer voor het verwezenlijken van de tweede en derde component van betrouwbare KI (ethische en robuuste KI). Deze beweringen zijn dus niet bedoeld als juridisch advies of als hulp bij het naleven van de toepasselijke wetgeving, hoewel vele ervan tot op zekere hoogte al wel in de bestaande wetgeving zijn opgenomen. Zie met betrekking daartoe punt 21 en verder.

² Zie de artikelen 24 tot en met 27 van het Handvest van de grondrechten van de EU (het Handvest), waarin wordt ingegaan op de rechten van kinderen en ouderen, de integratie van mensen met een beperking en de rechten van werknemers. Zie ook artikel 38 over de bescherming van consumenten.

Essentiële richtsnoeren uit hoofdstuk II:

- ✓ Zorg ervoor dat de ontwikkeling, de installatie en het gebruik van KI-systemen voldoen aan de vereisten voor betrouwbare KI: 1) menselijke controle en menselijk toezicht, 2) technische robuustheid en veiligheid, 3) privacy en datagovernance, 4) transparantie, 5) diversiteit, non-discriminatie en rechtvaardigheid, 6) milieu- en maatschappelijk welzijn en 7) verantwoordingsplicht.
- ✓ Bestudeer technische en niet-technische methoden om te waarborgen dat die vereisten worden verwezenlijkt.
- ✓ Stimuleer onderzoek en innovatie om bij te dragen aan het controleren van KI-systemen en het verwezenlijken van de vereisten te bevorderen. Verspreid de resultaten en open vragen onder een breder publiek en leid systematisch een nieuwe generatie deskundigen op het gebied van KI-ethiek op.
- ✓ Verspreid informatie over de mogelijkheden en beperkingen van een KI-systeem en over de manier waarop de vereisten worden verwezenlijkt, op een heldere en proactieve manier onder belanghebbenden om onrealistische verwachtingen te voorkomen. Wees open over het feit dat ze met een KI-systeem te maken hebben.
- ✓ Faciliteer de traceerbaarheid en controleerbaarheid van KI-systemen, met name in kritieke omgevingen of situaties.
- ✓ Betrek belanghebbenden bij de volledige levenscyclus van het KI-systeem. Stimuleer onderwijs en opleiding, zodat alle belanghebbenden zich bewust zijn van betrouwbare KI en opgeleid zijn op dat gebied.
- ✓ Wees bedacht op conflicterende beginselen en vereisten. Zoek, evalueer en documenteer deze afwegingen en de oplossingen voortdurend en maak ze kenbaar.

III. In hoofdstuk III wordt een concrete en niet-uitputtende controlelijst voor betrouwbare KI gegeven, die gericht is op de operationalisering van de vereisten uit hoofdstuk II. De **controlelijst** moet worden afgestemd op de specifieke gebruikssituatie van het KI-systeem³.

Essentiële richtsnoeren uit hoofdstuk III:

- ✓ Stel een controlelijst voor betrouwbare KI vast bij de ontwikkeling, de installatie of het gebruik van KI-systemen en stem deze af op de specifieke situatie waarin het systeem wordt gebruikt.
- ✓ Bedenk dat een dergelijke controlelijst nooit uitputtend zal zijn. Bij betrouwbare KI gaat het niet om het afvinken van elementen op een lijst, maar om het voortdurend vaststellen en verwezenlijken van vereisten, afwegen van oplossingen en waarborgen van betere resultaten gedurende de hele levenscyclus van het KI-systeem, en om het betrekken van de belanghebbenden bij dit proces.

- (3) Het laatste gedeelte van het document is bedoeld om een aantal van de binnen het kader besproken kwesties concreet te maken door voorbeelden te geven van voordelige mogelijkheden die moeten worden nagestreefd, en punten van zorg die KI-systemen met zich meebrengen en die zorgvuldig moeten worden afgewogen.
- (4) Hoewel deze richtsnoeren bedoeld zijn als ondersteuning voor KI-toepassingen in het algemeen door een horizontale grondslag te creëren voor de bewerkstelling van betrouwbare KI, brengen verschillende situaties verschillende uitdagingen met zich mee. Daarom moet worden onderzocht of er, gezien de contextgebondenheid van KI-systemen, naast dit horizontale kader ook een sectorale benadering nodig is.
- (5) Het is niet de bedoeling dat deze richtsnoeren in de plaats komen van beleid of regelgeving, huidig of toekomstig en in welke vorm dan ook, noch dat zij de invoering daarvan ontmoedigen. Ze moeten worden

³ Overeenkomstig de bij punt 2 beschreven reikwijdte van het kader biedt deze controlelijst geen advies over het naleven van de wetgeving (wettige KI), maar uitsluitend richtsnoeren voor het verwezenlijken van de tweede en derde component van betrouwbare KI (ethische en robuuste KI).

gezien als een dynamisch document, dat periodiek moet worden herzien en bijgewerkt om relevant te blijven naarmate de technologie, onze sociale omgeving en onze kennis zich ontwikkelen. Dit document wordt beschouwd als een uitgangspunt voor het debat over "betrouwbare KI voor Europa"⁴. Buiten Europa zijn deze richtsnoeren bedoeld om het onderzoek naar, de bezinning op en het debat over een ethisch kader voor KI-systemen op mondiaal niveau te stimuleren.

⁴ Het is de bedoeling dat dit ideaal van toepassing is op KI-systemen die in de EU-lidstaten worden ontwikkeld, geïnstalleerd en gebruikt, alsook op systemen die elders worden ontwikkeld of geproduceerd, maar die in de EU worden geïnstalleerd en gebruikt. Wanneer er in dit document naar "Europa" wordt verwezen, worden daarmee de EU-lidstaten bedoeld. Het streven is echter dat deze richtsnoeren ook buiten de EU relevant zijn. Met betrekking daartoe kan ook worden opgemerkt dat zowel Noorwegen als Zwitserland deel uitmaakt van het Gecoördineerd plan inzake KI, dat in december 2018 door de Commissie en de lidstaten is overeengekomen en bekendgemaakt.

A. INLEIDING

- (6) In haar mededelingen van 25 april 2018 en 7 december 2018 heeft de Europese Commissie (de Commissie) haar visie op kunstmatige intelligentie (KI) uiteengezet, waarin "ethische, veilige en geavanceerde KI in Europa" wordt ondersteund⁵. De visie van de Commissie rust op drie pijlers: i) het verhogen van publieke en particuliere investeringen in KI om het gebruik ervan te stimuleren, ii) het anticiperen op sociaaleconomische veranderingen en iii) het waarborgen van een passend ethisch en juridisch kader ter versterking van Europese waarden.
- (7) Om de uitvoering van deze visie te ondersteunen heeft de Commissie de deskundigengroep op hoog niveau inzake kunstmatige intelligentie (AI HLEG) opgericht, een onafhankelijke groep belast met het opstellen van twee documenten: 1) ethische richtsnoeren voor KI en 2) aanbevelingen voor beleid en investeringen.
- (8) Dit document bevat de ethische richtsnoeren voor KI, die zijn herzien na verdere besprekingen binnen onze groep in het kader van de bij de openbare raadpleging ontvangen feedback op de ontwerpversie die op 18 december 2018 was gepubliceerd. In de richtsnoeren wordt voortgebouwd op de werkzaamheden van de Europese Groep ethiek van de exacte wetenschappen en de nieuwe technologieën⁶ en wordt uitgegaan van andere soortgelijke inspanningen⁷.
- (9) De afgelopen maanden hebben de 52 leden van de deskundigengroep vergaderd, gediscussieerd en contact gehad, met het oog op het Europese motto: in verscheidenheid verenigd. Wij zijn ervan overtuigd dat KI de potentie heeft om de samenleving sterk te veranderen. KI is geen doel, maar veeleer een veelbelovend middel om de mens tot bloei te laten komen en zo het individuele en maatschappelijke welzijn te verbeteren, het algemeen belang te dienen en vooruitgang en innovatie teweeg te brengen. In het bijzonder kunnen KI-systemen bijdragen aan het verwezenlijken van de duurzameontwikkelingsdoelstellingen van de VN, zoals het bevorderen van het genderevenwicht, het aanpakken van de klimaatverandering, het rationaliseren van ons gebruik van natuurlijke hulpbronnen, het verbeteren van onze gezondheid, mobiliteit en productieprocessen en het ondersteunen van de manier waarop we de voortgang bijhouden aan de hand van indicatoren van duurzaamheid en sociale cohesie.
- (10) Daarvoor moet bij KI-systemen⁸ **de mens centraal staan** en moet ernaar worden gestreefd deze systemen te gebruiken in dienst van de mensheid en het algemeen belang, met als doel de verbetering van het welzijn en de vrijheid van de mens. KI-systemen bieden geweldige mogelijkheden, maar brengen ook bepaalde risico's met zich mee, waar passend en in proportie mee moet worden omgegaan. Er ligt nu een belangrijke kans om richting te geven aan de ontwikkeling van deze systemen. Wij willen ervoor zorgen dat we de sociaaleconomische situaties waarin ze worden ingebed, kunnen vertrouwen, en wij willen dat de producenten van de KI-systemen een concurrentievoordeel verkrijgen door betrouwbare KI in hun producten en diensten te verwerken. Hiervoor is het nodig **de voordelen van KI-systemen te maximaliseren** en tegelijk **de risico's ervan te voorkomen en tot een minimum te beperken**.
- (11) Wij zijn ervan overtuigd dat het in een context van snelle technologische verandering essentieel is dat vertrouwen het cement blijft van samenlevingen, gemeenschappen, economieën en duurzame ontwikkeling. Daarom kiezen wij **betrouwbare KI als onze fundamentele ambitie**, want mensen en gemeenschappen zullen

⁵ COM(2018) 237 final en COM(2018) 795 final. Merk op dat de term "in Europa" ("made in Europe") in de hele mededeling van de Commissie wordt gebruikt. Deze richtsnoeren beogen echter om niet alleen de in Europa gemaakte KI-systemen te omvatten, maar ook de systemen die elders worden ontwikkeld en in Europa worden geïnstalleerd of gebruikt. In dit document streven we er steeds naar betrouwbare KI "voor" Europa te stimuleren.

⁶ De Europese Groep ethiek van de exacte wetenschappen en de nieuwe technologieën (EGE) is een adviesgroep van de Commissie.

⁷ Zie punt 3.3 van COM(2018) 237.

⁸ In de woordenlijst aan het einde van dit document wordt een definitie van KI-systemen gegeven zoals gebruikt binnen dit document. Deze definitie wordt verder uitgewerkt in een speciaal document dat door de AI HLEG is opgesteld en dat bij deze richtsnoeren is gevoegd, getiteld "Een definitie van KI: de belangrijkste capaciteiten en wetenschappelijke disciplines".

alleen vertrouwen kunnen hebben in de technologische ontwikkelingen en de toepassingen daarvan als er een helder en volledig kader bestaat om de betrouwbaarheid ervan tot stand te brengen.

- (12) Dat is in onze optiek de manier waarop Europa zich kan positioneren als thuisbasis en voortrekker voor geavanceerde en ethische technologie. Wij, als Europese burgers, zullen de vruchten plukken van betrouwbare KI op een manier die strookt met onze basiswaarden van respect voor mensenrechten, democratie en de rechtsstaat.

Betrouwbare KI

- (13) Betrouwbaarheid is een randvoorwaarde voor mensen en samenlevingen om KI-systemen te kunnen ontwikkelen, installeren en gebruiken. Als KI-systemen – en de mensen erachter – niet aantoonbaar te vertrouwen zijn, kan dit ongewilde gevolgen hebben en kan het gebruik ervan worden belemmerd, waardoor de potentieel enorme sociale en economische voordelen van KI-systemen niet kunnen worden verwezenlijkt. Om Europa te helpen deze voordelen te verwezenlijken is onze visie het gebruik van ethiek als kernpijler om betrouwbare KI te garanderen en af te stemmen.
- (14) Vertrouwen in de ontwikkeling, de installatie en het gebruik van KI-systemen heeft niet alleen te maken met de inherente kenmerken van de technologie, maar ook met de eigenschappen van de sociaal-technische systemen waarvoor KI-systemen worden gebruikt⁹. Net als bij (verlies van) vertrouwen in de luchtvaart, kernenergie of voedselveiligheid zijn het niet slechts componenten van het KI-systeem die al dan niet vertrouwen wekken, maar het systeem in de algehele context. Streven naar betrouwbare KI heeft niet alleen te maken met de betrouwbaarheid van het KI-systeem zelf, maar vereist een holistische en systemische benadering, waarbij de betrouwbaarheid van alle actoren en processen die gedurende de volledige levenscyclus van het systeem deel uitmaken van de sociaal-technische context ervan, wordt meegenomen.
- (15) Betrouwbare KI bestaat uit **drie componenten**, waaraan gedurende de volledige levenscyclus van het systeem moet worden voldaan:
1. de KI moet **wettig** zijn, door te zorgen dat aan alle toepasselijke wet- en regelgeving wordt voldaan;
 2. de KI moet **ethisch** zijn, door te zorgen dat de ethische beginselen en waarden worden nageleefd; en
 3. de KI moet **robuust** zijn, uit zowel technisch als sociaal oogpunt, aangezien KI-systemen ongewild schade kunnen aanrichten, zelfs al zijn de bedoelingen goed.
- (16) Elk van deze drie componenten is nodig, maar op zichzelf niet voldoende om betrouwbare KI te bewerkstelligen¹⁰. In het ideale scenario sluiten ze alle drie op elkaar aan en valt de werking ervan gedeeltelijk samen. In de praktijk kunnen er echter spanningen bestaan tussen deze elementen (soms kunnen bijvoorbeeld het toepassingsgebied en de inhoud van de bestaande wetgeving niet in lijn zijn met de ethische normen). Als samenleving hebben wij de individuele en collectieve verantwoordelijkheid ervoor te zorgen dat alle drie de componenten bijdragen aan de totstandbrenging van betrouwbare KI¹¹.
- (17) Een betrouwbare benadering is essentieel om "verantwoordelijke concurrentie" mogelijk te maken. Hiermee wordt de grondslag geboden op basis waarvan iedereen die met KI-systemen te maken heeft, erop kan vertrouwen dat het ontwerp, de ontwikkeling en het gebruik ervan wettig, ethisch en robuust zijn. Met deze richtsnoeren, die zijn bedoeld om verantwoordelijke en duurzame KI-innovatie in Europa te stimuleren, wordt beoogd van ethiek een kernpijler te maken op basis waarvan een unieke benadering van KI tot stand komt, een benadering die erop gericht is de ontwikkeling van zowel individuele personen als het algemeen belang van de samenleving te bevorderen, versterken en beschermen. Wij zijn ervan overtuigd dat Europa zich

⁹ Deze systemen omvatten mensen, staatsactoren, ondernemingen, infrastructuur, software, protocollen, normen, governance, bestaande wetten, toezichtsmechanismen, stimuleringsstructuren, controleprocedures, beste verslagleggingspraktijken en meer.

¹⁰ Dit sluit niet uit dat aanvullende voorwaarden noodzakelijk kunnen zijn/worden.

¹¹ Dit kan ook betekenen dat de wetgevers of beleidsmakers de geschiktheid van de bestaande wetgeving moeten onderzoeken, indien deze mogelijk niet meer in lijn is met de ethische beginselen.

hierdoor zal kunnen profileren als mondiale voortrekker op het gebied van geavanceerde KI die ons individuele en collectieve vertrouwen verdient. Alleen door betrouwbaarheid te garanderen zullen Europeanen ten volle kunnen profiteren van de voordelen van KI-systemen, in de wetenschap dat er maatregelen zijn getroffen om hen tegen de potentiële gevaren ervan te beschermen.

- (18) Zoals het gebruik van KI-systemen niet ophoudt bij de nationale grens, geldt dit ook voor de gevolgen ervan. Daarom zijn er mondiale oplossingen nodig voor de mondiale kansen en uitdagingen die KI met zich meebrengt. Wij moedigen alle belanghebbenden dan ook aan om te werken aan een mondiaal kader voor betrouwbare KI door internationale consensus te creëren en tegelijk onze op grondrechten gebaseerde benadering te bevorderen en te handhaven.

Doelpubliek en toepassingsgebied

- (19) Deze richtsnoeren zijn gericht aan alle KI-belanghebbenden die KI ontwerpen, ontwikkelen, installeren, toepassen, gebruiken of erdoor worden beïnvloed, met inbegrip van onder meer bedrijven, organisaties, onderzoekers, overheidsdiensten, overheidsinstanties, instellingen, maatschappelijke organisaties, individuele personen, werknemers en consumenten. Belanghebbenden die zich inzetten voor de totstandbrenging van betrouwbare KI, kunnen vrijwillig kiezen om deze richtlijnen te gebruiken als methode voor de operationalisering van hun inzet, met name door de praktische controlelijst uit hoofdstuk III te gebruiken bij het ontwikkelings- en installatieproces van hun KI-systemen. Deze controlelijst kan ook een aanvulling vormen op – en dus worden verwerkt in – bestaande controleprocessen.
- (20) Deze richtsnoeren zijn bedoeld als ondersteuning voor KI-toepassingen in het algemeen waarbij wordt voorzien in een horizontale grondslag voor de totstandbrenging van betrouwbare KI. **Verschillende situaties brengen echter verschillende uitdagingen met zich mee.** KI-aanbevelingssystemen voor muziek leveren niet dezelfde ethische vragen op als KI-systemen waarin kritieke medische behandelingen worden voorgesteld. Het gebruik van KI-systemen binnen een relatie tussen bedrijf en consument, tussen bedrijven, tussen werkgever en werknemer of tussen overheid en burger – of, meer in het algemeen: in verschillende sectoren of situaties – levert eveneens verschillende kansen en uitdagingen op. Gezien de contextgebondenheid van KI-systemen wordt daarom erkend dat de uitvoering van deze richtsnoeren op de betreffende KI-toepassing moet worden afgestemd. Bovendien moet de noodzaak van een aanvullende sectorale benadering als aanvulling op het algemenere horizontale kader dat in dit document wordt voorgesteld, worden onderzocht.

Om meer inzicht te krijgen in de manier waarop deze richtsnoeren op horizontaal niveau kunnen worden uitgevoerd, alsook in de zaken waarvoor een sectorale benadering nodig is, nodigen we alle belanghebbenden uit om de controlelijst voor betrouwbare KI (hoofdstuk III), waarin dit kader wordt geoperationaliseerd, te testen en ons feedback te geven. Op basis van de middels deze testfase verzamelde feedback zullen wij de controlelijst uit deze richtsnoeren per begin 2020 herzien. De testfase gaat in de zomer van 2019 van start en duurt tot het einde van het jaar. Alle geïnteresseerde belanghebbenden kunnen deelnemen door hun belangstelling kenbaar te maken via de Europese KI-alliantie.

B. EEN KADER VOOR BETROUWBARE KI

- (21) In deze richtsnoeren wordt een kader uiteengezet voor de totstandbrenging van betrouwbare KI op basis van grondrechten zoals vervat in het Handvest van de grondrechten van de Europese Unie (het Handvest) en in de relevante internationale mensenrechtenwetgeving. Hieronder bespreken we kort de drie componenten van betrouwbare KI.

Wettige KI

- (22) KI-systemen zijn niet werkzaam binnen een wetteloze wereld. Een aantal juridisch bindende regels op Europees, nationaal en internationaal niveau is al van toepassing op of relevant voor de ontwikkeling, de

installatie en het gebruik van KI-systemen. Relevante wettelijke bronnen zijn onder meer het primaire EU-recht (de verdragen van de Europese Unie en het Handvest van de grondrechten van de Europese Unie) en het secundaire EU-recht (zoals de algemene verordening gegevensbescherming (AVG), anti-discriminatie richtlijnen, de machinerichtlijn, de richtlijn inzake productaansprakelijkheid, de verordening betreffende het vrije verkeer van niet-persoonsgebonden gegevens, het consumentenrecht en de richtlijnen inzake veiligheid en gezondheid op het werk), maar ook de mensenrechtenverdragen van de VN en de verdragen van de Raad van Europa (zoals het Europees Verdrag tot bescherming van de rechten van de mens) en allerlei wetgeving van de EU-lidstaten. Naast horizontaal toepasselijke regels bestaan er ook verschillende domeinspecifieke regels die van toepassing zijn op specifieke KI-toepassingen (zoals de verordening betreffende medische hulpmiddelen in de gezondheidszorgsector).

- (23) De wetgeving bevat zowel positieve als negatieve verplichtingen. Dat betekent dat deze niet alleen moet worden uitgelegd in verband met wat er *niet mag*, maar ook in verband met wat er *moet*. In de wetgeving worden niet alleen bepaalde handelingen verboden, maar ook andere mogelijk gemaakt. Met betrekking daartoe kan worden opgemerkt dat het Handvest artikelen bevat over de "vrijheid van ondernemerschap" en de "vrijheid van kunsten en wetenschappen", alsook artikelen waarin wordt ingegaan op gebieden waarmee wij beter bekend zijn als we de betrouwbaarheid van KI proberen te waarborgen, zoals gegevensbescherming en non-discriminatie.
- (24) In deze richtsnoeren wordt niet expliciet ingegaan op de eerste component van betrouwbare KI (wettige KI). Er wordt veeleer beoogd richtsnoeren te bieden voor het bevorderen en waarborgen van de tweede en derde component (ethische en robuuste KI). Hoewel die laatste twee tot op zekere hoogte al in de bestaande wetgeving zijn opgenomen, kan de volledige realisatie ervan verder gaan dan de bestaande wettelijke verplichtingen.
- (25) Niets in dit document mag worden uitgelegd of opgevat als juridisch advies of richtsnoeren voor de manier waarop toepasselijke bestaande wettelijke normen en voorschriften kunnen worden nageleefd. Door niets in dit document worden aan derden wettelijke rechten toegekend of wettelijke verplichtingen opgelegd. Wij wijzen er echter op dat het de plicht van iedere natuurlijke of rechtspersoon is om de wetgeving na te leven – ongeacht of die op dit moment al geldt of in de toekomst wordt vastgesteld op basis van de ontwikkeling van KI. Voor het vervolg van deze richtsnoeren wordt verondersteld dat **alle wettelijke rechten en verplichtingen die van toepassing zijn op de processen en activiteiten in verband met het ontwikkelen, installeren en gebruiken van KI, blijven gelden en in acht moeten worden genomen.**

Ethische KI

- (26) Voor de verwezenlijking van betrouwbare KI is naleving van de wetgeving slechts een van de drie benodigde componenten. De wetgeving is niet altijd actueel in verband met technologische ontwikkelingen, komt soms niet overeen met ethische normen of kan simpelweg niet geschikt zijn voor bepaalde kwesties. Om betrouwbaar te zijn moeten KI-systemen daarom ook ethisch zijn, door afstemming op ethische normen te waarborgen.

