Sigra

Vereisten van betrouwbare AI

Privacy bewustwordingscampagne

2024

Samen voor betere zorg & welzijn

Expertisecentrum Privacy & Informatieveiligheid

Privacy & Informatieveiligheid Bewustwordingscampagne

Digitale ethiek, balanceren tussen technologie en ethiek
Sigra organiseert jaarlijks in oktober de Privacy & Informatieveiligheid Bewustwordingscampagne. In 2024 is de campagne gericht op de **ethische kant** van het werken met nieuwe digitale zorg en welzijnstechnologieën. Digitale ethiek in zorg en welzijn richt zich op privacy, transparantie, data-ethiek, AI-algoritmen en de impact van technologie op patiënten en zorgprofessionals.

Campagnemiddelen
Voor Sigra-leden hebben we diverse middelen gecreëerd waarmee we professionals informeren en inspireren over digitale ethiek. Deze digitale brochure is een verdieping van de richtlijnen.

Bekijk **Sigra.nl** voor meer middelen om in je eigen organisatie te verspreiden.

Sigra

Expertisecentrum Privacy & Informatieveiligheid

Ethische richtlijnen voor betrouwbare AI

De ethische richtlijnen voor betrouwbare AI, opgesteld door de deskundigengroep op hoog niveau inzake AI van de Europese Commissie, omvatten een reeks principes en voorwaarden.

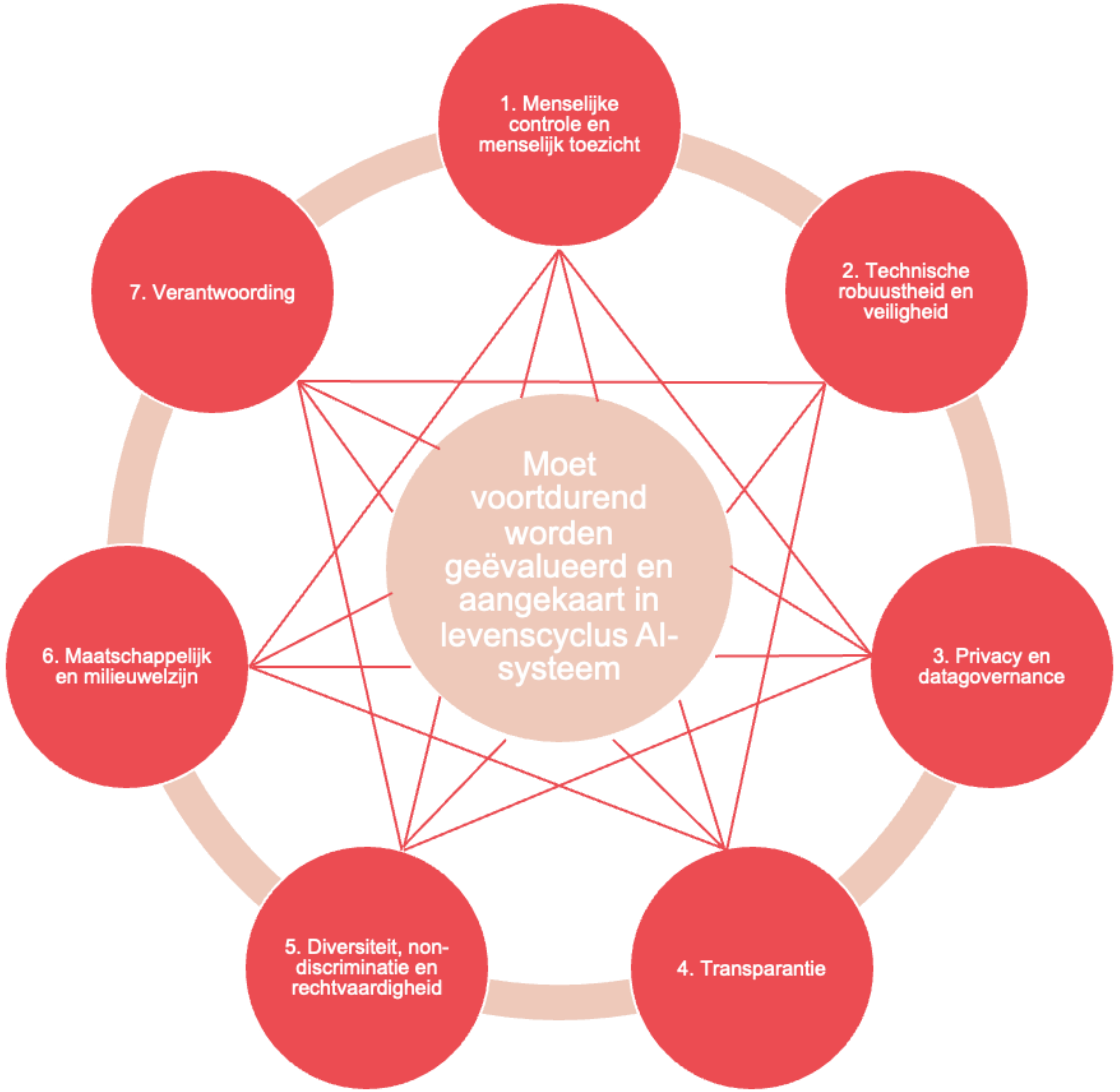
Deze voorwaarden moeten ervoor zorgen dat AI-systemen niet alleen technisch robuust en veilig zijn, maar ook ethisch verantwoord en in overeenstemming met de fundamentele rechten en waarden van de Europese Unie. Deze richtlijnen zijn bedoeld om te waarborgen dat AI-systemen de menselijke autonomie respecteren, transparant en eerlijk zijn en bijdragen aan het welzijn van de samenleving en het milieu.