Robuuste KI

- (27) Zelfs als een ethisch doel wordt gewaarborgd, moeten individuele personen en de samenleving ervan overtuigd zijn dat KI-systemen niet ongewild schade zullen aanrichten. Deze systemen moeten op een veilige, zekere en betrouwbare manier werken en er moeten voorzorgsmaatregelen worden getroffen om onbedoelde negatieve gevolgen te voorkomen. Het is dus belangrijk om ervoor te zorgen dat KI-systemen robuust zijn. Dit is zowel vanuit technisch oogpunt (zorgen voor de technische robuustheid van het systeem zoals gepast in een bepaalde context, bijvoorbeeld het toepassingsgebied of de fase van de levenscyclus) als vanuit sociaal oogpunt (rekening houdend met de context en omgeving waarin het systeem werkzaam is) nodig. Ethische en robuuste KI zijn dus nauw met elkaar verweven en vullen elkaar aan. De in hoofdstuk I beschreven beginselen en de daaruit in hoofdstuk II afgeleide vereisten hebben betrekking op beide componenten.

Het kader

- (28) De richtsnoeren in dit document worden in drie lagen met verschillend abstractieniveau gegeven: van het meest abstract in hoofdstuk I tot het meest concreet in hoofdstuk III:
- I) Grondslagen van betrouwbare KI.** In hoofdstuk I worden de grondslagen van betrouwbare KI en de benadering op basis van grondrechten¹² uiteengezet. De ethische beginselen die moeten worden gevolgd om ethische en robuuste KI te waarborgen, worden vastgesteld en beschreven.
- II) Betrouwbare KI verwezenlijken.** In hoofdstuk II worden deze ethische beginselen vertaald naar zeven vereisten die bij KI-systemen moeten worden toegepast en waaraan ze gedurende hun volledige levenscyclus moeten voldoen. Ook worden er zowel technische als niet-technische methoden aangereikt die voor de toepassing ervan kunnen worden gebruikt.
- III) Betrouwbare KI controleren.** Beroepsbeoefenaars op het gebied van KI verwachten concrete richtsnoeren. Daarom wordt in hoofdstuk III een voorlopige en niet-uitputtende controlelijst voor betrouwbare KI gegeven waarmee de vereisten uit hoofdstuk II kunnen worden geoperationaliseerd. Deze controle moet worden afgestemd op de toepassing van het specifieke systeem.
- (29) In het laatste gedeelte van het document worden voordelige kansen en punten van zorg in verband met KI-systemen uiteengezet die in overweging moeten worden genomen en waarover we verdere discussies willen stimuleren.
- (30) De structuur van de richtsnoeren wordt hieronder in figuur 1 geïllustreerd.

¹² Grondrechten liggen ten grondslag aan zowel internationale als EU-mensenrechtenwetgeving en onderbouwen de juridisch afdwingbare rechten die door de EU-verdragen en het Handvest van de grondrechten van de EU worden gegarandeerd. Omdat ze juridisch bindend zijn, valt naleving van de grondrechten onder de eerste component van betrouwbare KI: "wettige KI". Grondrechten kunnen echter ook worden begrepen als een weergave van bijzondere morele rechten die alle personen hebben op grond van hun menselijkheid, ongeacht de juridisch bindende status ervan. In die zin zijn ze ook onderdeel van de tweede component van betrouwbare KI: "ethische KI".



Figuur 1: De richtsnoeren als een kader voor betrouwbare KI

I. Hoofdstuk I: Grondslagen van betrouwbare KI

- (31) In dit hoofdstuk worden de grondslagen van betrouwbare KI uiteengezet, die gebaseerd zijn op de grondrechten en worden weergegeven in vier ethische beginselen die moeten worden gevolgd om ethische en robuuste KI te waarborgen. Dit hoofdstuk is sterk gebaseerd op het domein van de ethiek.
- (32) KI-ethiek is een subdomein van toegepaste ethiek dat gericht is op de ethische kwesties rondom de ontwikkeling, de installatie en het gebruik van KI. Het belangrijkste is om te bepalen hoe KI het leven van mensen kan verbeteren of zorgen kan opleveren, of het nu gaat om de levenskwaliteit of om de menselijke autonomie en vrijheid die nodig zijn voor een democratische samenleving.
- (33) Ethische reflectie op KI-technologie kan meerdere doelen hebben. Ten eerste kan hiermee worden gestimuleerd dat er wordt nagedacht over de noodzaak om personen en groepen op het meest basale niveau te beschermen. Ten tweede kunnen hiermee nieuwe soorten innovatie worden gestimuleerd waarmee ethische waarden worden bevorderd, zoals innovaties die bijdragen aan de verwezenlijking van de duurzameontwikkelingsdoelstellingen van de VN¹³, die stevig verankerd zijn in de komende EU-agenda voor 2030¹⁴. Hoewel dit document voornamelijk gaat over het eerste van deze doelen, moet het mogelijke belang van ethiek voor het tweede doel niet worden onderschat. Door betrouwbare KI kunnen mensen sterker tot bloei komen en kan het collectieve welzijn worden verbeterd, doordat er welvaart, waardecreatie en vermogensmaximalisatie worden gegenereerd. Betrouwbare KI kan bijdragen aan de verwezenlijking van een rechtvaardige samenleving door de verbetering van de gezondheid en het welzijn van burgers zo te stimuleren dat gelijkheid bij de verdeling van economische, maatschappelijke en politieke kansen wordt bevorderd.
- (34) Daarom is het noodzakelijk dat we begrijpen hoe we de ontwikkeling, de installatie en het gebruik van KI het best kunnen ondersteunen als we willen zorgen dat iedereen goed gedijt in een op KI gebaseerde wereld, en als we een betere toekomst willen creëren, maar ook wereldwijd willen blijven concurreren. Zoals elke krachtige technologie brengt ook het gebruik van KI-systemen in onze samenleving een aantal ethische uitdagingen met zich mee, bijvoorbeeld in verband met het effect ervan op mensen en de samenleving, de capaciteit voor besluitvorming en de veiligheid. Als we steeds vaker de hulp gaan inroepen van of beslissingen gaan uitbesteden aan KI-systemen, moeten we zorgen dat deze systemen rechtvaardig zijn wat hun effect op het leven van mensen betreft, dat ze in overeenstemming zijn met onbetwistbare waarden en in staat in overeenstemming daarmee te functioneren, en dat er geschikte verantwoordingsprocessen zijn die daarvoor zorgen.
- (35) Europa moet bepalen welke normatieve visie op een toekomst vol KI het wil realiseren, en vervolgens welke definitie van KI er in Europa moet worden onderzocht, ontwikkeld, geïnstalleerd en gebruikt om deze visie te verwezenlijken. Met dit document willen wij daaraan bijdragen door het begrip "betrouwbare KI" te introduceren, dat in onze optiek de juiste manier vormt om te bouwen aan een toekomst met KI. In een toekomst waarin democratie, de rechtsstaat en grondrechten de basis vormen voor KI-systemen en waarin die systemen de democratische cultuur voortdurend verbeteren en beschermen, zal ook een omgeving kunnen worden gecreëerd waarin innovatie en verantwoordelijke concurrentie kunnen gedijen.
- (36) Een domeinspecifieke ethische code – hoe consistent, ontwikkeld en verfijnd toekomstige versies ervan ook mogen zijn – kan nooit dienen als vervanging van ethisch denken, aangezien dit altijd gevoelig moet blijven voor contextuele details die niet in algemene richtsnoeren kunnen worden vastgelegd. Voor betrouwbare KI moeten we niet alleen een reeks voorschriften ontwikkelen, maar ook een ethische cultuur en houding creëren en onderhouden door middel van openbare discussies, onderwijs en praktische vormen van leren.

1. Grondrechten als morele en wettelijke rechten

¹³ https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_nl

¹⁴ <https://sustainabledevelopment.un.org/?menu=1300>

- (37) Wij geloven in een benadering van KI-ethiek op basis van de grondrechten die zijn vastgelegd in de EU-verdragen¹⁵, het Handvest van de grondrechten van de EU (het Handvest) en de internationale mensenrechtenwetgeving¹⁶. Eerbiediging van grondrechten, binnen een kader van democratie en de rechtsstaat, biedt de meest veelbelovende grondslagen voor het vaststellen van abstracte ethische beginselen en waarden die in het kader van KI kunnen worden geoperationaliseerd.
- (38) In de EU-verdragen en het Handvest wordt een reeks grondrechten voorgeschreven. De lidstaten en instellingen van de EU zijn wettelijk verplicht deze te eerbiedigen bij de uitvoering van EU-wetgeving. Deze rechten worden in het Handvest beschreven op basis van waardigheid, vrijheden, gelijkheid en solidariteit, de rechten van burgers en gerechtigheid. Van de gemeenschappelijke grondslag van deze rechten kan men zeggen dat deze is geworteld in respect voor de menselijke waardigheid, waarmee een benadering wordt weergegeven waarbij de mens centraal staat doordat deze een uniek en onvervreemdbaar moreel primaat heeft binnen het maatschappelijk, politieke, economische en sociale domein¹⁷.
- (39) Hoewel de in het Handvest beschreven rechten juridisch bindend zijn¹⁸, is het belangrijk om te erkennen dat grondrechten niet in alle gevallen volledige wettelijke bescherming bieden. Voor het Handvest is het bijvoorbeeld belangrijk om te benadrukken dat het toepassingsgebied ervan beperkt is tot bepaalde domeinen van de EU-wetgeving. De internationale mensenrechtenwetgeving en in het bijzonder het Europees Verdrag tot bescherming van de rechten van de mens zijn juridisch bindend voor de EU-lidstaten, ook op gebieden die buiten het toepassingsgebied van de EU-wetgeving vallen. Tegelijkertijd moet er worden benadrukt dat grondrechten ook worden verleend aan personen en (tot op zekere hoogte) aan groepen op grond van hun morele status als mens, los van de rechtskracht ervan. Opgevat als juridisch afdwingbare rechten vallen grondrechten daarom onder de eerste component van betrouwbare KI (wettige KI), die de naleving van de wetgeving waarborgt. Opgevat als de rechten van alle mensen, geworteld in de inherente morele status van mensen, ondersteunen ze ook de tweede component van betrouwbare KI (ethische KI), die te maken heeft met ethische normen die niet per definitie juridisch bindend zijn, maar wel cruciaal voor betrouwbaarheid. Omdat dit document niet is bedoeld als richtsnoer voor de eerste component, hebben verwijzingen naar grondrechten in het kader van deze niet-bindende richtsnoeren betrekking op de tweede component.

2. Van grondrechten naar ethische beginselen

2.1 Grondrechten als een basis voor betrouwbare KI

- (40) Van de volledige reeks in de internationale mensenrechtenwetgeving, de EU-verdragen en het Handvest beschreven ondeelbare rechten zijn de onderstaande groepen grondrechten in het bijzonder geschikt om toe te passen op KI-systemen. Veel van deze rechten zijn, in bepaalde omstandigheden, juridisch afdwingbaar in de EU, dus naleving van de voorwaarden ervan is wettelijk verplicht. Zelfs wanneer naleving van de juridisch afdwingbare grondrechten is verwezenlijkt, kan ethische reflectie ons helpen te begrijpen welke rol grondrechten en de onderliggende waarden daarvan spelen bij de ontwikkeling, de installatie en het gebruik van KI. Ook kan deze reflectie verfijndere richtsnoeren opleveren om te bepalen wat we met technologie *zouden moeten* doen in plaats van wat we er (momenteel) mee *kunnen* doen.
- (41) **Respect voor de menselijke waardigheid.** Onder menselijke waardigheid valt het idee dat ieder mens een

¹⁵ De EU is gebaseerd op een constitutionele verbintenis om de grondrechten en ondeelbare rechten van mensen te beschermen, eerbiediging van de rechtsstaat te garanderen, democratische vrijheid te waarborgen en het algemeen belang te bevorderen. Deze rechten zijn terug te vinden in de artikelen 2 en 3 van het Verdrag betreffende de Europese Unie en in het Handvest van de grondrechten van de EU.

¹⁶ In andere juridische instrumenten worden deze verbintenissen weergegeven en verder uitgewerkt, bijvoorbeeld in het Europees Sociaal Handvest van de Raad van Europa en in specifieke wetgeving zoals de algemene verordening gegevensbescherming van de EU.

¹⁷ Er moet worden opgemerkt dat voor toewijding aan KI waarbij de mens centraal staat, en voor de verankering daarvan in grondrechten, collectieve maatschappelijke en constitutionele grondslagen nodig zijn waarbij individuele vrijheid en respect voor de menselijke waardigheid zowel praktisch mogelijk als zinvol zijn en er geen onnodig individualistisch beeld van de mens wordt geschetst.

¹⁸ Het Handvest is, op grond van artikel 51 ervan, van toepassing op de EU-instellingen en EU-lidstaten bij de uitvoering van EU-wetgeving.

"intrinsieke waarde" heeft die nooit door anderen – of door nieuwe technologieën als KI-systemen – mag worden beperkt, in gevaar worden gebracht of onderdrukt.¹⁹ In het kader van KI houdt respect voor de menselijke waardigheid in dat alle mensen worden behandeld met het respect waar zij als morele *subjecten*, niet slechts *objecten* die kunnen worden ontleed, gesorteerd, gecijferd, gedreven, geconditioneerd of gemanipuleerd, recht op hebben. KI-systemen moeten dus worden ontwikkeld op een manier waarbij de fysieke en geestelijke integriteit van mensen, hun persoonlijke en culturele gevoel van identiteit en de vervulling van hun basisbehoeften worden gerespecteerd, gediend en beschermd²⁰.

- (42) **Vrijheid van het individu.** Mensen moeten vrij zijn om hun eigen levenskeuzen te maken. Dat betekent vrijheid van soevereine inmenging, maar er is ook ingrijpen voor nodig door de overheid en niet-gouvernementele organisaties, om te zorgen dat personen die het risico lopen te worden uitgesloten, gelijke toegang hebben tot de voordelen en mogelijkheden van KI. In het kader van KI is voor vrijheid van het individu beperking nodig van (in)directe illegale dwang, bedreigingen voor de mentale autonomie en geestelijke gezondheid, ongerechtvaardigd toezicht, misleiding en oneerlijke manipulatie. Sterker nog, vrijheid van het individu betekent een verbintenis om mensen *nóg* meer controle over hun eigen leven te geven, met (onder andere) rechten als bescherming van de vrijheid van ondernemerschap, de vrijheid van kunsten en wetenschappen, de vrijheid van meningsuiting, het recht op een privéleven en privacy en de vrijheid van vergadering en vereniging.
- (43) **Respect voor democratie, justitie en de rechtsstaat.** Alle macht van de overheid in constitutionele democratieën moet wettelijk worden toegekend en beperkt. KI-systemen moeten democratische processen onderhouden en bevorderen en de pluraliteit van de waarden en levenskeuzen van mensen respecteren. Ze mogen democratische processen, menselijk overleg of democratische kiesstelsels niet ondermijnen. Ook moet in KI-systemen een verbintenis worden verwerkt om te zorgen dat ze niet op manieren werken waardoor op de rechtsstaat berustende basisbeginselen en de toepasselijke wet- en regelgeving worden ondermijnd en dat een eerlijke procesgang en gelijkheid voor de wet worden gewaarborgd.
- (44) **Gelijkheid, non-discriminatie en solidariteit – inclusief de rechten van mensen die het risico lopen te worden uitgesloten.** Gelijk respect voor de morele waarde en waardigheid van alle mensen moet worden gewaarborgd. Dat gaat verder dan non-discriminatie, waarbij het wordt getolereerd verschil te maken tussen ongelijke situaties op basis van objectieve rechtvaardigingen. In het kader van KI betekent gelijkheid dat de activiteiten van het systeem geen vertekende resultaten mogen opleveren (de gegevens die worden gebruikt om KI-systemen te trainen, moeten bijvoorbeeld zo inclusief mogelijk zijn, zodat verschillende bevolkingsgroepen worden vertegenwoordigd). Hiervoor is ook passend respect nodig voor potentieel kwetsbare personen en groepen²¹, zoals werknemers, vrouwen, mensen met een beperking, etnische minderheden, kinderen, consumenten of anderen die het risico lopen te worden uitgesloten.
- (45) **Rechten van burgers.** Burgers hebben allerlei rechten, waaronder kiesrecht, het recht op behoorlijk bestuur, het recht op inzage in openbare documenten en het recht een verzoekschrift tot de overheid te richten. KI-systemen bieden veel potentieel voor de verbetering van de schaal en efficiëntie van de overheid bij het leveren van publieke goederen en diensten aan de maatschappij. Tegelijk kunnen KI-toepassingen ook negatieve gevolgen hebben voor de rechten van burgers en moeten deze rechten worden beschermd. Wanneer hier de term "rechten van burgers" wordt gebruikt, wordt daarmee niet beoogd de rechten van inwoners van derde landen en van personen die onregelmatig (of illegaal) in de EU verblijven, die ook rechten hebben op grond van internationale wetgeving en dus ook op het gebied van KI, te ontkennen of verwaarlozen.

¹⁹ C. McCrudden, Human Dignity and Judicial Interpretation of Human Rights, *EJIL*, 19(4), 2008.

²⁰ Zie voor een vergelijkbare interpretatie van "menselijke waardigheid" E. Hilgendorf, Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity, in: D. Grimm, A. Kemmerer, C. Möllers (red.), *Human Dignity in Context. Explorations of a Contested Concept*, 2018, blz. 325 ff.

²¹ Zie de woordenlijst voor een beschrijving van de term zoals deze in dit document wordt gebruikt.

2.2 Ethische beginselen in het kader van KI-systemen²²

- (46) Veel publieke, particuliere en maatschappelijke organisaties hebben inspiratie voor de creatie van ethische kaders voor KI uit grondrechten gehaald²³. In de EU heeft de Europese Groep ethiek van de exacte wetenschappen en de nieuwe technologieën (EGE) een reeks van negen basisbeginselen voorgesteld op basis van de fundamentele waarden die zijn vastgelegd in de EU-verdragen en het Handvest van de grondrechten van de EU.²⁴ Wij bouwen voort op dit werk. Daarbij erkennen we het grootste gedeelte van de tot dusver door verschillende groepen voorgestelde beginselen en verhelderen we de doelen die men met alle beginselen beoogt te stimuleren en ondersteunen. Deze ethische beginselen kunnen inspiratie opleveren voor nieuwe en specifieke regelgevingsinstrumenten, hulp bieden bij de uitleg van grondrechten naarmate onze sociaal-technische omgeving in de loop van de tijd verandert, en richting geven aan de motivering voor de ontwikkeling, het gebruik en de toepassing van KI-systemen – waarbij ze zich dynamisch aanpassen naarmate de samenleving zelf verandert.
- (47) KI-systemen moeten het individuele en collectieve welzijn verbeteren. In dit gedeelte worden **vier ethische beginselen** genoemd die zijn geworteld in grondrechten en die moeten worden nageleefd om te zorgen dat KI-systemen op betrouwbare wijze worden ontwikkeld, geïnstalleerd en gebruikt. Ze worden gespecificeerd als **ethische geboden**, en beroepsbeoefenaars op het gebied van KI moeten er altijd naar streven deze na te leven. Zonder een hiërarchie te willen opleggen vermelden we de beginselen hieronder in dezelfde volgorde waarin de grondrechten waarop ze zijn gebaseerd, in het Handvest staan²⁵.
- (48) Het betreft de beginselen van:
- (i) respect voor de menselijke autonomie;
 - (ii) preventie van schade;
 - (iii) rechtvaardigheid;
 - (iv) verantwoording.
- (49) Veel van deze beginselen zijn al grotendeels weergegeven in de bestaande wettelijke voorschriften die verplicht moeten worden nageleefd, en vallen dus ook binnen de reikwijdte van "wettige KI", de eerste component van betrouwbare KI²⁶. Zoals hierboven beschreven gaat naleving van de ethische beginselen, hoewel deze in veel wettelijke verplichtingen worden weerspiegeld, echter verder dan formele naleving van bestaande wetgeving²⁷.

- Het beginsel van respect voor de menselijke autonomie

- (50) De grondrechten waarop de EU is gefundeerd, zijn erop gericht respect voor de vrijheid en autonomie van

²² Deze beginselen zijn ook van toepassing op de ontwikkeling, de installatie en het gebruik van andere technologieën en hebben dus niet specifiek betrekking op KI-systemen. Wij hebben geprobeerd in dit gedeelte de specifieke relevantie ervan in een KI-gerelateerde context uiteen te zetten.

²³ Door op grondrechten te bouwen kan ook de onzekerheid worden beperkt wat regelgeving betreft, aangezien men kan voortbouwen op tientallen jaren van bescherming van grondrechten in de EU. Daardoor ontstaat duidelijkheid, leesbaarheid en voorzienbaarheid.

²⁴ Recenter heeft de taskforce van AI4People een enquête gehouden over de genoemde EGE-beginselen, alsook over 36 andere ethische beginselen die tot dusver naar voren zijn gebracht, en deze beginselen onder vier overkoepelende beginselen ingedeeld: L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), "AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", *Minds and Machines* 28(4): blz. 689-707.

²⁵ Respect voor de menselijke autonomie wordt sterk geassocieerd met het recht op menselijke waardigheid en vrijheid (weergegeven in de artikelen 1 en 6 van het Handvest). De preventie van schade houdt sterk verband met de bescherming van de fysieke of geestelijke integriteit (weergegeven in artikel 3). Rechtvaardigheid is nauw verbonden aan het recht op non-discriminatie, solidariteit en gerechtigheid (weergegeven in artikel 21 en verder). Verantwoording en verantwoordelijkheid zijn nauw verbonden aan de rechten die verband houden met rechtspleging (zoals weergegeven in artikel 47).

²⁶ Denk bijvoorbeeld aan de AVG of de EU-wetgeving inzake bescherming van consumenten.

²⁷ Zie voor meer informatie over dit onderwerp bijvoorbeeld L. Floridi, *Soft Ethics and the Governance of the Digital*, *Philosophy & Technology*, maart 2018, jaargang 31, editie 1, blz. 1–8.

mensen te waarborgen. Mensen die met KI-systemen werken, moeten hun volledige en effectieve zelfbeschikking kunnen behouden en kunnen deelnemen aan het democratische proces. KI-systemen mogen mensen niet onterecht onderwerpen, dwingen, misleiden, manipuleren, conditioneren of drijven. Ze moeten veeleer worden ontworpen om de menselijke cognitieve, sociale en culturele vaardigheden te vergroten, aan te vullen en te versterken. De verdeling van functies tussen mensen en KI-systemen moet gebeuren volgens ontwerpbeginselen waarbij de mens centraal staat, en moet zinvolle mogelijkheden voor menselijke keuze openlaten. Er moet dus worden gezorgd voor menselijk toezicht²⁸ en menselijke controle op de werkprocessen in KI-systemen. De werksfeer kan door KI-systemen ook fundamenteel veranderen. Deze moet mensen ondersteunen in de werkomgeving en gericht zijn op de creatie van zinvol werk.

- Het beginsel van preventie van schade

- (51) KI-systemen mogen geen schade²⁹ veroorzaken of vergroten of anderszins negatieve gevolgen hebben voor mensen³⁰. Dat betekent bescherming van de waardigheid, alsook de geestelijke en fysieke integriteit van mensen. KI-systemen en de omgeving waarin zij werken, moeten veilig en zeker zijn. Ze moeten technisch robuust zijn en er moet voor worden gezorgd dat ze geen ruimte bieden voor kwaadwillig gebruik. Kwetsbare personen moeten meer aandacht krijgen en moeten worden betrokken bij de ontwikkeling en installatie van KI-systemen. Er moet specifiek aandacht worden besteed aan situaties waarin KI-systemen negatieve gevolgen kunnen veroorzaken of vergroten vanwege ongelijkheid wat betreft macht of beschikking over informatie, bijvoorbeeld tussen werkgevers en werknemers, tussen bedrijven en consumenten of tussen overheden en burgers. Preventie van schade betekent ook dat er rekening moet worden gehouden met de natuurlijke omgeving en alle levende wezens.

- Het beginsel van rechtvaardigheid

- (52) De ontwikkeling, de installatie en het gebruik van KI-systemen moeten rechtvaardig zijn. Wij erkennen dat er veel verschillende interpretaties van rechtvaardigheid bestaan, maar zijn ervan overtuigd dat rechtvaardigheid zowel een inhoudelijke als een procedurele dimensie heeft. De inhoudelijke dimensie impliceert een toezegging om de gelijke en rechtvaardige verdeling van zowel voordelen als kosten te waarborgen en ervoor te zorgen dat personen en groepen vrij zijn van onrechtvaardige vertekening, discriminatie en stigmatisering. Indien onrechtvaardige vertekening kan worden voorkomen, zouden KI-systemen zelfs de maatschappelijke rechtvaardigheid kunnen vergroten. Gelijke kansen wat betreft toegang tot onderwijs, goederen, diensten en technologie moeten ook worden bevorderd. Daarnaast mag het gebruik van KI-systemen nooit tot gevolg hebben dat de (eind)gebruikers worden misleid of worden beperkt in hun keuzevrijheid. Verder impliceert rechtvaardigheid dat beroepsbeoefenaars op het gebied van KI het beginsel van evenredigheid tussen middelen en doelen moeten eerbiedigen en zorgvuldig moeten afwegen hoe ze tegengestelde belangen en doelstellingen in evenwicht kunnen brengen.³¹ De procedurele dimensie van rechtvaardigheid omvat het vermogen om beslissingen die worden genomen door KI-systemen en door de mensen die deze systemen beheren, aan te vechten en er effectief beroep tegen in te stellen.³² Om dat te kunnen doen moet de entiteit

²⁸ Het concept van menselijk toezicht wordt hieronder bij punt 65 verder uitgewerkt.

²⁹ Schade kan individueel of collectief zijn en kan ontastbare schade aan sociale, culturele en politieke omgevingen omvatten.

³⁰ Hieronder valt ook de leefwijze van personen of sociale groepen, om bijvoorbeeld culturele schade te voorkomen.

³¹ Dit heeft te maken met het evenredigheidsbeginsel (weergegeven in de stelregel dat men niet "met een kanon op een mug moet schieten"). Maatregelen die worden genomen om een bepaald doel te bereiken (bijv. de gegevensextractiemaatregelen die worden uitgevoerd om de KI-optimalisatiefunctie te realiseren), moeten worden beperkt tot het strikt noodzakelijke. Het houdt ook in dat, wanneer meerdere maatregelen in aanmerking komen voor het bereiken van een doel, de voorkeur moet worden gegeven aan de maatregel die het minst ingaat tegen de grondrechten en ethische normen (KI-ontwikkelaars moeten bijvoorbeeld altijd de voorkeur geven aan gegevens uit de publieke sector boven persoonsgegevens). Er kan ook worden verwezen naar de evenredigheid tussen gebruiker en installateur, met de rechten van bedrijven (zoals intellectuele eigendom en vertrouwelijkheid) enerzijds en de rechten van de gebruiker anderzijds.

³² Onder meer door gebruik te maken van hun recht van vereniging door lid te worden van een vakvereniging in een werkomgeving, overeenkomstig artikel 12 van het Handvest van de grondrechten van de EU.

die verantwoordelijk is voor de beslissing, identificeerbaar zijn en moet het besluitvormingsproces verklaarbaar zijn.

- Het beginsel van verantwoording

- (53) Verantwoording is cruciaal voor het scheppen en behouden van het vertrouwen van gebruikers in KI-systemen. Dat betekent dat processen transparant moeten zijn, dat de capaciteiten en het doel van KI-systemen openlijk kenbaar moeten worden gemaakt en dat beslissingen – voor zover mogelijk – verklaarbaar moeten zijn aan degenen die er direct of indirect de gevolgen van ondervinden. Zonder die informatie kan een beslissing niet naar behoren worden aangevochten. Het is niet altijd mogelijk om te verklaren waarom een model een bepaald resultaat of een bepaalde beslissing heeft opgeleverd (en welke combinatie van inputfactoren daaraan heeft bijgedragen). Deze gevallen worden "blackbox"-algoritmen genoemd en vereisen speciale aandacht. In die situaties kunnen andere verantwoordingsmaatregelen (zoals traceerbaarheid, controleerbaarheid en transparante communicatie over de capaciteiten van het systeem) nodig zijn, mits het systeem als geheel de grondrechten eerbiedigt. De mate waarin verantwoording nodig is, hangt sterk af van de context en de ernst van de gevolgen, mocht het resultaat onjuist of anderszins onnauwkeurig zijn.³³

2.3 Spanningen tussen de beginselen

- (54) Er kunnen spanningen ontstaan tussen bovengenoemde beginselen en er is geen vaste oplossing om deze te verhelpen. Overeenkomstig de fundamentele toewijding van de EU aan democratische betrokkenheid, een eerlijke procesgang en open politieke participatie moeten er methoden van verantwoordelijk overleg worden vastgesteld om met dergelijke spanningen om te gaan. In verschillende toepassingsdomeinen kunnen *het beginsel van preventie van schade* en *het beginsel van menselijke autonomie* bijvoorbeeld met elkaar in strijd zijn. Neem bijvoorbeeld het gebruik van KI-systemen voor "voorspellend beleid", dat kan helpen de criminaliteit terug te dringen, maar wel op manieren waarbij toezichtsactiviteiten komen kijken die inbreuk maken op de individuele vrijheid en privacy. Verder moeten de algehele voordelen van KI-systemen aanzienlijk groter zijn dan de te voorziene individuele risico's. Hoewel deze beginselen absoluut richtsnoeren bieden voor oplossingen, blijven ze abstracte ethische voorschriften. Van beroepsbeoefenaars op het gebied van KI kan dus niet worden verwacht dat zij de juiste oplossing vinden op basis van bovenstaande beginselen, maar zij moeten ethische dilemma's en afwegingen benaderen aan de hand van logische, empirische reflectie, niet van intuïtie of een willekeurig oordeel. Er kunnen echter situaties zijn waarin er geen ethisch acceptabele compromissen te vinden zijn. Bepaalde grondrechten en bijbehorende beginselen zijn absoluut en kunnen geen onderdeel uitmaken van een compromis (bijv. menselijke waardigheid).

Essentiële richtsnoeren uit hoofdstuk I:

- ✓ Ontwikkel, installeer en gebruik KI-systemen op dusdanige wijze dat de volgende ethische beginselen worden nageleefd: *respect voor menselijke autonomie, preventie van schade, rechtvaardigheid en verantwoording*. Erken de potentiële spanningen tussen deze beginselen en pak deze aan.
- ✓ Schenk bijzondere aandacht aan situaties waarbij kwetsbaardere groepen betrokken zijn, zoals kinderen, mensen met een beperking en andere groepen die historisch gezien kansarm zijn of het risico lopen te worden uitgesloten, en/of aan situaties die worden gekenmerkt door ongelijkheid wat betreft macht of beschikking over informatie, bijvoorbeeld tussen werkgevers en werknemers of tussen bedrijven en consumenten³⁴.
- ✓ Erken en wees u ervan bewust dat KI-systemen de potentie hebben om de mens en de samenleving vele grote voordelen op te leveren, maar dat sommige ook negatieve gevolgen met zich mee kunnen brengen,

³³ Als een KI-systeem onnauwkeurige koopaanbevelingen voortbrengt, levert dat bijvoorbeeld weinig ethische zorgen op, in tegenstelling tot KI-systemen die evalueren of iemand die voor een strafbaar feit is veroordeeld, voorwaardelijk moet worden vrijgelaten.

³⁴ Zie de artikelen 24 tot en met 27 van het Handvest, waarin wordt ingegaan op de rechten van kinderen en ouderen, de integratie van mensen met een beperking en de rechten van werknemers. Zie ook artikel 38 over de bescherming van consumenten.

die mogelijk lastig te voorspellen, vast te stellen of te meten zijn (bijv. gevolgen voor de democratie, de rechtsstaat en de verdelende rechtvaardigheid of voor de menselijke geest). Neem waar nodig passende maatregelen, evenredig aan de omvang van het risico, om deze risico's te beperken.

II. Hoofdstuk II: Betrouwbare KI verwezenlijken

- (55) In dit hoofdstuk worden richtsnoeren geboden voor de toepassing en verwezenlijking van betrouwbare KI, in de vorm van een lijst van zeven vereisten waaraan moet worden voldaan, voortbouwend op de in hoofdstuk I beschreven beginselen. Daarnaast worden de momenteel beschikbare technische en niet-technische methoden voor de uitvoering van deze vereisten in de volledige levenscyclus van het KI-systeem beschreven.

1. Vereisten van betrouwbare KI

- (56) De in hoofdstuk I beschreven beginselen moeten worden vertaald naar concrete vereisten voor de verwezenlijking van betrouwbare KI. Deze vereisten zijn van toepassing op verschillende belanghebbenden die deel uitmaken van de levenscyclus van KI-systemen: ontwikkelaars, installateurs en eindgebruikers, alsook de samenleving in bredere zin. Met ontwikkelaars doelen wij op degenen die onderzoek doen naar KI-systemen en deze ontwerpen en/of ontwikkelen. Met installateurs doelen we op publieke of particuliere organisaties die gebruikmaken van KI-systemen binnen hun bedrijfsprocessen en om producten en diensten aan anderen aan te bieden. Eindgebruikers zijn degenen die, direct of indirect, met het KI-systeem te maken hebben. Onder de samenleving in bredere zin vallen tot slot alle anderen die direct of indirect door KI-systemen worden beïnvloed.
- (57) De verschillende categorieën belanghebbenden spelen verschillende rollen om te zorgen dat aan de vereisten wordt voldaan:
- ontwikkelaars moeten de vereisten uitvoeren en toepassen op de ontwerp- en ontwikkelingsprocessen;
 - installateurs moeten zorgen dat de systemen die zij gebruiken en de producten en diensten die zij aanbieden, aan de vereisten voldoen;
 - eindgebruikers en de samenleving in bredere zin moeten in kennis worden gesteld van deze vereisten en de mogelijkheid hebben te verzoeken dat eraan wordt voldaan.
- (58) De onderstaande lijst van vereisten is niet uitputtend³⁵ en bevat systemische, individuele en maatschappelijke aspecten:

1 Menselijke controle en menselijk toezicht

Omvat grondrechten, menselijke controle en menselijk toezicht

2 Technische robuustheid en veiligheid

Omvat weerbaarheid tegen aanvallen en beveiliging, een uitwijkplan en algemene veiligheid, nauwkeurigheid, betrouwbaarheid en reproduceerbaarheid

3 Privacy en datagovernance

Omvat respect voor privacy, de kwaliteit en integriteit van gegevens en toegang tot gegevens

4 Transparantie

Omvat traceerbaarheid, verklaarbaarheid en communicatie

5 Diversiteit, non-discriminatie en rechtvaardigheid

Omvat het voorkomen van onrechtvaardige vertekening, toegankelijkheid en universeel ontwerp en

³⁵ Zonder een hiërarchie te willen opleggen vermelden we de beginselen hieronder in dezelfde volgorde waarin de beginselen en rechten waarmee ze verband houden, in het Handvest staan.

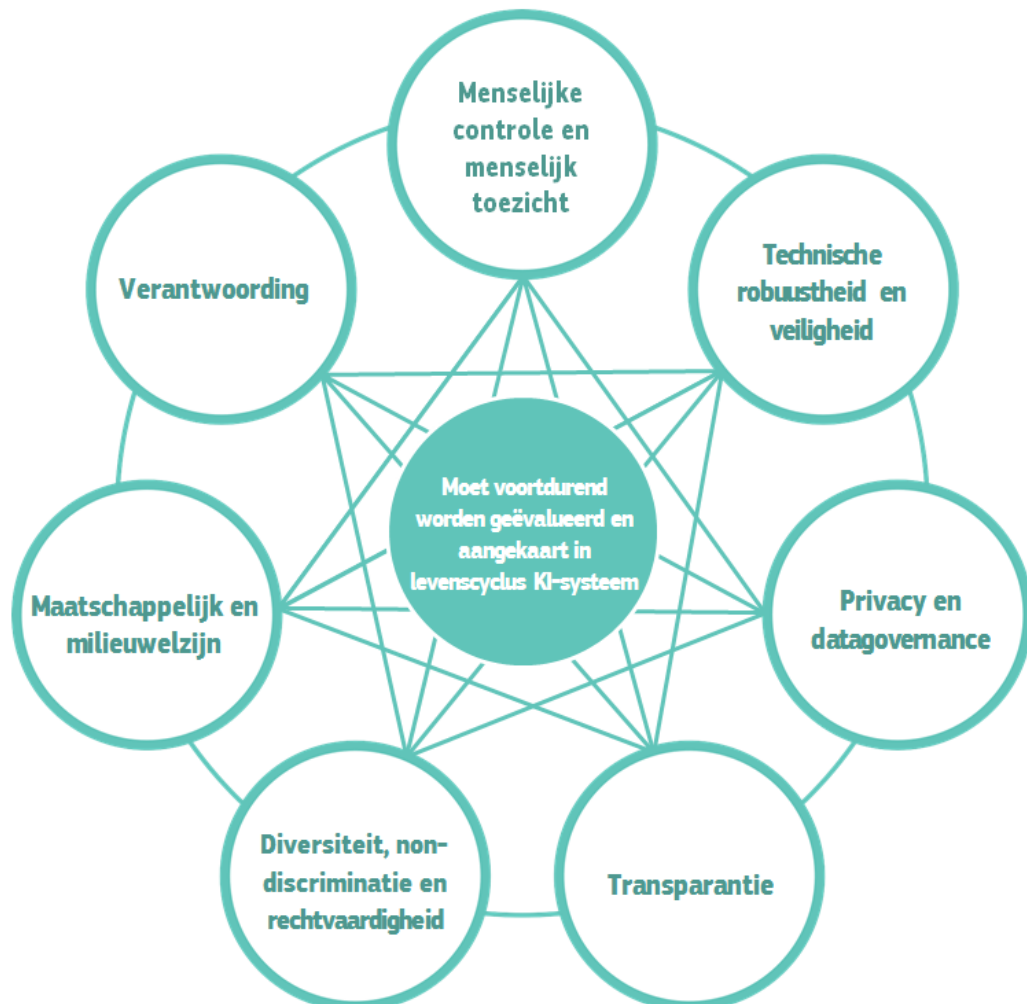
participatie van belanghebbenden

6 Maatschappelijk en milieuwelzijn

Omvat duurzaamheid en milieuvriendelijkheid, sociale gevolgen, de samenleving en de democratie

7 Verantwoording

Omvat controleerbaarheid, minimalisering en verslaglegging van negatieve gevolgen, afwegingen en beroep.



Figuur 2: Onderlinge verhouding van de zeven vereisten: alle vereisten zijn even belangrijk, ondersteunen elkaar en moeten gedurende de levenscyclus van een KI-systeem worden uitgevoerd en geëvalueerd

- (59) Hoewel alle vereisten even belangrijk zijn, moet er rekening worden gehouden met de context en de mogelijke spanningen ertussen wanneer ze in verschillende domeinen en sectoren worden toegepast. De uitvoering van deze vereisten moet plaatsvinden gedurende de volledige levenscyclus van het KI-systeem en hangt af van de specifieke toepassing. De meeste vereisten zijn op alle KI-systemen van toepassing, maar er wordt bijzondere aandacht besteed aan de vereisten die direct of indirect van invloed zijn op personen. Daarom kunnen ze voor sommige toepassingen (bijvoorbeeld in een industriële context) minder belangrijk zijn.
- (60) De bovenstaande vereisten bevatten elementen die in sommige gevallen al in de bestaande wetgeving worden weergegeven. Wij herhalen dat het – overeenkomstig de eerste component van betrouwbare KI – de verantwoordelijkheid van de ontwikkelaars en installateurs van KI-systemen is om te zorgen dat ze voldoen aan hun wettelijke verplichtingen wat betreft zowel horizontaal toepasselijke regels als domeinspecifieke

regelgeving.

- (61) Bij de volgende punten worden alle vereisten verder uitgewerkt.

1. Menselijke controle en menselijk toezicht

- (62) KI-systemen moeten menselijke autonomie en beslissingen ondersteunen, zoals voorgeschreven door het beginsel van *respect voor menselijke autonomie*. Daarvoor is het nodig dat KI-systemen een democratische, florerende en gelijkwaardige samenleving mogelijk maken door de controle van de gebruiker te ondersteunen en grondrechten te bevorderen, en dat ze de mogelijkheid bieden voor menselijk toezicht.
- (63) **Grondrechten.** Zoals veel technologieën kunnen KI-systemen grondrechten evenzeer bevorderen als belemmeren. Mensen kunnen er bijvoorbeeld voordeel van hebben doordat de systemen hen helpen hun persoonsgegevens bij te houden of doordat onderwijs toegankelijker wordt en hun recht op onderwijs dus wordt ondersteund. Gezien de reikwijdte en capaciteit van KI-systemen kunnen deze echter ook negatieve gevolgen hebben voor grondrechten. In situaties waar dergelijke risico's bestaan, moet een effectbeoordeling worden uitgevoerd wat grondrechten betreft. Deze moet voorafgaand aan de ontwikkeling van het systeem worden uitgevoerd en moet een evaluatie omvatten van de vraag of de risico's kunnen worden beperkt of gerechtvaardigd zoals noodzakelijk in een democratische samenleving om de rechten en vrijheden van anderen te respecteren. Daarnaast moeten er mechanismen worden ingesteld om externe feedback te ontvangen ten aanzien van KI-systemen die mogelijk inbreuk maken op grondrechten.
- (64) **Menselijke controle.** Gebruikers moeten onderbouwde, autonome beslissingen kunnen nemen ten aanzien van KI-systemen. Ze moeten de kennis en hulpmiddelen ontvangen om KI-systemen te kunnen begrijpen en er in bevredigende mate mee te kunnen omgaan, en ze moeten, indien nodig, in staat worden gesteld het systeem redelijkerwijs zelf te controleren of aan te vechten. KI-systemen moeten mensen ondersteunen bij het maken van betere, meer onderbouwde keuzen die overeenkomen met hun doelen. Soms kunnen KI-systemen worden geïnstalleerd om menselijk gedrag te vormen en beïnvloeden via mechanismen die lastig te detecteren zijn, omdat ze gebruikmaken van onbewuste processen, met inbegrip van verschillende vormen van onrechtvaardige manipulatie, misleiding, drijven en conditionering, allemaal bedreigingen voor de individuele autonomie. Het algehele beginsel van gebruikersautonomie moet centraal staan bij de functionaliteit van het systeem. Cruciaal daarbij is het recht niet te worden onderworpen aan een uitsluitend op geautomatiseerde verwerking gebaseerd besluit waaraan voor gebruikers rechtsgevolgen zijn verbonden of dat hen anderszins in aanmerkelijke mate treft³⁶.
- (65) **Menselijk toezicht.** Menselijk toezicht helpt om te zorgen dat een KI-systeem de menselijke autonomie niet ondermijnt en geen andere negatieve effecten veroorzaakt. Toezicht kan worden verwezenlijkt via governancemechanismen, zoals een benadering met "human-in-the-loop" (HITL), "human-on-the-loop" (HOTL) of "human-in-command" (HIC). HITL verwijst naar de capaciteit voor menselijke interventie in elke besluitcyclus van het systeem, die in veel gevallen mogelijk noch wenselijk is. HOTL verwijst naar de capaciteit voor menselijke interventie gedurende de ontwerpcyclus van het systeem en het monitoren van de werking van het systeem. HIC verwijst naar de capaciteit om de algehele activiteit van het KI-systeem te overzien (inclusief zijn bredere economische, maatschappelijke, juridische en ethische gevolgen) en het vermogen om te kiezen wanneer en hoe het systeem in iedere specifieke situatie wordt gebruikt. Daarbij kan het bijvoorbeeld gaan om de keuze om een KI-systeem in een bepaalde situatie niet te gebruiken, om een bepaalde mate van menselijk oordeel vast te stellen tijdens het gebruik van het systeem of om te garanderen dat een door het systeem genomen beslissing kan worden herroepen. Verder moet ervoor worden gezorgd dat openbare handhavingsinstanties in staat zijn toezicht te houden overeenkomstig hun mandaat. Toezichtsmechanismen kunnen in verschillende mate nodig zijn om andere veiligheids- en

³⁶

Hier kan worden verwezen naar artikel 22 van de AVG, waar dit recht al is vastgelegd.

controlemaatregelen te ondersteunen, afhankelijk van het toepassingsgebied van het KI-systeem en het potentiële risico. Indien alle andere omstandigheden gelijk blijven, moet een KI-systeem uitgebreider worden getest en is er strengere governance nodig naarmate er minder menselijk toezicht mogelijk is op het systeem.

2. Technische robuustheid en veiligheid

- (66) Een cruciale component van de verwezenlijking van betrouwbare KI is technische robuustheid, die nauw verbonden is met het *beginsel van preventie van schade*. Voor technische robuustheid moeten KI-systemen worden ontwikkeld met een preventieve benadering van risico's en op dusdanige wijze dat ze zich betrouwbaar en zoals bedoeld gedragen, terwijl onbedoelde en onverwachte schade zo veel mogelijk wordt beperkt en onacceptabele schade wordt voorkomen. Dit moet ook gelden voor potentiële veranderingen in de omgeving waarin ze werken of de aanwezigheid van andere actoren (menselijk en kunstmatig) die mogelijk op conflicterende wijze met het systeem in aanraking komen. Daarnaast moet de fysieke en geestelijke integriteit van mensen worden gewaarborgd.
- (67) **Weerbaarheid tegen aanvallen en beveiliging.** Zoals alle softwaresystemen moeten KI-systemen worden beschermd tegen kwetsbaarheden waardoor ze door kwaadwillenden kunnen worden misbruikt (bijv. gehackt). Aanvallen kunnen gericht zijn op de gegevens (datavergiftiging), het model (lek in het model) of de onderliggende infrastructuur, zowel software als hardware. Als een KI-systeem wordt aangevallen, bijvoorbeeld bij vijandige aanvallen, kunnen zowel de gegevens als het gedrag van het systeem worden veranderd, waardoor het systeem andere beslissingen neemt of volledig wordt afgesloten. Systemen en gegevens kunnen ook worden aangetast door kwade opzet of door blootstelling aan onverwachte situaties. Ontoereikende beveiligingsprocessen kunnen ook leiden tot onjuiste beslissingen of zelfs fysieke schade. KI-systemen worden pas veilig geacht³⁷ wanneer er rekening is gehouden met mogelijke onbedoelde toepassingen van KI (bijv. toepassingen voor dubbel gebruik) en potentieel misbruik van een KI-systeem door kwaadwillende actoren en er stappen zijn ondernomen om deze dingen te voorkomen en te beperken.³⁸
- (68) **Uitwijkplan en algemene veiligheid.** KI-systemen moeten over waarborgen beschikken die in geval van problemen een uitwijkplan in werking stellen. Dat kan betekenen dat KI-systemen wisselen van een statistische naar een op regels gebaseerde procedure of dat ze om een menselijke beheerder vragen voor ze hun werk voortzetten.³⁹ Er moet worden gezorgd dat het systeem doet wat het moet doen zonder schade te berokkenen aan levende wezens of het milieu. Daaronder valt ook de minimalisering van onbedoelde gevolgen en fouten. Daarnaast moeten er processen worden ingesteld om potentiële risico's in verband met het gebruik van KI-systemen in verschillende toepassingsdomeinen te verduidelijken en beoordelen. Het benodigde niveau van veiligheidsmaatregelen hangt af van de omvang van het risico dat een KI-systeem met zich meebrengt, die op zijn beurt weer afhangt van de capaciteiten van het systeem. Indien er kan worden voorzien dat het ontwikkelingsproces of het systeem zelf bijzonder hoge risico's zal opleveren, is het cruciaal dat er proactief veiligheidsmaatregelen worden ontwikkeld en getest.
- (69) **Nauwkeurigheid.** Nauwkeurigheid heeft betrekking op het vermogen van een KI-systeem om correcte afwegingen te maken, bijvoorbeeld door informatie in de juiste categorieën in te delen, of op het vermogen ervan om correcte voorspellingen of aanbevelingen te doen of beslissingen te nemen op basis van gegevens of modellen. Door een expliciet en gedegen ontwikkelings- en evaluatieproces kunnen onbedoelde risico's vanwege onjuiste voorspellingen worden beperkt en gecorrigeerd. Indien onnauwkeurige voorspellingen zo nu

³⁷ Zie bijv. de overwegingen onder punt 2.7 van het Gecoördineerd plan inzake kunstmatige intelligentie van de Europese Unie.

³⁸ Voor de beveiliging van KI-systemen kan het hard nodig zijn om binnen onderzoek en ontwikkeling een heilzame cyclus te kunnen ontwikkelen tussen het doorgronden van aanvallen, het ontwikkelen van passende bescherming en de verbetering van evaluatiemethoden. Voor de verwezenlijking daarvan moet convergentie tussen de KI-gemeenschap en de beveiligingsgemeenschap worden gestimuleerd. Daarnaast is het de verantwoordelijkheid van alle betrokken actoren om gemeenschappelijke, grensoverschrijdende veiligheids- en beveiligingsnormen te bepalen en een omgeving van wederzijds vertrouwen te creëren waarin internationale samenwerking wordt bevorderd. Zie voor mogelijke maatregelen *Malicious Use of AI* (Avin S., Brundage M. et. al., 2018).

³⁹ Scenario's waarbij menselijke interventie niet direct mogelijk zou zijn, moeten ook in overweging worden genomen.

en dan niet kunnen worden voorkomen, is het belangrijk dat het systeem kan aangeven hoe groot de kans op dergelijke fouten is. Een hoge mate van nauwkeurigheid is in het bijzonder belangrijk in situaties waarin het KI-systeem rechtstreeks gevolgen heeft voor de levens van mensen.

- (70) **Betrouwbaarheid en reproduceerbaarheid.** Het is essentieel dat de resultaten van KI-systemen reproduceerbaar en betrouwbaar zijn. Een betrouwbaar KI-systeem is een systeem dat goed werkt met allerlei soorten input en in allerlei situaties. Dit is nodig om een KI-systeem te controleren en om onbedoelde schade te voorkomen. Reproduceerbaarheid beschrijft of een KI-experiment hetzelfde gedrag vertoont wanneer het onder gelijke omstandigheden wordt herhaald. Hierdoor kunnen wetenschappers en beleidsmakers nauwkeurig beschrijven wat KI-systemen doen. Replicatiebestanden⁴⁰ kunnen het proces van het testen en reproduceren van gedrag vergemakkelijken.

3. Privacy en datagovernance

- (71) Nauw verbonden met het *beginsel van preventie van schade* is privacy, een grondrecht waarop KI-systemen in het bijzonder van invloed zijn. Voor de preventie van privacyschade is ook geschikte datagovernance noodzakelijk die betrekking heeft op de kwaliteit en integriteit van de gebruikte gegevens, de relevantie ervan binnen het domein waarop de KI-systemen worden geïnstalleerd, de toegangsprotocollen en de capaciteit om gegevens op dusdanige wijze te verwerken dat de privacy wordt beschermd.
- (72) **Privacy en gegevensbescherming.** KI-systemen moeten privacy en gegevensbescherming garanderen gedurende de volledige levenscyclus van het systeem⁴¹. Dit omvat de informatie die oorspronkelijk door de gebruiker is aangeleverd, alsook de informatie die in de loop van zijn/haar interactie met het systeem over de gebruiker is gegenereerd (bijv. resultaten die het KI-systeem heeft gegenereerd voor specifieke gebruikers of de manier waarop gebruikers op bepaalde aanbevelingen hebben gereageerd). Uit de digitale documentatie van menselijk gedrag kunnen KI-systemen niet alleen de voorkeuren van mensen afleiden, maar ook hun seksuele geaardheid, leeftijd, geslacht en religieuze of politieke standpunten. Om te zorgen dat mensen het proces van gegevensverzameling kunnen vertrouwen, moet ervoor worden gezorgd dat de over hen verzamelde gegevens niet worden gebruikt om hen onwettig of onrechtvaardig te discrimineren.
- (73) **Kwaliteit en integriteit van gegevens.** De kwaliteit van de gebruikte gegevenssets is cruciaal voor de prestaties van KI-systemen. Wanneer gegevens worden verzameld, kunnen ze sociaal geconstrueerde vertekening, onnauwkeurigheden, fouten en vergissingen bevatten. Deze moeten worden verholpen vóórdat het systeem met een bepaalde gegevensset wordt getraind. Daarnaast moet de integriteit van de gegevens worden gewaarborgd. Als er kwaadwillige gegevens in een KI-systeem worden ingevoerd, kan het gedrag ervan veranderen, met name bij zelflerende systemen. De gebruikte processen en gegevenssets moeten bij elke stap, zoals het plannen, trainen, testen en installeren, worden getest en gedocumenteerd. Dit moet ook gelden voor KI-systemen die niet intern zijn ontwikkeld, maar elders zijn verkregen.
- (74) **Toegang tot gegevens.** In iedere organisatie die gegevens van personen verwerkt (ongeacht of iemand een gebruiker van het systeem is of niet), moeten gegevensprotocollen worden ingesteld om de toegang tot gegevens te beheren. In deze protocollen moet worden beschreven wie onder welke omstandigheden gegevens kan inzien. Uitsluitend gekwalificeerd personeel met de bevoegdheid en noodzaak om de gegevens van personen in te zien, mag hiertoe de mogelijkheid krijgen.

4. Transparantie

- (75) Deze vereiste houdt nauw verband met het *beginsel van verantwoording* en omvat de transparantie van

⁴⁰ Hierbij gaat het om bestanden die elke stap van het ontwikkelingsproces van het KI-systeem herhalen, van het onderzoek en de initiële gegevensverzameling tot de resultaten.

⁴¹ Hier kan worden verwezen naar bestaande privacywetgeving, zoals de AVG of de komende e-privacyverordening.

elementen die relevant zijn voor een KI-systeem: de gegevens, het systeem en de bedrijfsmodellen.

- (76) **Traceerbaarheid.** De gegevenssets en de processen waaruit de beslissing van het KI-systeem voortkomt, met inbegrip van die van de verzameling en indeling van gegevens, alsook de gebruikte algoritmen, moeten zo goed mogelijk worden gedocumenteerd om ze traceerbaar te maken en de transparantie te vergroten. Dit geldt ook voor de door het KI-systeem gemaakte beslissingen. Hierdoor wordt het mogelijk vast te stellen waarom een KI-beslissing onjuist was, waardoor vervolgens toekomstige fouten kunnen worden voorkomen. Traceerbaarheid maakt dus controleerbaarheid en verklaarbaarheid mogelijk.
- (77) **Verklaarbaarheid.** Verklaarbaarheid heeft te maken met het vermogen om zowel de technische processen van een KI-systeem als de daaraan gerelateerde menselijke beslissingen (bijv. de toepassingsgebieden van een KI-systeem) te verklaren. Voor technische verklaarbaarheid is het nodig dat de door een KI-systeem gemaakte beslissingen door mensen kunnen worden begrepen en getraceerd. Daarnaast moeten er mogelijk afwegingen worden gemaakt tussen het verbeteren van de verklaarbaarheid van een systeem (waardoor de nauwkeurigheid zou kunnen afnemen) of het vergroten van de nauwkeurigheid ervan (ten koste van de verklaarbaarheid). Wanneer een KI-systeem significante gevolgen heeft voor het leven van mensen, moet het altijd mogelijk zijn om te vragen om een geschikte verklaring van het besluitvormingsproces van het KI-systeem. Een dergelijke verklaring moet tijdig worden gegeven en zijn afgestemd op de mate van deskundigheid van de betrokken belanghebbende (bijv. leek, regelgever of onderzoeker). Verder moeten er verklaringen beschikbaar zijn van de mate waarin een KI-systeem het organisatorische besluitvormingsproces, de ontwerpkeuzen van het systeem en de motivering voor de installatie ervan beïnvloedt en vormt (zodat de transparantie van het bedrijfsmodel wordt gewaarborgd).
- (78) **Communicatie.** KI-systemen mogen zich tegenover gebruikers niet als mensen voordoen. Mensen hebben het recht ervan in kennis te worden gesteld dat ze met een KI-systeem te maken hebben. Dat houdt in dat KI-systemen als zodanig herkenbaar moeten zijn. Daarnaast moet, waar dit nodig is om naleving van de grondrechten te waarborgen, de optie worden geboden om te besluiten menselijke interactie aan te gaan in plaats van deze interactie. Bovendien moeten de capaciteiten en beperkingen van het KI-systeem op een voor de betreffende situatie gepaste manier aan beroepsbeoefenaars op het gebied van KI of aan eindgebruikers worden gemeld. Daaronder kan communicatie vallen over de mate van nauwkeurigheid van het systeem, alsook over de beperkingen ervan.

5. Diversiteit, non-discriminatie en rechtvaardigheid

- (79) Om betrouwbare KI te verwezenlijken moeten we inclusie en diversiteit mogelijk maken gedurende de gehele levenscyclus van het KI-systeem. Daarbij moet niet alleen gedurende het hele proces met alle betrokken belanghebbenden rekening worden gehouden, maar moet ook worden gezorgd voor gelijke toegang via inclusieve ontwerpprocessen, alsook voor gelijke behandeling. Deze vereiste is nauw verbonden met *het beginsel van rechtvaardigheid*.
- (80) **Voorkomen van onrechtvaardige vertekening.** In door KI-systemen gebruikte gegevenssets (zowel voor training als voor werkzaamheden) kan sprake zijn van onbedoelde historische vertekening, onvolledigheid of slechte governance modellen. Als dergelijke vertekening wordt behouden, kunnen er onbedoelde (in)directe vooroordelen en discriminatie⁴² tegen bepaalde groepen of mensen ontstaan, waardoor vooroordelen en marginalisering mogelijk worden versterkt. Er kan ook schade ontstaan als gevolg van de onbedoelde uitbuiting van vertekening (door consumenten) of door oneerlijke mededinging, zoals de homogenisering van

⁴² Zie voor een definitie van directe en indirecte discriminatie bijvoorbeeld artikel 2 van Richtlijn 2000/78/EG van de Raad van 27 november 2000 tot instelling van een algemeen kader voor gelijke behandeling in arbeid en beroep. Zie ook artikel 21 van het Handvest van de grondrechten van de EU.

prijzen door middel van samenspanning of een niet-transparante markt.⁴³ Aanwijsbare en discriminerende vertekening moet in de verzamelingsfase waar mogelijk worden verwijderd. Er kan ook sprake zijn van vertekening in de manier waarop KI-systemen worden ontwikkeld (bijv. het programmeren van algoritmen). Deze vertekening kan worden tegengegaan door toezichtsprocessen in te stellen om het doel, de beperkingen, de vereisten en de beslissingen van het systeem op een heldere en transparante manier te analyseren en te behandelen. Daarnaast kan de diversiteit van meningen worden gewaarborgd door personeel met diverse achtergronden en uit verschillende culturen en disciplines in dienst te nemen, dus dit moet worden aangemoedigd.

- (81) ***Toegankelijkheid en universeel ontwerp.*** Met name in domeinen tussen bedrijf en consument moeten bij systemen de gebruikers centraal staan en moeten ze zo worden ontworpen dat mensen KI-producten of -diensten kunnen gebruiken ongeacht hun leeftijd, geslacht, vermogens of eigenschappen. Het is in het bijzonder van belang dat deze technologie toegankelijk is voor mensen met een beperking, die in alle groepen van de samenleving voorkomen. In KI-systemen moet geen gebruik worden gemaakt van één vaste aanpak voor alle situaties, maar er moet rekening worden gehouden met beginselen van universeel ontwerp⁴⁴ die zijn gericht op het breedst mogelijke scala aan gebruikers, volgens relevante toegankelijkheidsnormen.⁴⁵ Daardoor worden voor alle mensen gelijke toegang tot en actieve participatie aan bestaande en nieuwe door de computer ondersteunde menselijke activiteiten en ten aanzien van ondersteunende technologieën mogelijk.⁴⁶
- (82) ***Participatie van belanghebbenden.*** Om betrouwbare KI-systemen te ontwikkelen is het raadzaam om belanghebbenden te raadplegen die direct of indirect met het systeem te maken hebben gedurende de levenscyclus ervan. Het is nuttig om, ook na de installatie, regelmatig om feedback te vragen en langetermijnmechanismen voor de participatie van belanghebbenden in te stellen, bijvoorbeeld door de informatie voor en de raadpleging en participatie van werknemers te garanderen gedurende het volledige toepassingsproces van KI-systemen bij organisaties.

6. Maatschappelijk en milieuwelzijn

- (83) Overeenkomstig de *beginselen van rechtvaardigheid* en *preventie van schade* moeten de samenleving in bredere zin, andere wezens met gevoel en het milieu gedurende de KI-levenscyclus ook als belanghebbenden worden beschouwd. Duurzaamheid en ecologische verantwoordelijkheid van KI-systemen moeten worden aangemoedigd en onderzoek naar KI-oplossingen die ingaan op mondiale punten van zorg, zoals de duurzameontwikkelingsdoelstellingen, moet worden gestimuleerd. In het ideale scenario wordt KI gebruikt ten voordele van alle mensen, inclusief toekomstige generaties.
- (84) ***Duurzame en milieuvriendelijke KI.*** KI-systemen hebben de potentie om een aantal van de meest urgente maatschappelijke zorgen te verhelpen, maar er moet wel worden gezorgd dat dit zo milieuvriendelijk mogelijk gebeurt. Het ontwikkelings-, installatie- en gebruiksproces van het systeem, alsook de volledige toeleveringsketen, moet met betrekking daartoe worden gecontroleerd, bijvoorbeeld door via een kritisch onderzoek naar het gebruik van hulpbronnen en de energieconsumptie tijdens de training te kiezen voor minder schadelijke opties. Maatregelen om de milieuvriendelijkheid van de volledige toeleveringsketen van het KI-systeem te waarborgen, moeten worden aangemoedigd.

⁴³ Cf. paper van het Bureau van de Europese Unie voor de grondrechten:

"BigData: Discrimination in data-supported decision making (2018)" <http://fra.europa.eu/en/publication/2018/big-data-discrimination>.

⁴⁴ Op grond van artikel 42 van de richtlijn inzake het plaatsen van overheidsopdrachten moet bij technische specificaties rekening worden gehouden met toegankelijkheid en ontwerp voor iedereen.

⁴⁵ Bijvoorbeeld EN 301 549.

⁴⁶ Deze vereiste houdt verband met het Verdrag van de Verenigde Naties inzake de rechten van personen met een handicap.

- (85) **Sociale gevolgen.** Als we op alle terreinen van ons leven aan sociale KI-systemen⁴⁷ worden blootgesteld (of het nou gaat om onderwijs, werk, zorg of vermaak), kan onze opvatting van sociale controle veranderen of kan dit gevolgen hebben voor onze sociale relaties en hechting. Hoewel KI-systemen kunnen worden gebruikt om sociale vaardigheden te vergroten,⁴⁸ kunnen ze evenzeer bijdragen aan de verslechtering ervan. Dat zou ook gevolgen kunnen hebben voor het fysieke en geestelijke welzijn van mensen. De effecten van deze systemen moeten daarom zorgvuldig worden gemonitord en afgewogen.
- (86) **Samenleving en democratie.** Naast de beoordeling van de gevolgen van de ontwikkeling, de installatie en het gebruik van een KI-systeem voor individuele personen, moeten deze gevolgen ook vanuit maatschappelijk oogpunt worden beoordeeld, waarbij rekening moet worden gehouden met het effect op instellingen, de democratie en de samenleving als geheel. Het gebruik van KI-systemen moet zorgvuldig worden afgewogen, met name in situaties die verband houden met het democratische proces, met inbegrip van niet alleen politieke besluitvorming, maar ook verkiezingen.

7. Verantwoording

- (87) De vereiste van verantwoording vormt een aanvulling op de bovenstaande vereisten en is nauw verbonden aan het *beginsel van rechtvaardigheid*. Op grond van deze vereiste moeten mechanismen worden ingesteld om de verantwoordelijkheid en verantwoording voor KI-systemen en de resultaten daarvan te garanderen, zowel voor als na de toepassing.
- (88) **Controleerbaarheid.** Controleerbaarheid houdt in dat het mogelijk wordt gemaakt de algoritmen, gegevens en ontwerpprocessen te controleren. Dat betekent niet noodzakelijkerwijs dat informatie over bedrijfsmodellen en intellectuele eigendom in verband met het KI-systeem altijd openbaar beschikbaar moet zijn. Evaluatie door interne en externe controleurs en de beschikbaarheid van dergelijke evaluatieverslagen kunnen bijdragen aan de betrouwbaarheid van de technologie. Bij toepassingen die van invloed zijn op grondrechten, met inbegrip van veiligheidskritieke toepassingen, moeten KI-systemen onafhankelijk kunnen worden gecontroleerd.
- (89) **Minimalisering en verslaglegging van negatieve gevolgen.** Het vermogen om verslag te doen van handelingen of beslissingen die bijdragen aan een bepaald resultaat van het systeem, alsook het vermogen om op de gevolgen van een dergelijk resultaat te reageren, moet worden gewaarborgd. De vaststelling, beoordeling, verslaglegging en minimalisering van de potentiële negatieve effecten van KI-systemen is in het bijzonder cruciaal voor degenen die er (in)direct de gevolgen van ondervinden. Er moet gedegen bescherming beschikbaar zijn voor klokkenluiders, ngo's, vakverenigingen of andere entiteiten wanneer zij melding maken van legitieme zorgen over een op KI gebaseerd systeem. Het gebruik van effectbeoordelingen (bijv. red teaming of vormen van algoritmische effectbeoordeling), zowel voor als tijdens de ontwikkeling, de installatie en het gebruik van KI-systemen, kan nuttig zijn voor het minimaliseren van negatieve gevolgen. Deze beoordelingen moeten in verhouding staan tot het risico dat de KI-systemen met zich meebrengen.
- (90) **Afwegingen.** Bij de uitvoering van de bovenstaande vereisten kunnen er spanningen tussen ontstaan, die tot onvermijdelijke afwegingen kunnen leiden. Dergelijke afwegingen moeten op een rationele en methodologische manier worden benaderd binnen de stand van de technologie. Dat houdt in dat er moet worden vastgesteld welke relevante belangen en waarden door het KI-systeem in het gedrang komen en dat,

⁴⁷ Hierbij gaat het om KI-systemen die met mensen communiceren en contact hebben door sociaal gedrag te simuleren in interactie met een menselijke robot (belichaamde KI) of als avatars in een virtuele werkelijkheid. Daardoor hebben die systemen de potentie om onze sociaal-culturele praktijken en de structuur van ons sociale leven te veranderen.

⁴⁸ Zie bijvoorbeeld het door de EU gefinancierde project waarbij software op basis van KI wordt ontwikkeld om robots in staat te stellen tot effectievere interactie met autistische kinderen tijdens door mensen geleide sessies en de kinderen zo te helpen hun sociale en communicatieve vaardigheden te verbeteren:
http://ec.europa.eu/research/infocentre/article_en.cfm?id=research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968

in geval van een conflict, de afwegingen ertussen uitdrukkelijk moeten worden erkend en geëvalueerd wat betreft het risico voor ethische beginselen, met inbegrip van grondrechten. In situaties waarin geen ethisch acceptabele compromissen kunnen worden gevonden, mogen de ontwikkeling, de installatie en het gebruik van het KI-systeem niet in die vorm worden voortgezet. Alle beslissingen over te maken afwegingen moeten worden onderbouwd en goed worden gedocumenteerd. De beslisser moet aansprakelijk zijn voor de manier waarop de passende afweging wordt gemaakt, en moet de gepastheid van de genomen beslissing voortdurend controleren, zodat waar nodig de noodzakelijke aanpassingen aan het systeem kunnen worden gemaakt.⁴⁹

- (91) **Beroep.** Wanneer zich een negatief effect voordoet, moeten er toegankelijke mechanismen bestaan waarmee geschikte mogelijkheden om beroep in te stellen worden gewaarborgd⁵⁰. De wetenschap dat het mogelijk is om beroep in te stellen wanneer er iets verkeerd gaat, is essentieel voor het vertrouwen. Er moet bijzondere aandacht worden besteed aan kwetsbare personen of groepen.

2. Technische en niet-technische methoden voor de verwezenlijking van betrouwbare KI

- (92) Voor de uitvoering van de bovenstaande vereisten kunnen zowel technische als niet-technische methoden worden gebruikt. Hieronder vallen alle fasen van de levenscyclus van een KI-systeem. De methoden die worden gebruikt om de vereisten uit te voeren, moeten voortdurend worden geëvalueerd. Ook moet er voortdurend verslag worden gedaan van de aanpassingen aan de uitvoeringsprocessen en moeten deze worden gerechtvaardigd.⁵¹ Omdat KI-systemen voortdurend veranderen en binnen een dynamische omgeving werkzaam zijn, is de verwezenlijking van betrouwbare KI een doorlopend proces, dat hieronder in figuur 3 wordt weergegeven.



Figuur 3: Verwezenlijking van betrouwbare KI gedurende de gehele levenscyclus van het systeem

- (93) De volgende methoden kunnen worden beschouwd als aanvulling op of alternatief voor elkaar, aangezien voor verschillende vereisten – en verschillende gevoeligheden – mogelijk verschillende uitvoeringsmethoden nodig zijn. Dit overzicht beoogt niet volledig, uitputtend of verplicht te zijn. Het doel ervan is veeleer om een lijst te geven van voorgestelde methoden die kunnen helpen bij de verwezenlijking van betrouwbare KI.

1. Technische methoden

- (94) In dit gedeelte worden de technische methoden beschreven voor de waarborging van betrouwbare KI die in de

⁴⁹ Dit kan met behulp van verschillende governance modellen worden verwezenlijkt. De aanwezigheid van een interne en/of externe ethische (en sectorspecifieke) deskundige of raad kan bijvoorbeeld nuttig zijn om potentiële conflictgebieden uit te lichten en manieren voor te stellen waarop dat conflict het best kan worden opgelost. Zinnvolle raadpleging van en gesprekken met belanghebbenden, met inbegrip van degenen die het risico lopen negatieve gevolgen te ondervinden van een KI-systeem, is ook nuttig. Europese universiteiten moeten een voortrekkersrol spelen bij de opleiding van de benodigde ethiekdeskundigen.

⁵⁰ Zie ook het advies van het Bureau van de Europese Unie voor de grondrechten: "Improving access to remedy in the area of business and human rights at the EU level" (2017), <https://fra.europa.eu/en/opinion/2017/business-human-rights>.

⁵¹ Hieronder valt bijvoorbeeld rechtvaardiging van de bij het ontwerp, de ontwikkeling en de installatie van het systeem gemaakte keuzen om de bovenstaande vereisten erin te verwerken.

ontwerp-, ontwikkelings- en gebruiksfase van een KI-systeem kunnen worden verwerkt. De onderstaande methoden variëren in mate van volwassenheid⁵².

- *Architecture voor betrouwbare KI*

- (95) Vereisten voor betrouwbare KI moeten worden vertaald naar procedures en/of beperkingen van procedures, die in de architectuur van het KI-systeem moeten worden verankerd. Dit kan worden bereikt door middel van een "witte lijst" met een reeks regels (gedragingen of toestanden) die het systeem altijd moet volgen, een "zwarte lijst" met beperkingen van gedragingen of toestanden die het systeem nooit mag overtreden, en combinaties daarvan of complexere, aantoonbare garanties ten aanzien van het gedrag van het systeem. De monitoring van de naleving van deze beperkingen door het systeem gedurende de werkzaamheden kan via een afzonderlijk proces worden verwezenlijkt.
- (96) KI-systemen met leercapaciteiten die hun gedrag dynamisch kunnen aanpassen, kunnen worden gezien als non-deterministische systemen die mogelijk onverwacht gedrag vertonen. Naar deze systemen wordt vaak gekeken vanuit het theoretische oogpunt van een cyclus van "waarnemen-plannen-handelen". Om deze architectuur aan te passen voor de verwezenlijking van betrouwbare KI moeten de vereisten worden geïntegreerd bij alle drie de stappen van de cyclus: i) bij de stap "waarnemen" moet het systeem zo worden ontwikkeld dat het alle omgevingselementen herkent die nodig zijn om naleving van de vereisten te waarborgen; ii) bij de stap "plannen" moet het systeem uitsluitend plannen overwegen die aansluiten bij de vereisten; iii) bij de stap "handelen" moeten de handelingen van het systeem beperkt blijven tot gedrag waarmee de vereisten worden verwezenlijkt.
- (97) De architectuur zoals hierboven geschetst, is algemeen en biedt voor de meeste KI-systemen slechts een imperfecte beschrijving. Toch biedt zij aanknopingspunten voor beperkingen en beleidspunten die in specifieke modules moeten worden weergegeven, zodat er een geheel systeem ontstaat dat betrouwbaar is en ook zo wordt gezien.

- *Ethiek en rechtsstaat door ontwerp (X door ontwerp)*

- (98) Methoden om waarden door ontwerp te waarborgen verbinden de abstracte beginselen waaraan het systeem moet voldoen en de specifieke uitvoeringsbeslissingen nauwkeurig en expliciet. Het idee dat naleving van normen kan worden opgenomen in het ontwerp van het KI-systeem is voor deze methode essentieel. Het is de verantwoordelijkheid van bedrijven om de gevolgen van hun KI-systemen en de normen waaraan hun KI-systeem moet voldoen om negatieve gevolgen te voorkomen, vanaf het allereerste begin vast te stellen. Verschillende "door-ontwerp"-concepten worden al veelvuldig gebruikt, zoals *privacy door ontwerp* en *beveiliging door ontwerp*. Zoals hierboven aangegeven, moeten de processen, gegevens en resultaten van KI veilig zijn, wil het systeem vertrouwen winnen. Ook moet het systeem robuust worden ontworpen om kwaadwillige gegevens en aanvallen te weerstaan. Er moet een mechanisme voor veiligheidsuitschakeling worden opgenomen, en het systeem moet zijn werkzaamheden kunnen vervolgen na gedwongen uitschakeling (bijv. een aanval).

- *Verklaringsmethoden*

- (99) Een systeem kan pas betrouwbaar zijn als we kunnen begrijpen waarom het zich op een bepaalde manier heeft gedragen en waarom het een bepaalde interpretatie heeft gegeven. Er is een heel onderzoeksgebied, verklaarbare KI (VKI), gewijd aan deze kwestie om de onderliggende mechanismen van het systeem beter te begrijpen en oplossingen te vinden. Momenteel is dit nog een open uitdaging voor KI-systemen op basis van neurale netwerken. Trainingsprocessen met neurale netwerken kunnen leiden tot netwerkparameters die zijn

⁵² Hoewel sommige van deze methoden momenteel al beschikbaar zijn, is voor andere nog verder onderzoek nodig. De gebieden waarop verder onderzoek nodig is, worden ook gebruikt als input voor het tweede product van de AI HLEG: de aanbevelingen voor beleid en investeringen.

ingesteld op numerieke waarden die lastig te verenigen zijn met resultaten. Bovendien kunnen kleine veranderingen in de gegevenswaarden soms leiden tot drastische veranderingen in de interpretatie, waardoor het systeem bijvoorbeeld een schoolbus met een struisvogel verwacht. Van deze kwetsbaarheid kan ook gebruik worden gemaakt tijdens aanvallen op het systeem. Methoden met VKI-onderzoek zijn cruciaal, niet alleen om het gedrag van het systeem te verklaren tegenover gebruikers, maar ook om betrouwbare technologie te installeren.

- *Testen en valideren*

(100) Vanwege het non-deterministische en contextgebonden karakter van KI-systemen is traditioneel testen niet voldoende. Fouten van de door het systeem gebruikte concepten en representaties komen mogelijk alleen voor wanneer een programma op gegevens wordt toegepast die realistisch genoeg zijn. Om de verwerking van gegevens te controleren en valideren moet het onderliggende model daarom tijdens zowel de training als de installatie zorgvuldig worden gemonitord op stabiliteit, robuustheid en werking binnen heldere en voorspelbare grenzen. Er moet worden gezorgd dat het resultaat van het planningsproces consistent is met de input en dat de beslissingen op dusdanige wijze worden genomen dat het onderliggende proces kan worden gevalideerd.

(101) Het testen en valideren van het systeem moet zo vroeg mogelijk gebeuren om te zorgen dat het systeem zich gedurende de volledige levenscyclus en met name na de installatie gedraagt zoals bedoeld. Alle componenten van een KI-systeem moeten erin worden opgenomen, met inbegrip van gegevens, vooraf getrainde modellen, omgevingen en het gedrag van het systeem als geheel. Het systeem moet worden ontworpen en uitgevoerd door een zo divers mogelijke groep mensen. Er moeten meerdere maatstaven worden ontwikkeld voor de categorieën die vanuit verschillende oogpunten worden getest. Kwaadwillig testen door vertrouwde en diverse "red teams" die opzettelijk proberen het systeem te "breken" om kwetsbaarheden te vinden, alsook "bug bounties" die buitenstaanders stimuleren om fouten en zwakheden in het systeem op te sporen en op verantwoorde wijze te melden, kan worden overwogen. Tot slot moet er worden gezorgd dat de resultaten of handelingen consistent zijn met de resultaten van de voorgaande processen door ze te vergelijken met het eerder vastgestelde beleid, zodat dit niet wordt geschonden.

- *Kwaliteit van de dienstindicatoren*

(102) Er kan worden beschreven wat de geschikte kwaliteit van de dienstindicatoren is, om te zorgen dat men in de basis begrijpt of ze zijn getest en ontwikkeld met beveiligings- en veiligheidsoverwegingen in het achterhoofd. Bij deze indicatoren kunnen maatregelen worden opgenomen om het testen en trainen van algoritmen te evalueren, alsook maatstaven van traditionele software voor: functionaliteit; prestaties; bruikbaarheid; betrouwbaarheid; veiligheid; en onderhoudbaarheid.

2. Niet-technische methoden

(103) In dit gedeelte wordt een scala aan niet-technische methoden beschreven die een waardevolle rol kunnen spelen bij het waarborgen en in stand houden van betrouwbare KI. Ook deze methoden moeten **voortdurend** worden geëvalueerd.

- *Regelgeving*

(104) Zoals hierboven genoemd bestaat er momenteel al regelgeving ter ondersteuning van betrouwbare KI – denk bijvoorbeeld aan wetgeving inzake productveiligheid en kaders voor aansprakelijkheid. Voor zover er in onze optiek regelgeving moet worden herzien, aangepast of ingevoerd – om betrouwbare KI te waarborgen of mogelijk te maken – geven we dit aan in ons tweede product, dat bestaat uit aanbevelingen voor KI-beleid en -investeringen.

- *Gedragscodes*

(105) Organisaties en belanghebbenden kunnen zich aansluiten bij deze richtsnoeren door hun handvest van maatschappelijke verantwoordelijkheid, hun kernprestatie-indicatoren (KPI's), hun gedragscodes of hun internationale beleidsdocumenten aan te passen door toe te voegen dat ze streven naar betrouwbare KI. Meer in het algemeen kunnen organisaties die aan een KI-systeem werken, hun bedoelingen documenteren en deze onderschrijven met normen voor bepaalde wenselijke waarden, zoals grondrechten, transparantie en het voorkomen van schade.

- *Normalisatie*

(106) Normen, bijvoorbeeld voor ontwerp-, productie- en bedrijfspraktijken, kunnen voor KI-gebruikers, consumenten, organisaties, onderzoeksinstellingen en overheden als een kwaliteitsbeheersysteem fungeren door de mogelijkheid te bieden tot het herkennen en stimuleren van ethisch gedrag via hun aankoopbeslissingen. Naast traditionele normen bestaan er ook medereguleringsbenaderingen: accreditatiesystemen, professionele ethische codes of normen voor ontwerp conform de grondrechten. Huidige voorbeelden zijn ISO-normen of de normen uit de IEEE P7000-reeks, maar in de toekomst zou een label voor "betrouwbare KI" nuttig kunnen zijn om, aan de hand van specifieke technische normen, te bevestigen dat het systeem bijvoorbeeld voldoet aan de normen voor veiligheid, technische robuustheid en verklaarbaarheid.

- *Certificering*

(107) Omdat niet kan worden verwacht dat iedereen de werking en gevolgen van KI-systemen volledig kan begrijpen, kunnen er organisaties in overweging worden genomen die aan het grotere publiek kunnen bevestigen dat een KI-systeem transparant, verantwoordelijk en rechtvaardig is⁵³. Voor deze certificeringen zouden normen worden gebruikt die voor verschillende toepassingsdomeinen en KI-technieken zijn ontwikkeld en goed zijn afgestemd op de sectorale en maatschappelijke normen van de betreffende context. Certificering kan echter nooit in de plaats komen van verantwoordelijkheid. Daarom moet deze worden aangevuld met verantwoordingskaders, met inbegrip van disclaimers en mechanismen voor beoordeling en sanering⁵⁴.

- *Verantwoording via governancekaders*

(108) Organisaties moeten zowel interne als externe governancekaders creëren om verantwoording te waarborgen voor de ethische dimensies van beslissingen in verband met de ontwikkeling, de installatie en het gebruik van KI. Daarbij kan het bijvoorbeeld gaan om de aanstelling van iemand die gaat over ethische kwesties in verband met KI, of van een interne/externe ethische commissie of raad. Een van de mogelijke taken van een dergelijke persoon, commissie of raad bestaat uit het houden van toezicht en het geven van advies. Zoals hierboven beschreven kunnen certificeringsspecificaties hier ook een rol in spelen. Er moet worden gezorgd voor communicatiekanalen met de bedrijfstak en/of groepen voor publiek toezicht, zodat beste praktijken kunnen worden uitgewisseld, dilemma's kunnen worden besproken of opkomende ethisch zorgelijke kwesties kunnen worden gemeld. Dergelijke mechanismen kunnen dienen ter aanvulling op, maar niet ter vervanging van wettelijk toezicht (bijv. in de vorm van de aanstelling van een functionaris voor gegevensbescherming of vergelijkbare maatregelen die wettelijk verplicht zijn op grond van het gegevensbeschermingsrecht).

- *Onderwijs en bewustzijn ter bevordering van een ethische denkwijze*

(109) Voor betrouwbare KI is het belangrijk dat alle belanghebbenden deelnemen en goed geïnformeerd zijn. Communicatie, onderwijs en opleiding zijn daarbij van groot belang om te zorgen dat kennis van de potentiële gevolgen van KI-systemen wijdverbreid is en om mensen bewust te maken van het feit dat ze de

⁵³ Zoals voorgesteld door bijvoorbeeld het initiatief Ethically Aligned Design (Ethisch afgestemd ontwerp) van het IEEE: <https://standards.ieee.org/industry-connections/ec/autonomous-systems.html>.

⁵⁴ Zie voor meer informatie over de beperkingen van certificering: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

maatschappelijke ontwikkeling mede vorm kunnen geven. Het gaat daarbij om alle belanghebbenden, zoals degenen die betrokken zijn bij het maken van de producten (de ontwerpers en ontwikkelaars), de gebruikers (bedrijven of individuele personen) en andere beïnvloede groepen (die misschien geen KI-systeem aanschaffen of gebruiken, maar voor wie beslissingen worden gemaakt door een KI-systeem, alsook de samenleving als geheel). Basiskennis op het gebied van KI moet in de gehele samenleving worden bevorderd. Voor het onderwijzen van het grote publiek is het noodzakelijk dat ethici van gedegen vaardigheden en opleiding worden voorzien op dit gebied.

- *Participatie van belanghebbenden en sociale dialoog*

(110) KI heeft veel voordelen en Europa moet zorgen dat deze voor iedereen beschikbaar zijn. Daarvoor is een open debat en de betrokkenheid van sociale partners, belanghebbenden en het grote publiek nodig. Veel organisaties maken al gebruik van panels van belanghebbenden voor de bespreking van het gebruik van KI-systemen en gegevensanalyse. In deze panels zetelen verschillende soorten leden, zoals juridische deskundigen, technische deskundigen, ethici, vertegenwoordigers van consumenten en werknemers. Door participatie en de dialoog op het gebied van het gebruik en de gevolgen van KI-systemen te stimuleren wordt de evaluatie van resultaten en methoden ondersteund. Dit kan in het bijzonder nuttig zijn in complexe gevallen.

- *Diversiteit en inclusieve ontwerpteams*

(111) Diversiteit en inclusie spelen een cruciale rol bij het ontwikkelen van KI-systemen voor gebruik in de echte wereld. Naarmate KI-systemen meer taken zelfstandig uitvoeren, is het essentieel dat de teams die deze systemen ontwerpen, ontwikkelen, testen, onderhouden, installeren en/of aanschaffen, dezelfde diversiteit vertonen als de gebruikers en de samenleving in het algemeen. Op die manier wordt bijgedragen aan de objectiviteit en wordt er rekening gehouden met verschillende perspectieven, behoeften en doelen. In het ideale scenario zijn de teams niet alleen divers wat geslacht, cultuur en leeftijd betreft, maar ook als het gaat om professionele achtergrond en vaardigheden.

Essentiële richtsnoeren uit hoofdstuk II:

- ✓ Zorg dat het KI-systeem gedurende de gehele levenscyclus voldoet aan de vereisten voor betrouwbare KI: 1) menselijke controle en menselijk toezicht, 2) technische robuustheid en veiligheid, 3) privacy en datagovernance, 4) transparantie, 5) diversiteit, non-discriminatie en rechtvaardigheid, 6) milieu- en maatschappelijk welzijn en 7) verantwoordingsplicht.
- ✓ Bestudeer technische en niet-technische methoden om te waarborgen dat die vereisten worden verwezenlijkt.
- ✓ Stimuleer onderzoek en innovatie om aan het controleren van KI-systemen bij te dragen en het verwezenlijken van de vereisten te bevorderen. Verspreid de resultaten en open vragen onder een breder publiek en leid systematisch een nieuwe generatie deskundigen op het gebied van KI-ethiek op.
- ✓ Verspreid informatie over de mogelijkheden en beperkingen van een KI-systeem en over de manier waarop de vereisten worden verwezenlijkt, op een heldere en proactieve manier onder belanghebbenden om onrealistische verwachtingen te voorkomen. Wees open over het feit dat ze met een KI-systeem te maken hebben.
- ✓ Faciliteer de traceerbaarheid en controleerbaarheid van KI-systemen, met name in kritieke omgevingen en situaties.
- ✓ Betrek belanghebbenden bij de volledige levenscyclus van het KI-systeem. Stimuleer onderwijs en opleiding, zodat alle belanghebbenden zich bewust zijn van betrouwbare KI en opgeleid zijn op dat gebied.
- ✓ Wees bedacht op conflicterende beginselen en vereisten. Zoek, evalueer en documenteer deze afwegingen en de oplossingen voortdurend en maak ze kenbaar.

III. Hoofdstuk III: Betrouwbare KI controleren

- (112) Op basis van de kernvereisten uit hoofdstuk II wordt in dit hoofdstuk een niet-uitputtende **controlelijst voor betrouwbare KI** (testversie) gegeven voor de **operationalisering van betrouwbare KI**. Deze lijst is in het bijzonder van toepassing op KI-systemen met rechtstreekse interactie met gebruikers en is met name bedoeld voor ontwikkelaars en installateurs van KI-systemen (ongeacht of zij het systeem zelf hebben ontwikkeld of het van derden hebben verkregen). In deze controlelijst wordt niet ingegaan op de operationalisering van de eerste component van betrouwbare KI (wettige KI). Het volgen van deze controlelijst vormt geen bewijs dat de wetgeving wordt nageleefd. De lijst is ook niet bedoeld als richtsnoer voor de naleving van de toepasselijke wetgeving. Omdat KI-systemen sterk toepassingsgebonden zijn, moet de controlelijst worden afgestemd op de specifieke gebruikssituatie en -context waarbinnen een systeem werkzaam is. Verder wordt in dit hoofdstuk een algemene aanbeveling gedaan over de manier waarop de controlelijst voor betrouwbare KI kan worden toegepast via een governancestructuur waarin zowel het operationele als het bestuurlijke niveau is opgenomen.
- (113) De controlelijst en governancestructuur worden ontwikkeld in nauwe samenwerking met belanghebbenden uit de publieke en particuliere sector. Het proces wordt uitgevoerd als een testproces, zodat er uitgebreide feedback mogelijk is vanuit twee parallelle processen:
- a. een kwalitatief proces dat de representativiteit waarborgt, waarbij een klein aantal geselecteerde bedrijven, organisaties en instellingen (uit verschillende sectoren en van verschillende grootte) zich aanmeldt om de controlelijst en de governancestructuur in de praktijk te testen en diepgaande feedback te geven;
 - b. een kwantitatief proces, waarbij alle geïnteresseerde belanghebbenden zich kunnen aanmelden om de controlelijst te testen en feedback te geven via een openbare raadpleging.
- (114) Na de testfase integreren wij de resultaten van het feedbackproces in de controlelijst en stellen we begin 2020 een herziene versie op. Het doel is om een kader te creëren dat horizontaal kan worden gebruikt voor alle toepassingen en dat dus een grondslag biedt voor de verwezenlijking van betrouwbare KI in alle domeinen. Wanneer een dergelijk kader eenmaal is vastgesteld, zou er een sectoraal of toepassings specifiek kader kunnen worden ontwikkeld.

Governance

- (115) Het kan nuttig zijn voor bedrijven, organisaties en instellingen om na te denken over de manier waarop de controlelijst voor betrouwbare KI binnen hun organisatie kan worden toegepast. Dit kan door het controleproces in de bestaande governancemechanismen op te nemen of door nieuwe processen in te voeren. Die keuze hangt af van de interne structuur, de omvang en de beschikbare middelen van de organisatie.
- (116) Uit onderzoek⁵⁵ is gebleken dat aandacht van het hoogste management essentieel is om verandering teweeg te brengen. Ook is aangetoond dat de acceptatie en relevantie van het invoeren van nieuwe processen (al dan niet technologisch) kunnen worden bevorderd door alle belanghebbenden binnen een bedrijf, organisatie of instelling erbij te betrekken⁵⁶. Daarom raden wij aan om een proces in te voeren waarbij zowel het operationele niveau als het hoogste bestuurlijke niveau betrokken zijn.

⁵⁵ <https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation>

⁵⁶ Zie bijvoorbeeld A. Bryson, E. Barth en H. Dale-Olsen, The Effects of Organisational change on worker well-being and the moderating role of trade unions, *ILRRReview*, 66(4), juli 2013; Jirjahn, U. en Smith, S.C. (2006). "What Factors Lead Management to Support or Oppose Employee Participation—With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations", 45(4), blz. 650–680; Michie, J. en Sheehan, M. (2003). "Labour market deregulation, 'flexibility' and innovation", *Cambridge Journal of Economics*, 27(1), blz. 123–143.

Niveau	Relevante taken (afhankelijk van de organisatie)
Management en raad	Het hoogste management bespreekt en evalueert de ontwikkeling, installatie of aankoop van KI en dient als escalatieraad voor de evaluatie van alle KI-innovaties en -toepassingen wanneer er punten van zorg naar voren komen. Het betreft degenen op wie de eventuele invoering van KI-systemen van invloed zou zijn (bijv. werknemers), alsook hun vertegenwoordigers, bij het hele proces via informatie-, raadplegings- en participatieprocedures.
Afdeling naleving/Juridische afdeling/Afdeling maatschappelijke verantwoordelijkheid	De afdeling verantwoordelijkheid monitort het gebruik van de controlelijst en de noodzakelijke ontwikkeling ervan voor aanpassing aan de veranderingen in technologie of regelgeving. De afdeling werkt de normen of het interne beleid inzake KI-systemen bij en zorgt ervoor dat bij het gebruik van dergelijke systemen het huidige wettelijke en regelgevingskader en de waarden van de organisatie worden gevolgd.
Product- en dienstenontwikkeling of gelijkwaardig	De afdeling product- en dienstenontwikkeling gebruikt de controlelijst om op KI gebaseerde producten en diensten te evalueren en houdt alle resultaten bij. Deze resultaten worden op managementniveau besproken. Daar worden de nieuwe of herziene op KI gebaseerde toepassingen uiteindelijk ook goedgekeurd.
Kwaliteitsborging	De afdeling kwaliteitsborging (of gelijkwaardig) waarborgt en controleert de resultaten van de controlelijst en neemt maatregelen om een kwestie naar een hoger niveau te tillen indien het resultaat niet bevredigend is of er onvoorziene resultaten worden ontdekt.
HR	De HR-afdeling zorgt voor de juiste combinatie van vaardigheden en diversiteit binnen de profielen van de ontwikkelaars van KI-systemen. Ook zorgt de afdeling ervoor dat het juiste niveau van opleiding op het gebied van betrouwbare KI wordt verzorgd binnen de organisatie.
Inkoop	De afdeling inkoop zorgt ervoor dat in het inkoopproces voor op KI gebaseerde producten en diensten een controle op betrouwbare KI is opgenomen.
Dagelijkse werkzaamheden	Ontwikkelaars en projectmanagers nemen de controlelijst op in hun dagelijkse werkzaamheden en documenteren de resultaten en uitkomsten van de controle.

De controlelijst voor betrouwbare KI gebruiken

- (117) Bij het gebruik van de controlelijst in de praktijk raden wij aan niet alleen aandacht te besteden aan de punten van zorg, maar ook aan de vragen die niet (gemakkelijk) kunnen worden beantwoord. Eén potentieel probleem is het gebrek aan diversiteit wat betreft de vaardigheden en competenties binnen het team dat het KI-systeem ontwikkelt en test. Daarom kan het nodig zijn andere belanghebbenden van binnen of buiten de organisatie erbij te betrekken. Het is sterk aan te raden om alle resultaten bij te houden in zowel technische als bestuurlijke termen, zodat de aanpak van het probleem op alle niveaus van de governancestructuur kan worden begrepen.
- (118) De controlelijst is bedoeld als leidraad voor beroepsbeoefenaars op het gebied van KI bij het ontwikkelen, installeren en gebruiken van betrouwbare KI. De controle moet op evenredige wijze worden afgestemd op de specifieke gebruikssituatie. Gedurende de testfase kunnen er specifieke gevoelige gebieden naar voren komen. Bij de volgende stap wordt de noodzaak van verdere specificatie in dergelijke gevallen

geëvalueerd. Deze controlelijst biedt geen concrete antwoorden op de ontstane vragen, maar stimuleert wel reflectie op de stappen die kunnen bijdragen aan de verwezenlijking van de betrouwbaarheid van KI-systemen, en op de potentiële stappen die met betrekking daartoe moeten worden genomen.

Verband met bestaande wetgeving en processen

- (119) Het is ook belangrijk dat degenen die betrokken zijn bij de ontwikkeling, de installatie en het gebruik van KI, erkennen dat er verschillende bestaande wetten zijn op grond waarvan bepaalde processen verplicht zijn en bepaalde resultaten verboden, en die (gedeeltelijk) samen kunnen vallen met enkele van de maatregelen op de controlelijst. De wetgeving inzake gegevensbescherming bevat bijvoorbeeld een reeks wettelijke voorschriften die moeten worden nageleefd door degenen die betrokken zijn bij de verzameling en verwerking van persoonsgegevens. Voor betrouwbare KI is het echter ook nodig dat gegevens ethisch worden verwerkt. Met interne procedures en intern beleid gericht op het waarborgen van de naleving van wetgeving inzake gegevensbescherming kan dit mogelijk worden gemaakt en kunnen zo bestaande wettelijke processen worden aangevuld. Het volgen van deze controlelijst vormt echter *geen* bewijs dat de wetgeving wordt nageleefd. De lijst is ook niet bedoeld als richtsnoer voor de naleving van toepasselijke wetgeving. Het doel van de controlelijst is veeleer het aanleveren van een reeks specifieke vragen voor ontvangers die proberen te zorgen dat hun benadering van de ontwikkeling en installatie van KI gericht is op (de verwezenlijking van) betrouwbare KI.
- (120) Veel beroepsbeoefenaars op het gebied van KI beschikken ook al over bestaande controlehulpmiddelen en softwareontwikkelingsprocessen om naleving, ook van niet-wettelijke normen, te waarborgen. De onderstaande controle hoeft niet noodzakelijkerwijs los te worden uitgevoerd, maar kan in dergelijke bestaande praktijken worden verwerkt.

CONTROLELIJST VOOR BETROUWBARE KI (TESTVERSIE)

1. Menselijke controle en menselijk toezicht

Grondrechten:

- ✓ Hebt u, in de gebruikssituaties waarin er mogelijk sprake kon zijn van negatieve gevolgen voor de grondrechten, een effectbeoordeling op het gebied van grondrechten uitgevoerd? Hebt u mogelijke afwegingen tussen de verschillende beginselen en rechten vastgesteld en gedocumenteerd?
- ✓ Is er sprake van een wisselwerking tussen het KI-systeem en de beslissingen van menselijke eindgebruikers (bijv. aanbevolen handelingen of beslissingen, aanbod van opties)?
 - Is er in die gevallen een risico dat het KI-systeem de menselijke autonomie aantast door het beslissingsproces van de eindgebruiker op een onbedoelde manier te beïnvloeden?
 - Hebt u overwogen of het KI-systeem aan de gebruikers moet aangeven dat inhoud of een beslissing, advies of uitkomst het resultaat is van een algoritmische beslissing?
 - Indien het KI-systeem een chatbot of gesprekssysteem bevat: zijn de menselijke eindgebruikers in kennis gesteld van het feit dat ze te maken hebben met een niet-menselijke actor?

Menselijke controle:

- ✓ Indien het KI-systeem in de werk- en arbeidsprocessen is opgenomen: hebt u nagedacht over de taakverdeling tussen het KI-systeem en de menselijke werknemers om te zorgen voor zinvolle

interactie en een gepaste mate van menselijk toezicht en menselijke controle?

- Versterkt of vergroot het KI-systeem menselijke capaciteiten?
- Hebt u voorzorgsmaatregelen genomen om overmatig vertrouwen in of op het KI-systeem bij werkprocessen te voorkomen?

Menselijk toezicht:

- ✓ Hebt u nagedacht over de gepaste mate van menselijke controle voor dit specifieke KI-systeem en deze specifieke gebruikssituatie?
 - Kunt u, indien van toepassing, de mate van menselijke controle of betrokkenheid beschrijven? Wie is de "human in control" en wat zijn de momenten of hulpmiddelen voor menselijke interventie?
 - Hebt u mechanismen en maatregelen ingesteld om dergelijke potentiële menselijke controle of menselijk toezicht te waarborgen of om te zorgen dat beslissingen onder de algehele verantwoordelijkheid van mensen worden genomen?
 - Hebt u maatregelen genomen om controle mogelijk te maken en problemen in verband met het beheren van KI-autonomie te verhelpen?
- ✓ Indien het KI-systeem of de gebruikssituatie zelflerend of autonoom is: hebt u specifiekere controle- en toezichtsmechanismen ingesteld?
 - Wat voor detectie- en responsmechanismen hebt u ingesteld om te controleren of er iets mis kan gaan?
 - Hebt u gezorgd voor een "stopknop" of procedure waarmee een activiteit indien nodig veilig kan worden afgebroken? Wordt met deze procedure het proces volledig afgebroken, wordt het gedeeltelijk afgebroken of wordt de controle overgedragen aan een mens?

2. Technische robuustheid en veiligheid

Weerbaarheid tegen aanvallen en beveiliging:

- ✓ Hebt u potentiële soorten aanvallen onderzocht waarvoor het KI-systeem kwetsbaar zou kunnen zijn?
 - Hebt u in het bijzonder nagedacht over verschillende soorten kwetsbaarheid, zoals gegevensverontreiniging, fysieke infrastructuur en cyberaanvallen?
- ✓ Hebt u maatregelen of systemen ingesteld om de integriteit van het KI-systeem en de weerbaarheid tegen potentiële aanvallen te waarborgen?
- ✓ Hebt u onderzocht hoe uw systeem zich in onverwachte situaties en omgevingen gedraagt?
- ✓ Hebt u overwogen of en in hoeverre uw systeem geschikt zou kunnen zijn voor dubbel gebruik? Zo ja, hebt u geschikte preventieve maatregelen genomen tegen deze situatie (met inbegrip van bijvoorbeeld het niet publiceren van het onderzoek of het niet installeren van het systeem)?

Uitwijkplan en algemene veiligheid:

- ✓ Hebt u gezorgd dat uw systeem over een toereikend uitwijkplan beschikt, mocht het te maken krijgen met kwaadwillige aanvallen of andere onverwachte situaties (bijv. technische omschakelingsprocedures of vragen om een menselijke beheerder alvorens verder te gaan)?
- ✓ Hebt u nagedacht over de omvang van het risico dat het KI-systeem in deze specifieke gebruikssituatie met zich meebrengt?
 - Hebt u een proces ingesteld om de risico's en de veiligheid te meten en te controleren?
 - Hebt u de noodzakelijke informatie aangeleverd in geval van een risico voor de menselijke fysieke integriteit?
 - Hebt u een verzekering overwogen voor eventuele schade door het KI-systeem?
 - Hebt u de potentiële veiligheidsrisico's vastgesteld van (andere) voorzienbare toepassingen van de technologie, met inbegrip van onbedoeld of kwaadwillig misbruik ervan? Is er een plan om deze risico's te beperken of te beheeren?
- ✓ Hebt u onderzocht of er een goede kans is dat het KI-systeem schade veroorzaakt voor gebruikers of derden? Zo ja, hebt u de waarschijnlijkheid, de potentiële schade, het getroffen publiek en de ernst onderzocht?
 - Indien er een risico bestaat dat het KI-systeem schade veroorzaakt: hebt u nagedacht over regels voor aansprakelijkheid en bescherming van consumenten? Hoe hebt u daar rekening mee gehouden?
 - Hebt u nagedacht over de potentiële gevolgen of het veiligheidsrisico voor het milieu of voor dieren?
 - Hebt u in uw risicoanalyse meegenomen of beveiligings- of netwerkproblemen (zoals gevaren voor de cyberbeveiliging) veiligheidsrisico's of schade met zich meebrengen vanwege onbedoeld gedrag van het KI-systeem?
- ✓ Hebt u de vermoedelijke gevolgen ingeschat van een fout van uw KI-systeem met als gevolg de levering van verkeerde resultaten waardoor uw systeem niet meer beschikbaar is of maatschappelijk onaanvaardbare resultaten oplevert (bijv. discriminerende praktijken)?
 - Hebt u drempels en governance vastgesteld voor de bovenstaande scenario's, zodat er alternatieve of uitwijkplannen in werking worden gesteld?
 - Hebt u uitwijkplannen vastgesteld en getest?

Nauwkeurigheid

- ✓ Hebt u onderzocht welke mate en definitie van nauwkeurigheid nodig zouden zijn in de context van het KI-systeem en in de gebruikssituatie?
 - Hebt u onderzocht hoe de nauwkeurigheid wordt gemeten en gegarandeerd?
 - Hebt u maatregelen ingesteld om te zorgen dat de gebruikte gegevens volledig en actueel zijn?
 - Hebt u maatregelen ingesteld om te controleren of er aanvullende gegevens nodig zijn,

bijvoorbeeld om de nauwkeurigheid te vergroten of vertekening weg te nemen?

- ✓ Hebt u de schade onderzocht die zou worden veroorzaakt als het KI-systeem onnauwkeurige voorspellingen doet?
- ✓ Hebt u voor manieren gezorgd om te meten of uw systeem een onaanvaardbaar aantal onnauwkeurige voorspellingen doet?
- ✓ Indien er onnauwkeurige voorspellingen worden gedaan: hebt u een reeks stappen vastgesteld om het probleem te verhelpen?

Betrouwbaarheid en reproduceerbaarheid:

- ✓ Hebt u een strategie vastgesteld om te monitoren en testen of het KI-systeem in overeenstemming is met de doelen, doelstellingen en beoogde toepassingen?
 - Hebt u getest of er rekening moet worden gehouden met specifieke situaties of omstandigheden om de reproduceerbaarheid te waarborgen?
 - Hebt u controleprocessen of -methoden vastgesteld om verschillende aspecten van betrouwbaarheid en reproduceerbaarheid te meten en te waarborgen?
 - Hebt u processen vastgesteld om te beschrijven wanneer een KI-systeem bij bepaalde soorten instellingen een fout maakt?
 - Hebt u deze processen voor het testen en controleren van de betrouwbaarheid van KI-systemen duidelijk gedocumenteerd en geoperationaliseerd?

Hebt u gezorgd voor mechanismen of communicatie om (eind)gebruikers te garanderen dat het KI-systeem betrouwbaar is?

3. Privacy en datagovernance

Respect voor privacy en gegevensbescherming:

- ✓ Hebt u, afhankelijk van de gebruikssituatie, mechanismen vastgesteld waardoor anderen problemen in verband met privacy of gegevensbescherming kunnen melden die betrekking hebben op de processen van het KI-systeem in verband met gegevensverzameling (voor zowel training als werking) en gegevensverwerking?
- ✓ Hebt u het type en de reikwijdte van de gegevens in uw gegevenssets onderzocht (bijv. of ze persoonsgegevens bevatten)?
- ✓ Hebt u nagedacht over manieren om het KI-systeem te ontwikkelen of het model te trainen zonder of met zo weinig mogelijk gebruik van potentieel gevoelige gegevens of persoonsgegevens?
- ✓ Hebt u, afhankelijk van de gebruikssituatie, mechanismen ingebouwd voor het opmerken van en de controle over persoonsgegevens (zoals geldige toestemming en de mogelijkheid om deze in te trekken, indien van toepassing)?
- ✓ Hebt u maatregelen genomen om de privacy te vergroten, bijvoorbeeld via encryptie, anonimisering en aggregatie?

- ✓ Indien er een functionaris voor gegevensbescherming is: hebt u deze persoon in een vroeg stadium bij het proces betrokken?

Kwaliteit en integriteit van gegevens:

- ✓ Hebt u uw systeem afgestemd op de potentieel relevante normen (bijv. ISO, IEEE) of veelgebruikte protocollen voor uw dagelijkse gegevensbeheer en governance?
- ✓ Hebt u toezichtsmechanismen vastgesteld voor de verzameling, de opslag, de verwerking en het gebruik van gegevens?
- ✓ Hebt u onderzocht in hoeverre u controle heeft over de kwaliteit van de gebruikte externe gegevensbronnen?
- ✓ Hebt u processen vastgesteld om de kwaliteit en integriteit van uw gegevens te waarborgen? Hebt u andere processen overwogen? Hoe controleert u of uw gegevenssets niet zijn aangetast of gehackt?

Toegang tot gegevens:

- ✓ Welke protocollen, processen en procedures zijn er gevolgd voor het beheer en de waarborging van goede datagovernance?
 - Hebt u onderzocht wie er toegang heeft tot gebruikersgegevens en onder welke omstandigheden?
 - Hebt u gecontroleerd of deze personen gekwalificeerd zijn om de gegevens in te zien, of het noodzakelijk is dat zij de gegevens inzien en of zij over de benodigde vaardigheden beschikken om de details van het gegevensbeschermingsbeleid te begrijpen?
 - Hebt u gezorgd voor een toezichtsmechanisme om bij te houden wanneer, waar, hoe, door wie en met welk doel gegevens worden ingezien?

4. Transparantie

Traceerbaarheid:

- ✓ Hebt u maatregelen ingesteld waarmee de traceerbaarheid kan worden gewaarborgd? Daarbij kan het gaan om de documentatie van:
 - de voor het ontwerp en de ontwikkeling van het algoritmische systeem gebruikte methoden:
 - in geval van een KI-systeem op basis van regels moet de programmeermethode of de manier waarop het model is gebouwd, worden gedocumenteerd;
 - in geval van een KI-systeem op basis van leren, moet de trainingsmethode van het algoritme, met inbegrip van welke input er is verzameld en geselecteerd en de manier waarop dit is gebeurd, worden gedocumenteerd.
 - de voor het testen en valideren van het algoritmische systeem gebruikte methoden:
 - in geval van een KI-systeem op basis van regels, moeten de voor het testen en valideren gebruikte scenario's of situaties worden gedocumenteerd;
 - in geval van een model op basis van leren, moet de voor het testen en valideren

gebruikte informatie worden gedocumenteerd.

- de resultaten van het algoritmische systeem:
 - de resultaten of de door het algoritme genomen beslissingen, alsook potentiële andere beslissingen die in andere situaties zouden ontstaan (bijv. voor andere subgroepen of gebruikers) moeten worden gedocumenteerd.

Verklaarbaarheid:

- ✓ Hebt u onderzocht in hoeverre de door het KI-systeem genomen beslissingen en dus het resultaat kunnen worden begrepen?
- ✓ Hebt u gezorgd dat er voor alle gebruikers die een verklaring wensen, een verklaring begrijpelijk kan worden gemaakt van de redenen dat een systeem een bepaalde keuze heeft gemaakt die tot een bepaalde uitkomst heeft geleid?
- ✓ Hebt u onderzocht in hoeverre de beslissingen van het systeem van invloed zijn op het besluitvormingsproces binnen de organisatie?
- ✓ Hebt u onderzocht waarom dit specifieke systeem op dit specifieke gebied is ingezet?
- ✓ Hebt u het bedrijfsmodel van dit systeem onderzocht (bijv. hoe het waarde creëert voor de organisatie)?
- ✓ Hebt u het KI-systeem vanaf het begin ontworpen met interpreteerbaarheid in het achterhoofd?
 - Hebt u onderzocht welk model het eenvoudigst en het meest interpreteerbaar zou zijn voor de betreffende toepassing, en hebt u geprobeerd dit te gebruiken?
 - Hebt u gecontroleerd of u uw trainings- en testgegevens kunt analyseren? Kunt u deze in de loop van de tijd veranderen en bijwerken?
 - Hebt u onderzocht of u, na de training en ontwikkeling van het model, mogelijkheden hebt om de interpreteerbaarheid ervan te beoordelen en of u toegang hebt tot de interne werkstroom van het model?

Communicatie:

- ✓ Hebt u – via een disclaimer of op een andere manier – aan de (eind)gebruikers kenbaar gemaakt dat ze niet met een andere mens, maar met een KI-systeem te maken hebben? Hebt u uw KI-systeem als zodanig aangemerkt?
- ✓ Hebt u mechanismen ingesteld om gebruikers in kennis te stellen van de redenen en criteria achter de resultaten van het KI-systeem?
 - Is dit duidelijk en begrijpelijk aan de beoogde gebruikers kenbaar gemaakt?
 - Hebt u processen ingesteld waardoor rekening wordt gehouden met de feedback van gebruikers, en gebruikt u deze om het systeem aan te passen?
 - Hebt u ook gecommuniceerd over mogelijke of ervaren risico's, zoals vertekening?
 - Hebt u, afhankelijk van de gebruikssituatie, ook nagedacht over de communicatie en transparantie richting andere doelgroepen, derden of het grote publiek?
- ✓ Hebt u duidelijk gemaakt wat het doel van het KI-systeem is en wie of wat er profijt kan hebben van

het product of de dienst?

- Zijn de gebruiksscenario's van het product beschreven en duidelijk kenbaar gemaakt, en zijn daarbij ook alternatieve communicatievormen overwogen om te zorgen dat de boodschap begrijpelijk en geschikt is voor de gebruiker aan wie deze is gericht?
- Hebt u, afhankelijk van de gebruikssituatie, nagedacht over de menselijke psychologie en potentiële beperkingen, zoals het risico op verwarring, voorkeur voor bevestiging of cognitieve vermoeidheid?
- ✓ Hebt u de kenmerken, beperkingen en potentiële tekortkomingen van het KI-systeem duidelijk kenbaar gemaakt:
 - in geval van ontwikkeling: aan degene die het installeert in een product of dienst?
 - in geval van installatie: aan de eindgebruiker of consument?

5. Diversiteit, non-discriminatie en rechtvaardigheid

Voorkomen van onrechtvaardige vertekening:

- ✓ Hebt u gezorgd voor een strategie of een reeks procedures om te voorkomen dat er onrechtvaardige vertekening wordt gecreëerd of versterkt in het KI-systeem, met betrekking tot zowel het gebruik van inputgegevens als het ontwerp van het algoritme?
 - Hebt u de mogelijke beperkingen die voortkomen uit de samenstelling van de gebruikte gegevenssets, onderzocht en erkend?
 - Hebt u nagedacht over de diversiteit en de representativiteit van gebruikers in de gegevens? Hebt u getest op specifieke populaties of problematische gebruikssituaties?
 - Hebt u onderzoek gedaan naar technische hulpmiddelen om uw begrip van de gegevens, het model en de prestaties te verbeteren, en hebt u deze gebruikt?
 - Hebt u processen ingesteld voor het testen en monitoren van potentiële vertekening gedurende de ontwikkelings-, installatie- en gebruiksfase van het systeem?
- ✓ Hebt u, afhankelijk van de gebruikssituatie, gezorgd voor een mechanisme waardoor anderen problemen in verband met vertekening, discriminatie of slechte prestaties van het KI-systeem kunnen melden?
 - Hebt u nagedacht over duidelijke stappen en manieren om te communiceren over hoe en aan wie dergelijke problemen kunnen worden gemeld?
 - Hebt u niet alleen nagedacht over de (eind)gebruikers, maar ook over anderen die mogelijk indirect de gevolgen van het KI-systeem kunnen ondervinden?
- ✓ Hebt u onderzocht of er variabiliteit kan bestaan in de beslissingen die onder gelijke omstandigheden mogelijk zijn?
 - Zo ja, hebt u nagedacht over de mogelijke oorzaken daarvan?
 - In geval van variabiliteit: hebt u een meet- of controlemechanisme vastgesteld voor de potentiële

gevolgen van die variabiliteit voor de grondrechten?

- ✓ Hebt u gezorgd voor een geschikte werkdefinitie van "rechtvaardigheid" die u toepast bij het ontwerpen van KI-systemen?
 - Wordt uw definitie veel gebruikt? Hebt u andere definities overwogen voordat u deze koos?
 - Hebt u gezorgd voor een kwantitatieve analyse of maatstaf om de toegepaste definitie van rechtvaardigheid te meten en te testen?
 - Hebt u mechanismen ingesteld om de rechtvaardigheid in uw KI-systemen te waarborgen? Hebt u andere mogelijke mechanismen overwogen?

Toegankelijkheid en universeel ontwerp:

- ✓ Hebt u gezorgd dat het KI-systeem geschikt is voor een breed scala aan individuele voorkeuren en vermogens?
 - Hebt u onderzocht of het KI-systeem kan worden gebruikt door mensen met bijzondere behoeften of een beperking of door mensen die het risico lopen te worden uitgesloten? Hoe is dit in het ontwerp van het systeem verwerkt en hoe wordt het gecontroleerd?
 - Hebt u gezorgd dat de informatie over het KI-systeem ook toegankelijk is voor gebruikers van ondersteunende technologieën?
 - Hebt u deze gemeenschap betrokken bij de ontwikkelingsfase van het KI-systeem?
- ✓ Hebt u rekening gehouden met de gevolgen van uw KI-systeem voor de potentiële gebruikers?
 - Is het team dat betrokken is bij de bouw van het KI-systeem, representatief voor uw beoogde gebruikers? Is het representatief voor de bevolking als geheel, ook rekening houdend met andere groepen die er mogelijk indirect de gevolgen van ondervinden?
 - Hebt u onderzocht of er personen of groepen zijn die onevenredig kunnen worden getroffen door negatieve gevolgen?
 - Hebt u feedback gekregen van andere teams of groepen die verschillende achtergronden en ervaringen vertegenwoordigen?

Participatie van belanghebbenden:

- ✓ Hebt u nagedacht over een mechanisme om de participatie van verschillende belanghebbenden onderdeel te maken van de ontwikkeling en het gebruik van het KI-systeem?
- ✓ Hebt u de invoering van het KI-systeem in uw organisatie voorbereid door de getroffen werknemers en hun vertegenwoordigers vooraf in kennis te stellen en bij het proces te betrekken?

6. Maatschappelijk en milieuwelzijn

Duurzame en milieuvriendelijke KI:

- ✓ Hebt u mechanismen ingesteld om de gevolgen van de ontwikkeling, de installatie en het gebruik van het KI-systeem voor het milieu te meten (bijv. de door het datacentrum gebruikte energie, het type

energie dat de datacentra gebruiken enz.)?

- ✓ Hebt u gezorgd voor maatregelen om de gevolgen van de levenscyclus van uw KI-systeem voor het milieu te beperken?

Sociale gevolgen:

- ✓ Indien er sprake is van rechtstreekse interactie tussen het KI-systeem en mensen:
 - Hebt u onderzocht of het KI-systeem mensen stimuleert om verbondenheid en empathie jegens het systeem te ontwikkelen?
 - Hebt u gezorgd dat het KI-systeem duidelijk aangeeft dat zijn sociale interactie gesimuleerd is en dat het niet tot "begrip" en "gevoelens" in staat is?
- ✓ Hebt u gezorgd voor een goed begrip van de sociale gevolgen van het KI-systeem? Hebt u bijvoorbeeld onderzocht of er een risico bestaat op het verlies van banen of vaardigheden onder de arbeidskrachten? Welke stappen zijn er genomen om dergelijke risico's tegen te gaan?

Samenleving en democratie:

- ✓ Hebt u de bredere maatschappelijke gevolgen van het gebruik van het KI-systeem voorbij de individuele (eind)gebruiker onderzocht, zoals potentieel indirect getroffen belanghebbenden?

7. Verantwoording

Controleerbaarheid:

- ✓ Hebt u mechanismen ingesteld die de controleerbaarheid van het systeem door interne en/of externe onafhankelijke actoren mogelijk maken, bijvoorbeeld in de vorm van waarborging van de traceerbaarheid en het bijhouden van de processen en resultaten van het KI-systeem?

Minimalisering en verslaglegging van negatieve gevolgen:

- ✓ Hebt u een risico- of effectbeoordeling van het KI-systeem uitgevoerd waarin rekening wordt gehouden met de verschillende belanghebbenden die direct of indirect de gevolgen ervan ondervinden?
- ✓ Hebt u onderwijs- en opleidingskaders gecreëerd voor de ontwikkeling van verantwoordingspraktijken?
 - Welke werknemers of delen van het team zijn erbij betrokken? Gaat het verder dan de ontwikkelingsfase?
 - Wordt binnen deze opleiding ook het potentiële wettelijke kader onderwezen dat van toepassing is op het KI-systeem?
 - Hebt u overwogen een "onderzoeksraad voor ethische KI" of een vergelijkbaar mechanisme in te stellen voor de bespreking van de algehele verantwoording en ethische praktijken, met inbegrip van eventuele grijze gebieden?
- ✓ Is er, naast de interne initiatieven of kaders voor het toezicht op ethiek en verantwoording, ook sprake van een vorm van externe begeleiding of zijn er ook controleprocessen ingesteld?
- ✓ Zijn er processen ingesteld waarmee derden (bijv. leveranciers, consumenten,

distributeurs/verkopers) of werknemers potentiële kwetsbaarheden, risico's of vertekening in het KI-systeem/de KI-toepassing kunnen melden?

Documentatie van afwegingen:

- ✓ Hebt u een mechanisme ingesteld om relevante belangen en waarden die door het KI-systeem in het gedrang komen, alsook potentiële afwegingen daartussen, vast te stellen?
- ✓ Welk proces gebruikt u voor het maken van dergelijke afwegingen? Hebt u gezorgd dat de beslissing bij de afweging is gedocumenteerd?

Vermogen om beroep in te stellen:

- ✓ Hebt u een geschikte reeks mechanismen ingesteld om het mogelijk te maken om in geval van schade of negatieve gevolgen beroep in te stellen?
- ✓ Hebt u mechanismen ingesteld om (eind)gebruikers/derden te voorzien van informatie over de mogelijkheden om beroep in te stellen?

Wij nodigen alle belanghebbenden uit deze controlelijst in de praktijk te testen en feedback te geven op de toepasbaarheid, de volledigheid en de relevantie ervan voor de specifieke KI-toepassing of het specifieke KI-domein, alsook op de overlap of complementariteit met bestaande nalevings- of controleprocessen. Op basis van deze feedback wordt begin 2020 een herziene versie van de controlelijst voor betrouwbare KI aan de Commissie voorgelegd.

Essentiële richtsnoeren uit hoofdstuk III:

- ✓ Stel een **controlelijst** voor betrouwbare KI vast bij de ontwikkeling, de installatie of het gebruik van KI en stem deze af op de specifieke situatie waarin het systeem wordt gebruikt.
- ✓ Bedenk dat een dergelijke controlelijst **nooit uitputtend** zal zijn. Bij betrouwbare KI gaat het niet om het afvinken van elementen op een lijst, maar om het voortdurend vaststellen van vereisten, afwegen van oplossingen en waarborgen van betere resultaten gedurende de hele levenscyclus van het KI-systeem, en om het betrekken van de belanghebbenden bij dit proces.

C. VOORBEELDEN VAN DE MOGELIJKHEDEN EN PUNTEN VAN ZORG DIE KI MET ZICH MEEBRENGT

(121) In het volgende gedeelte geven wij voorbeelden van vormen van ontwikkeling en gebruik van KI die moeten worden aangemoedigd. Ook geven we voorbeelden van gevallen waarin de ontwikkeling, de installatie of het gebruik tegen onze waarden in kan gaan en specifieke punten van zorg met zich mee kan brengen. Er moet een evenwicht worden gevonden tussen wat er met KI moet en wat er *kán* worden gedaan, en er moet voldoende aandacht worden besteed aan wat er *niet* met KI moet worden gedaan.

1. Voorbeelden van de mogelijkheden van betrouwbare KI

(122) Betrouwbare KI kan een uitgelezen mogelijkheid bieden om de beperking te ondersteunen van urgente uitdagingen waarmee de samenleving te maken heeft, zoals een vergrijzende bevolking, toenemende sociale ongelijkheid en milieuvervuiling. Deze potentie is ook wereldwijd zichtbaar, bijvoorbeeld in de

duurzaam ontwikkelingsdoelstellingen van de VN⁵⁷. In het volgende gedeelte gaan we in op de manier waarop een Europese KI-strategie waarmee enkele van deze uitdagingen kunnen worden aangegaan, kan worden gestimuleerd.

a. Klimaatactie en duurzame infrastructuur

- (123) Het aanpakken van de klimaatverandering moet voor beleidsmakers over de hele wereld een topprioriteit zijn. Digitale transformatie en betrouwbare KI hebben een enorme potentie wat betreft het verminderen van het effect van mensen op het milieu en maken efficiënt en effectief gebruik van energie en natuurlijke hulpbronnen mogelijk⁵⁸. Betrouwbare KI kan bijvoorbeeld worden gekoppeld aan big data om de energiebehoeften nauwkeuriger vast te stellen, zodat de energie-infrastructuur en het energieverbruik efficiënter kunnen worden⁵⁹.
- (124) In sectoren als het openbaar vervoer kunnen KI-systemen voor intelligente vervoerssystemen⁶⁰ worden gebruikt om rijen te minimaliseren, routes te optimaliseren, visueel gehandicapten de mogelijkheid te geven zelfstandiger te zijn,⁶¹ bij te dragen aan het koolstofvrij maken van de sector door energie-efficiënte motoren te optimaliseren, en de ecologische voetafdruk te verkleinen, voor een groenere samenleving. Momenteel komt er wereldwijd elke 23 seconden iemand om het leven bij een auto-ongeluk⁶². Met KI-systemen kan het aantal sterfgevallen drastisch worden verlaagd, bijvoorbeeld door betere reactietijden en betere naleving van de regels⁶³.

b. Gezondheid en welzijn

- (125) Betrouwbare KI-technologieën kunnen worden (en worden al) gebruikt om behandelingen slimmer en gericht te maken en om bij te dragen aan de preventie van levensbedreigende ziekten⁶⁴. Artsen en medische beroepsbeoefenaars kunnen in potentie, zelfs voordat mensen ziek worden, een nauwkeurigere en gedetailleerdere analyse uitvoeren van de complexe gezondheidsgegevens van een patiënt en preventieve behandeling op maat verzorgen⁶⁵. In het kader van de vergrijzende Europese bevolking kunnen KI en robotica

⁵⁷ <https://sustainabledevelopment.un.org/?menu=1300>

⁵⁸ Een aantal EU-projecten is gericht op de ontwikkeling van slimme netwerken en energie-opslag, die de potentie hebben om bij te dragen aan een geslaagde digitaal ondersteunde energietransitie, onder meer via op KI gebaseerde en andere oplossingen. Als aanvulling op het werk binnen deze individuele projecten heeft de Commissie het initiatief Bridge opgezet, waardoor bij de lopende Horizon 2020-projecten op het gebied van slimme netwerken en energie-opslag een gemeenschappelijke visie kan worden ontwikkeld op horizontale kwesties: <https://www.h2020-bridge.eu/>.

⁵⁹ Zie bijvoorbeeld het project Encompass: <http://www.encompass-project.eu/>.

⁶⁰ Nieuwe oplossingen op basis van KI helpen steden bij de voorbereiding op de toekomst van mobiliteit. Zie bijvoorbeeld het door de EU gefinancierde project Fabulos: <https://fabulos.eu/>.

⁶¹ Zie bijvoorbeeld het project PRO4VIP, dat onderdeel is van de strategie Europese visie voor 2020 voor de bestrijding van vermijdbare blindheid, met name als gevolg van ouderdom. Mobiliteit en oriëntatie was een van de prioriteitsgebieden van het project.

⁶² <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.

⁶³ Het Europese project UB-Drive is bijvoorbeeld gericht op het aanpakken van de beschreven uitdagingen op het gebied van vervoer door bij te dragen aan het faciliteren van de geleidelijke automatisering van en samenwerking tussen voertuigen, zodat een veiliger, inclusiever en betaalbaarder vervoerssysteem mogelijk wordt: <https://up-drive.eu/>.

⁶⁴ Zie bijvoorbeeld het project Revolver (Repeated Evolution of Cancer – Herhaalde ontwikkeling van kanker): <https://www.healtheuropa.eu/personalised-cancer-treatment/87958/>, of het project Murab, waarbinnen nauwkeurigere biopsieën worden uitgevoerd en dat is gericht op snellere diagnosticering van kanker en andere ziekten: <https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story>.

⁶⁵ Zie bijvoorbeeld het project Live Incite: www.karolinska.se/en/live-incite. Dit consortium van inkopers uit de gezondheidszorg daagt de sector uit om slimme KI- en andere oplossingen te ontwikkelen waarmee leefstijlinterventies in het perioperatieve proces mogelijk worden gemaakt. Het doel is de ontwikkeling van nieuwe innovatieve oplossingen op het gebied van e-gezondheid die patiënten op een gepersonaliseerde manier kunnen stimuleren om zowel voor als na een operatie de benodigde actie te ondernemen in hun leefstijl en zo het resultaat van de gezondheidszorg te optimaliseren.

waardevolle hulpmiddelen zijn ter ondersteuning van zorgverleners en de ouderenzorg.⁶⁶ De gesteldheid van patiënten kan er in realtime mee in de gaten worden gehouden, waardoor er levens kunnen worden gered⁶⁷.

- (126) Betrouwbare KI kan ook op een bredere schaal ondersteuning bieden. Er kunnen bijvoorbeeld algemene tendensen in de gezondheidszorg- en behandelsector mee worden onderzocht en vastgesteld,⁶⁸ met eerdere vaststelling van ziekten, efficiëntere ontwikkeling van geneesmiddelen, gerichtere behandelingen⁶⁹ en uiteindelijk meer geredde levens tot gevolg.

c. Goed onderwijs en digitale transformatie

- (127) Vanwege de nieuwe technologische, economische en ecologische veranderingen moet de samenleving proactiever worden. Overheden, marktleiders, onderwijsinstellingen en vakverenigingen hebben de verantwoordelijkheid om de burgers het nieuwe, digitale tijdperk in te loodsen door te zorgen dat zij over de vaardigheden beschikken die nodig zijn voor de banen van de toekomst. Betrouwbare KI-technologieën kunnen helpen om nauwkeuriger te voorspellen welke banen en beroepen door de technologie zullen worden verstoord, welke nieuwe functies er zullen ontstaan en welke vaardigheden er nodig zullen zijn. Op die manier kunnen overheden, vakverenigingen en sectoren worden geholpen bij het plannen van de (om)scholing van werknemers. Ook kunnen burgers die bang zijn overbodig te zullen worden, zo een manier krijgen om zich te ontwikkelen, zodat ze een nieuwe rol kunnen vervullen.
- (128) Verder kan KI een geweldig hulpmiddel zijn voor het tegengaan van ongelijkheden in het onderwijs en het creëren van gepersonaliseerde en aanpasbare onderwijsprogramma's waarmee iedereen nieuwe kwalificaties, vaardigheden en competenties kan verwerven in overeenstemming met zijn of haar eigen leervermogen⁷⁰. Met behulp van KI zou zowel de leersnelheid als de kwaliteit van het onderwijs kunnen worden verhoogd – van de basisschool tot het hoger onderwijs.

2. Voorbeelden van punten van zorg die KI met zich meebrengt

- (129) Een punt van zorg in verband met KI ontstaat wanneer een van de componenten van betrouwbare KI wordt geschonden. Veel van de onderstaande zorgen vallen al binnen het toepassingsgebied van bestaande

⁶⁶ Binnen het door de EU gefinancierde project Caresses houdt men zich bezig met robots voor de ouderenzorg, waarbij het accent ligt op de culturele gevoeligheid ervan: ze passen hun gedrag en hun manier van spreken aan aan de cultuur en de gewoonten van de oudere die ze bijstaan: <http://caressesrobot.org/en/project/>. Zie ook de KI-toepassing genaamd Alfred, een virtuele assistent die ouderen helpt om actief te blijven: <https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active>. Daarnaast wordt er binnen het project Empattics (EMpowering PATients for a BeTTER Information and improvement of the Communication Systems – Betere informatie voor patiënten en verbetering van de communicatiesystemen) onderzocht en beschreven hoe beroepsbeoefenaars in de gezondheidszorg en patiënten ICT-technologieën, met inbegrip van KI-systemen, gebruiken om samen interventies te plannen en om de voortgang van hun fysieke en geestelijke gesteldheid van patiënten te monitoren: www.empattics.eu.

⁶⁷ Zie bijvoorbeeld de MyHealth Avatar (www.myhealthavatar.eu), die een digitale weergave van de gezondheidsstatus van een patiënt biedt. In het kader van dit onderzoeksproject zijn een app en een onlineplatform opgezet om digitale informatie over uw gezondheidsstatus op de lange termijn te verzamelen en voor u inzichtelijk te maken. Dit gebeurt in de vorm van een levenslange gezondheidsmetgezel ("avatar"). De MyHealth Avatar voorspelt ook uw risico op een beroerte, diabetes, hart- en vaatziekten en een te hoge bloeddruk.

⁶⁸ Zie bijvoorbeeld het project Enrichme (www.enrichme.eu), waarbinnen de geleidelijke daling van de cognitieve capaciteit binnen de vergrijzende bevolking wordt aangepakt. Met behulp van een geïntegreerd platform voor omgevingsondersteund wonen en een mobiele dienstrobot voor monitoring en interactie op de lange termijn kunnen ouderen langer zelfstandig en actief blijven.

⁶⁹ Zie bijvoorbeeld het gebruik van KI door Sophia Genetics, dat gebruikmaakt van statistische inferentie, patroonherkenning en automatisch leren om de waarde van genomica- en radiomicagegevens te maximaliseren: <https://www.sophiagenetics.com/home.html>.

⁷⁰ Zie bijvoorbeeld het project MaTHiSiS, dat is gericht op het bieden van een oplossing voor leren op basis van affect in een comfortabele leeromgeving bestaande uit hoogwaardige technologische apparatuur en algoritmen: (<http://mathisis-project.eu/>). Zie ook de Watson Classroom van IBM en het platform van Century Tech.

wettelijke voorschriften die verplicht zijn en dus moeten worden nageleefd. Zelfs in situaties waar is aangetoond dat de wettelijke voorschriften zijn nageleefd, is het echter mogelijk dat hiermee niet alle ethische punten van zorg die kunnen ontstaan, zijn verholpen. Aangezien ons begrip van de geschiktheid van regels en ethische beginselen altijd een ontwikkeling doormaakt en in de loop van de tijd kan veranderen, is het mogelijk dat de onderstaande, niet-uitputtende lijst van punten van zorg in de toekomst wordt ingekort, uitgebreid, aangepast of bijgewerkt.

a. Personen herkennen en volgen met KI

- (130) Met KI is almaar efficiëntere herkenning mogelijk van individuele personen door zowel publieke als particuliere entiteiten. Interessante voorbeelden van schaalbare KI-herkenningstechnologie zijn gezichtsherkenning en andere onvrijwillige herkenningmethoden met behulp van biometrische gegevens (zoals leugendetectie, persoonlijkheidsbeoordeling via micro-expressies en automatische stemherkenning). Soms is herkenning van personen het gewenste resultaat en in lijn met ethische beginselen (bijvoorbeeld bij het opsporen van fraude, witwassen of terrorismefinanciering). Automatische herkenning brengt echter grote juridische en ethische zorgen met zich mee, aangezien deze op veel psychologische en sociaal-culturele niveaus onverwachte gevolgen kan hebben. Een evenredig gebruik van controletechnieken voor KI is nodig om de autonomie van Europese burgers in stand te houden. Voor de verwezenlijking van betrouwbare KI is het essentieel om helder af te bakenen of, wanneer en hoe KI mag worden gebruikt voor de automatische herkenning van personen en om verschil te maken tussen de herkenning van een persoon en het traceren en volgen van een persoon, alsook tussen gericht toezicht en grootschalig toezicht. De toepassing van dergelijke technologieën moet duidelijk gerechtvaardigd zijn op grond van bestaande wetgeving⁷¹. Indien de wettelijke basis voor een dergelijke activiteit "toestemming" is, moeten er praktische middelen⁷² worden ontwikkeld waardoor de te geven zinvolle en gecontroleerde toestemming automatisch kan worden herkend door KI- of vergelijkbare technologieën. Hetzelfde geldt voor het gebruik van "anonieme" persoonsgegevens waarvan de anonimisering kan worden teruggedraaid.

b. Verborgene KI-systemen

- (131) Mensen moeten altijd weten of ze rechtstreeks contact hebben met een ander mens of met een machine. Het is de verantwoordelijkheid van beroepsbeoefenaars op het gebied van KI om dit op betrouwbare wijze te verwezenlijken. Zij moeten er daarom voor zorgen dat mensen bewust worden gemaakt van het feit dat – of kunnen navragen en controleren of – zij met een KI-systeem te maken hebben (bijvoorbeeld door middel van heldere en transparante disclaimers). Opgemerkt zij dat er grensgevallen bestaan die de zaak extra ingewikkeld maken (zoals een door een mens ingesproken stem die door een KI-systeem wordt gefilterd). Bedacht moet worden dat de verwarring tussen mensen en machines verschillende gevolgen kan hebben, zoals hechting, beïnvloeding of vermindering van de waarde van het mens-zijn.⁷³ De ontwikkeling van op mensen lijkende robots⁷⁴ moet daarom worden onderworpen aan zorgvuldige ethische controle.

c. Schending van grondrechten door beoordeling van burgers met behulp van KI

- (132) Samenlevingen moeten ernaar streven de vrijheid en autonomie van alle burgers te beschermen. Iedere vorm van beoordeling van burgers kan leiden tot verlies van deze autonomie en kan het beginsel van non-discriminatie in gevaar brengen. Beoordeling mag alleen worden gebruikt als deze duidelijk gerechtvaardigd is en als de maatregelen evenredig en rechtvaardig zijn. Normatieve beoordeling van burgers (algemene

⁷¹ Met betrekking hiertoe kan worden verwezen naar artikel 6 van de AVG, waarin onder andere is bepaald dat de verwerking alleen rechtmatig is indien deze een geldige rechtsgrond heeft.

⁷² Zoals blijkt uit de huidige mechanismen voor het geven van geïnformeerde toestemming op het internet, geven consumenten doorgaans toestemming zonder hier goed over na te denken. Deze mechanismen kunnen dus nauwelijks als praktisch worden beschouwd.

⁷³ Madary & Metzinger (2016). Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. Frontiers in Robotics and AI, 3(3).

⁷⁴ Dit geldt ook voor KI-gestuurde avatars.

beoordeling van "morele persoonlijkheid" of "ethische integriteit") op *alle* aspecten en op grote schaal door overheidsinstanties of particuliere actoren vormt een bedreiging voor deze waarden, met name wanneer het gebruik ervan niet in overeenstemming is met de grondrechten, niet evenredig is of geen afgebakend en kenbaar gemaakt doel heeft.

- (133) Beoordeling van burgers – op grote of kleinere schaal – wordt tegenwoordig al vaak gebruikt voor louter beschrijvende en domeinspecifieke beoordelingen (zoals schoolsystemen, e-learning en rijbewijzen). Zelfs bij deze smallere toepassingen moet er een volledig transparante procedure voor burgers beschikbaar worden gemaakt, met inbegrip van informatie over het proces, het doel en de methodologie van de beoordeling. Opgemerkt zij dat transparantie niet voldoende is om discriminatie te voorkomen, rechtvaardigheid te garanderen en alle problemen rondom beoordeling te verhelpen. Idealiter moet de mogelijkheid worden geboden om, indien mogelijk, zonder negatieve gevolgen te besluiten niet deel te nemen aan het beoordelingsmechanisme. Anders moeten er mechanismen worden ingesteld om de beoordelingen aan te vechten en te corrigeren. Dit is met name van belang in situaties waar de macht ongelijk tussen de partijen is verdeeld. In gevallen waarin dit nodig is om naleving van de grondrechten te verwezenlijken, moeten dergelijke onttrekkingsmogelijkheden in het ontwerp van de technologie worden gewaarborgd, wat noodzakelijk is in een democratische samenleving.

d. Dodelijke autonome wapensystemen

- (134) Momenteel houdt een onbekend aantal landen en sectoren zich bezig met onderzoek naar en de ontwikkeling van dodelijke autonome wapensystemen, variërend van raketten met de mogelijkheid tot selectieve aanvallen tot lerende machines met de cognitieve vaardigheden om zonder menselijke interventie te besluiten wanneer, waar en met wie ze vechten. Dit brengt ethische punten van zorg met zich mee, zoals het feit dat er een oncontroleerbare wapenwedloop van historisch ongekeerde omvang zou kunnen ontstaan en militaire situaties waarbij er vrijwel geen sprake meer is van menselijke controle en waarbij de risico's van storingen niet worden aangepakt. Het Europees Parlement heeft opgeroepen tot de urgente ontwikkeling van een gemeenschappelijk, juridisch bindend standpunt waarin wordt ingegaan op ethische en juridische vragen op het gebied van menselijke controle, toezicht, verantwoording en de uitvoering van internationale mensenrechtenwetgeving, internationale humanitaire wetgeving en militaire strategieën.⁷⁵ Wijzend op het doel van de Europese Unie om de vrede te bevorderen, zoals vastgelegd in artikel 3 van het Verdrag betreffende de Europese Unie, staan wij achter de resolutie van het Parlement van 12 september 2018 en alle verwante inspanningen op het gebied van dodelijke autonome wapensystemen, en beogen wij deze te ondersteunen.

e. Potentiële punten van zorg op de langere termijn

- (135) De ontwikkeling van KI is nog altijd domeinspecifiek en vereist dat goed opgeleide menselijke wetenschappers en technici de doelen ervan exact specificeren. Als we naar de toekomst kijken met een langere periode in gedachten, kunnen we echter speculeren over bepaalde kritieke punten van zorg op de langere termijn⁷⁶. Uit een op risico gebaseerde benadering is op te maken dat we rekening moeten blijven houden met deze zorgen met het oog op mogelijke onbekende factoren en "zwarte zwanen".⁷⁷ Omdat dat deze punten van zorg grote gevolgen kunnen hebben en gezien de huidige onzekerheid rondom de bijbehorende ontwikkelingen, moeten er regelmatig controles worden uitgevoerd op deze punten.

⁷⁵ Resolutie 2018/2752(RSP) van het Europees Parlement.

⁷⁶ Sommigen zijn van mening zijn dat volledige kunstmatige intelligentie, kunstmatig bewustzijn, kunstmatige morele actoren, een superintelligentie of transformatieve KI (die momenteel niet bestaan) voorbeelden van dergelijke langetermijnzorgen kunnen zijn, maar deze worden door veel anderen als onrealistisch beschouwd.

⁷⁷ Een zwarte zwaan is een zeer zeldzame gebeurtenis met grote gevolgen – zo zeldzaam dat deze mogelijk nog nooit is voorgekomen. Daarom kan gewoonlijk slechts met zeer weinig zekerheid worden vastgesteld hoe groot de kans erop is.

D. CONCLUSIE

- (136) Dit document bevat de ethische richtsnoeren voor KI die zijn ontwikkeld door de deskundigengroep op hoog niveau inzake kunstmatige intelligentie (AI HLEG).
- (137) Wij erkennen de positieve gevolgen die KI-systemen al teweeg hebben gebracht en zullen blijven brengen, zowel commercieel als maatschappelijk. Het gaat ons er echter evenzeer om dat de risico's en andere negatieve gevolgen die aan deze technologieën zijn verbonden, op correcte en evenredige wijze worden aangepakt in het kader van de KI-toepassing. KI is een technologie die zowel transformatief als verstorend is, en de ontwikkeling ervan is de afgelopen jaren mogelijk gemaakt door de beschikbaarheid van enorme hoeveelheden digitale gegevens, grote technologische vooruitgang op het gebied van rekenkracht en opslagcapaciteit en aanzienlijke wetenschappelijke en technische innovatie op het gebied van KI-methoden en -hulpmiddelen. KI-systemen zullen de samenleving en burgers blijven beïnvloeden op manieren die we ons nog niet kunnen voorstellen.
- (138) In het kader daarvan is het belangrijk om KI-systemen te creëren die te vertrouwen zijn, want mensen zullen alleen volledig en vol overtuiging de vruchten ervan kunnen plukken als de technologie, met inbegrip van de achterliggende processen en mensen, betrouwbaar is. Bij het opstellen van deze richtsnoeren is betrouwbare KI daarom onze kernambitie geweest.
- (139) Betrouwbare KI is opgebouwd uit drie componenten: de KI moet 1) wettig zijn, door eerbiediging van alle toepasselijke wet- en regelgeving te waarborgen, 2) ethisch zijn, door naleving van ethische beginselen en waarden te waarborgen, en 3) robuust zijn uit zowel technisch als sociaal oogpunt om te zorgen dat KI-systemen niet ongewild schade aanrichten, zelfs al zijn de bedoelingen goed. Elke component is nodig, maar op zichzelf niet voldoende om betrouwbare KI te bewerkstelligen. In het ideale scenario sluiten alle drie de componenten op elkaar aan en valt de werking ervan gedeeltelijk samen. Waar spanningen ontstaan, moeten we proberen de componenten op één lijn te brengen.
- (140) In hoofdstuk I hebben we de grondrechten uiteengezet, evenals een bijbehorende reeks ethische beginselen die cruciaal zijn in een KI-context. In hoofdstuk II hebben we zeven kernvereisten beschreven waaraan KI-systemen moeten voldoen om betrouwbare KI te verwezenlijken. We hebben technische en niet-technische methoden voorgesteld die van pas kunnen komen bij de uitvoering ervan. Tot slot hebben we in hoofdstuk III een controlelijst voor betrouwbare KI gegeven die kan helpen bij de operationalisering van de zeven vereisten. In een laatste gedeelte hebben we voorbeelden gegeven van voordelige kansen en punten van zorg in verband met KI-systemen waarover we verdere discussie hopen te stimuleren.
- (141) Europa heeft een uniek perspectief vanwege zijn focus op het centraal stellen van de burger bij al zijn inspanningen. Deze focus is verweven in het DNA van de Europese Unie middels de verdragen waarop deze is gebouwd. Dit document maakt onderdeel uit van een visie waarbij betrouwbare KI wordt bevorderd. Wij zijn ervan overtuigd dat betrouwbare KI de grondslag moet zijn op basis waarvan Europa een leidende positie kan verwerven op het gebied van innovatieve, geavanceerde KI-systemen. Deze ambitieuze visie zal ertoe bijdragen dat Europese burgers tot bloei komen, zowel individueel als collectief. Ons doel is het creëren van een cultuur van "betrouwbare KI voor Europa", zodat iedereen op dusdanige wijze kan profiteren van de voordelen van KI dat eerbiediging wordt gewaarborgd van onze basiswaarden: grondrechten, democratie en de rechtsstaat.

WOORDENLIJST

(142) Deze woordenlijst heeft betrekking op de richtsnoeren en is bedoeld als hulpmiddel om de in dit document gebruikte termen te begrijpen.

Systemen op basis van kunstmatige intelligentie of KI-systemen

(143) Systemen op basis van kunstmatige intelligentie (KI) zijn door mensen ontworpen⁷⁸ softwaresystemen (en mogelijk ook hardwaresystemen) die, met een complex doel, in de fysieke of digitale dimensie werken door via gegevensverzameling hun omgeving waar te nemen, de verzamelde gestructureerde of ongestructureerde gegevens te interpreteren, te redeneren op basis van de uit deze gegevens verkregen kennis of de verkregen informatie te verwerken en te beslissen met welke handeling(en) het gestelde doel het best kan worden bereikt. KI-systemen kunnen gebruikmaken van symbolische regels of een numeriek model leren en kunnen hun gedrag ook aanpassen door te analyseren welke invloed hun eerdere handelingen op de omgeving hebben.

(144) Als wetenschappelijke discipline omvat KI verschillende benaderingen en technieken, zoals automatisch leren (waarvan deep learning en reinforcement-lernen specifieke voorbeelden zijn), automatisch redeneren (waaronder plannen, inroosteren, kennisrepresentatie en redeneren, zoeken en optimaliseren) en robotica (waaronder controle, waarneming, sensoren en actuatoren, alsook de integratie van alle andere technieken in cyber-fysieke systemen).

(145) Tegelijk met dit document wordt ook een afzonderlijk door de AI HLEG opgesteld document gepubliceerd, waarin de in dit document gehanteerde definitie van "KI-systemen" verder wordt uitgewerkt. Dit document is getiteld: "Een definitie van KI: de belangrijkste capaciteiten en wetenschappelijke disciplines".

Beroepsbeoefenaars op het gebied van KI

(146) Met beroepsbeoefenaars op het gebied van KI bedoelen we alle personen of organisaties die KI-systemen ontwikkelen (met inbegrip van onderzoek naar en ontwerp van deze systemen of de aanlevering van gegevens ervoor), installeren (met inbegrip van de toepassing) of gebruiken, met uitzondering van de personen of organisaties die KI-systemen gebruiken als eindgebruiker of consument.

Levenscyclus van een KI-systeem

(147) De levenscyclus van een KI-systeem omvat de ontwikkelingsfase (met inbegrip van onderzoek, ontwerp, aanlevering van gegevens en beperkte testen), de installatiefase (met inbegrip van de toepassing) en de gebruiksfase.

Controleerbaarheid

(148) Controleerbaarheid verwijst naar het vermogen van een KI-systeem om een controle van de algoritmen, gegevens en ontwerpprocessen van het systeem te ondergaan en vormt een van de zeven vereisten waaraan een betrouwbaar KI-systeem moet voldoen. Dat betekent niet noodzakelijkerwijs dat informatie over bedrijfsmodellen en intellectuele eigendom in verband met het KI-systeem altijd openbaar beschikbaar moet zijn. Door de traceerbaarheid en registratiemechanismen vanaf de vroege ontwerpfase van het KI-systeem te waarborgen kan worden bijgedragen aan de controleerbaarheid van het systeem.

Vertekening

(149) Vertekening is een neiging om een persoon, object of standpunt te bevoordelen of te benadelen. Vertekening kan op veel manieren in KI-systemen voorkomen. Bij gegevensgestuurde KI-systemen, zoals de systemen die worden geproduceerd middels automatisch leren, kan vertekening in de gegevensverzameling en de training bijvoorbeeld leiden tot een KI-systeem dat vertekening vertoont. Bij KI-systemen op basis van logica, zoals op

⁷⁸

Mensen ontwerpen KI-systemen rechtstreeks, maar kunnen ook KI-technieken gebruiken om hun ontwerp te optimaliseren.

regels gebaseerde systemen, kan vertekening ontstaan door de manier waarop een kennistechnicus de regels ziet die in een bepaalde situatie gelden. Vertekening kan ook ontstaan door onlinelieren en aanpassing middels interactie, of door personalisering, waarbij gebruikers aanbevelingen of informatie krijgen die op hun smaak zijn afgestemd. Vertekening heeft niet noodzakelijkerwijs te maken met menselijke vooringenomenheid of met door de mens gestuurde gegevensverzameling, maar kan bijvoorbeeld ontstaan doordat een systeem slechts in een beperkt aantal situaties wordt gebruikt, waardoor het niet naar andere situaties kan worden uitgebreid. Vertekening kan goed of slecht en bedoeld of onbedoeld zijn. In bepaalde gevallen kan vertekening leiden tot discriminerende en/of onrechtvaardige resultaten. In dit document wordt dat onrechtvaardige vertekening genoemd.

Ethiek

- (150) Ethiek is een academische discipline die een tak vormt van de filosofie. In algemene termen wordt binnen de ethiek ingegaan op vragen als "wat is een goede handeling?", "wat is de waarde van een mensenleven?", "wat is rechtvaardigheid?" en "wat is het goede leven?". De academische ethiek kent vier grote onderzoeksgebieden: i) meta-ethiek, waarbij het vooral gaat om de betekenis en duiding van normatieve zinnen en de vraag hoe de waarheidswaarde ervan kan worden bepaald (als die er is); ii) normatieve ethiek, het praktische middel om een morele handelwijze te bepalen door de normen voor goede en slechte handelingen te onderzoeken en een waarde aan specifieke handelingen toe te kennen; iii) descriptieve ethiek, gericht op een empirisch onderzoek van het morele gedrag en de morele overtuigingen van mensen; en iv) toegepaste ethiek, waarbij het gaat om wat we verplicht moeten (of wat we mogen) doen in een specifieke (vaak historisch nieuwe) situatie of een bepaald domein van (vaak historisch ongekende) mogelijke handelingen. Binnen de toegepaste ethiek wordt ingegaan op situaties uit het echte leven waarin onder tijdsdruk en vaak met beperkte rationaliteit beslissingen moeten worden genomen. KI-ethiek wordt over het algemeen beschouwd als een voorbeeld van toegepaste ethiek en is gericht op de normatieve problemen die voortkomen uit het ontwerp, de ontwikkeling, de toepassing en het gebruik van KI.
- (151) Binnen ethische discussies worden vaak de termen "moreel" en "ethisch" gebruikt. De term "moreel" verwijst naar de concrete, feitelijke gedragspatronen, de gewoonten en de conventies die specifieke culturen, groepen of personen op een bepaald moment vertonen. De term "ethisch" verwijst naar een evaluatieve beoordeling van dergelijke concrete handelingen en gedragingen vanuit een systematisch, academisch perspectief.

Ethische KI

- (152) In dit document wordt met ethische KI verwezen naar dusdanige ontwikkeling en installatie en dusdanig gebruik van KI dat naleving van de ethische normen, met inbegrip van grondrechten als bijzondere morele rechten, ethische beginselen en verwante kernwaarden, wordt gewaarborgd. Dit is het tweede van de drie kernelementen die noodzakelijk zijn voor de verwezenlijking van betrouwbare KI.

KI waarbij de mens centraal staat

- (153) Bij de benadering van KI waarbij de mens centraal staat, wordt ernaar gestreefd te zorgen dat menselijke waarden centraal staan bij de manier waarop KI-systemen worden ontwikkeld, geïnstalleerd, gebruikt en gemonitord. Dit gebeurt door eerbiediging van de grondrechten te waarborgen, met inbegrip van de rechten die worden beschreven in de verdragen van de Europese Unie en het Handvest van de grondrechten van de Europese Unie. Al deze rechten zijn verenigd op grond van een gemeenschappelijke grondslag die is geworteld in respect voor de menselijke waardigheid en waarin de mens een unieke en onvervreembare morele status heeft. Bij deze benadering hoort ook het rekening houden met de natuurlijke omgeving en andere levende wezens die deel uitmaken van het menselijke ecosysteem, alsook een duurzame benadering om toekomstige generaties de mogelijkheid te geven tot bloei te komen.

Red teaming

- (154) Red teaming is de praktijk waarbij een "red team" of onafhankelijke groep een kwaadwillige rol of een kwaadwillig standpunt inneemt en zo een organisatie uitdaagt om haar effectiviteit te vergroten. Red teaming wordt voornamelijk gebruikt om potentiële kwetsbaarheden op het gebied van beveiliging vast te stellen en aan te pakken.

Reproduceerbaarheid

- (155) Reproduceerbaarheid beschrijft of een KI-experiment hetzelfde gedrag vertoont wanneer het onder gelijke omstandigheden wordt herhaald.

Robuuste KI

- (156) Onder de robuustheid van een KI-systeem valt zowel de technische robuustheid (gepastheid in een bepaalde context, zoals het toepassingsgebied of de fase van de levenscyclus) als de robuustheid uit sociaal oogpunt (zorgen dat het KI-systeem naar behoren rekening houdt met de context en omgeving waarin het werkzaam is). Deze robuustheid is cruciaal om te zorgen dat er geen ongewilde schade kan ontstaan, zelfs al zijn de bedoelingen goed. Robuustheid is de derde van de drie componenten die noodzakelijk zijn voor de verwezenlijking van betrouwbare KI.

Belanghebbenden

- (157) Met belanghebbenden bedoelen we iedereen die onderzoek doet naar KI of die deze ontwikkelt, installeert of gebruikt, alsook degenen die (direct of indirect) de gevolgen van KI ondervinden – met inbegrip van onder meer bedrijven, organisaties, onderzoekers, overheidsdiensten, instellingen, maatschappelijke organisaties, overheden, regelgevers, sociale partners, personen, burgers, werknemers en consumenten.

Traceerbaarheid

- (158) De traceerbaarheid van een KI-systeem verwijst naar het vermogen om de gegevens en de ontwikkelings- en installatieprocessen van het systeem bij te houden, gewoonlijk door middel van gedocumenteerde opgeslagen identificatie.

Vertrouwen

- (159) Wij gebruiken de volgende definitie uit de literatuur: "Vertrouwen wordt gezien als: 1) een reeks specifieke overtuigingen in verband met welwillendheid, bekwaamheid, integriteit en voorspelbaarheid (vertrouwensovertuigingen); 2) de bereidheid van één partij om in een riskante situatie op een ander te vertrouwen (vertrouwensintentie); of 3) de combinatie van deze elementen."⁷⁹ Hoewel "vertrouwen" gewoonlijk geen eigenschap is die aan machines wordt toegeschreven, willen we in dit document benadrukken hoe belangrijk het is er niet alleen op te kunnen vertrouwen dat KI-systemen in lijn zijn met wettelijke voorschriften en ethische normen en robuust zijn, maar ook dat dit vertrouwen kan worden toegeschreven aan alle mensen en processen die bij de levenscyclus van het KI-systeem betrokken zijn.

Betrouwbare KI

- (160) Betrouwbare KI is opgebouwd uit drie componenten: de KI moet 1) wettig zijn, door eerbiediging van alle toepasselijke wet- en regelgeving te waarborgen, 2) ethisch zijn, door respect voor ethische beginselen en waarden te tonen en naleving ervan te waarborgen, en 3) robuust zijn uit zowel technisch als sociaal oogpunt, aangezien KI-systemen ongewild schade kunnen aanrichten, zelfs al zijn de bedoelingen goed. Betrouwbare KI heeft niet alleen betrekking op de betrouwbaarheid van het KI-systeem zelf, maar omvat ook de betrouwbaarheid van alle processen en actoren die deel uitmaken van de levenscyclus van het systeem.

Kwetsbare personen en groepen

⁷⁹ Siau, K., Wang, W. (2018), Building Trust in Artificial Intelligence, Machine Learning, and Robotics, *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), S. blz. 47–53.

(161) Vanwege de heterogeniteit ervan bestaat er geen algemeen aanvaarde of breed gedragen definitie van kwetsbare personen. Wat een kwetsbare persoon of groep is, hangt van de context af. Tijdelijke levensgebeurtenissen (zoals jeugd of ziekte), marktfactoren (zoals ongelijkheid wat betreft beschikking over informatie of marktmacht), economische factoren (zoals armoede), factoren in verband met iemands identiteit (zoals geslacht, religie of cultuur) en andere factoren kunnen een rol spelen. In artikel 21 van het Handvest van de grondrechten van de EU over non-discriminatie worden de volgende gronden genoemd, die, naast andere gronden, een referentiepunt kunnen vormen: geslacht, ras, kleur, etnische of sociale afkomst, genetische kenmerken, taal, godsdienst of overtuigingen, politieke of andere denkbeelden, het behoren tot een nationale minderheid, vermogen, geboorte, een handicap, leeftijd en seksuele gerichtheid. In andere rechtsartikelen wordt ingegaan op de rechten van specifieke groepen, naast de bovenstaande rechten. Een dergelijke lijst is nooit uitputtend en kan in de loop van de tijd veranderen. Een kwetsbare groep is een groep personen die één of meerdere kenmerken van kwetsbaarheid gemeen hebben.

Dit document is opgesteld door de leden van de deskundigengroep op hoog niveau inzake

KI,

die hieronder in alfabetische volgorde worden genoemd.

Pekka Ala-Pietilä, voorzitter van de AI HLEG KI Finland, Huhtamaki, Sanoma	Pierre Lucas Orgalim – de technologische sectoren van Europa
Wilhelm Bauer Fraunhofer	Ieva Martinkenaite Telenor
Urs Bergmann – corapporteur Zalando	Thomas Metzinger – corapporteur JGU Mainz & Europese Vereniging van Universiteiten
Mária Bielíková Slowaakse Technische Universiteit in Bratislava	Catelijne Muller ALLAI Nederland & EESC
Cecilia Bonefeld-Dahl – corapporteur DigitalEurope	Markus Noga SAP
Yann Bonnet ANSSI	Barry O'Sullivan, vicevoorzitter van de AI HLEG University College Cork
Loubna Bouarfa OKRA	Ursula Pacht BEUC
Stéphane Brunessaux Airbus	Nicolas Petit – corapporteur Universiteit van Luik
Raja Chatila IEEE-initiatief Ethics of Intelligent and Autonomous Systems (Ethiek van intelligente en autonome systemen) & Sorbonne Université	Christoph Peylo Bosch
Mark Coeckelbergh Universiteit van Wenen	Iris Plöger BDI
Virginia Dignum – corapporteur Universiteit van Umeå	Stefano Quintarelli Garden Ventures
Luciano Floridi Universiteit van Oxford	Andrea Renda Staf Europacollege & CEPS
Jean-François Gagné – corapporteur Element AI	Francesca Rossi IBM
Chiara Giovannini ANEC	Cristina San José Europese Bankfederatie
Joanna Goodey Bureau voor de grondrechten	George Sharkov Digital SME Alliance
Sami Haddadin Munich School of Robotics and MI	Philipp Slusallek Duits onderzoekscentrum voor KI (DFKI)
Gry Hasselbalch De denkdootank DataEthics & Universiteit van Kopenhagen	Françoise Soulié Fogelman KI-consultant
Fredrik Heintz Universiteit van Linköping	Saskia Steinacker – corapporteur Bayer
Fanny Hidvegi Access Now	Jaana Tallinn Ambient Sound Investment
Eric Hilgendorf Universiteit van Würzburg	Thierry Tingaud STMicroelectronics
Klaus Höckner Hilfsgemeinschaft der Blinden und Sehschwachen	Jakob Uszkoreit Google
Mari-Noëlle Jégo-Laveissière Orange	Aimee Van Wynsberghe – corapporteur TU Delft
Leo Kärkkäinen Nokia Bell Labs	Thiébaut Weber EVV
Sabine Theresia Köszegi TU Wenen	Cecile Wendling AXA
Robert Kroplewski Advocaat & adviseur voor de Poolse overheid	Karen Yeung – corapporteur Universiteit van Birmingham
Elisabeth Ling RELX	

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe en Karen Yeung hebben als rapporteurs gefungeerd voor dit document.

Pekka Ala-Pietilä is voorzitter van de AI HLEG. Barry O'Sullivan is vicevoorzitter en coördineert het tweede product van de AI HLEG.

Nozha Boujemaa, die tot 1 februari 2019 vicevoorzitter was en het eerste product coördineerde, heeft ook bijgedragen aan de inhoud van dit document.

Nathalie Smuha heeft redactionele ondersteuning geboden.