Om dit te bereiken, hebben de richtlijnen zeven vereisten vastgesteld.

Sigra

Expertisecentrum Privacy & Informatieveiligheid

Vereisten van betrouwbare AI



Sigra

Expertisecentrum Privacy & Informatieveiligheid

Vereisten van betrouwbare AI

1. Menselijke controle en menselijk toezicht

AI-systemen moeten menselijke autonomie en beslissingen ondersteunen. Dit betekent dat ze een democratische, florerende en gelijkwaardige samenleving moeten bevorderen door gebruikerscontrole en grondrechten te ondersteunen, en menselijk toezicht mogelijk te maken.

Grondrechten

AI-systemen kunnen zowel positieve als negatieve effecten hebben op grondrechten. 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. Echter, vanwege hun reikwijdte en capaciteit kunnen ze ook negatieve gevolgen hebben. Daarom moet er een effectbeoordeling van grondrechten worden uitgevoerd vóór de ontwikkeling van het systeem, om te evalueren of risico's beperkt of gerechtvaardigd kunnen worden. Daarnaast moeten er mechanismen zijn om externe feedback te ontvangen over mogelijke inbreuken op grondrechten.

Menselijke controle

Gebruikers moeten autonome beslissingen kunnen nemen over AI-systemen en de kennis en hulpmiddelen krijgen om deze systemen te begrijpen en ermee om te gaan. Ze moeten het systeem kunnen controleren of aanvechten indien nodig. AI-systemen moeten mensen helpen betere keuzes te maken die overeenkomen met hun doelen. Het principe van gebruikersautonomie moet centraal staan, en gebruikers hebben het recht om niet onderworpen te worden aan beslissingen die uitsluitend op geautomatiseerde verwerking zijn gebaseerd en hen significant beïnvloeden.

Menselijk toezicht

Menselijk toezicht zorgt ervoor dat AI-systemen de menselijke autonomie niet ondermijnen en geen negatieve effecten veroorzaken. Dit kan via verschillende governancemechanismen zoals "human-in-the-loop" (HITL), "human-on-the-loop" (HOTL), en "human-in-command" (HIC). HITL staat voor menselijke interventie in elke besluitcyclus, HOTL voor interventie tijdens de ontwerpcyclus en monitoring, en HIC voor het overzien van de algehele activiteit van het AI-systeem. Toezichtsmechanismen moeten worden afgestemd op het toepassingsgebied en potentiële risico's van het AI-systeem, waarbij strengere governance nodig is naarmate er minder menselijk toezicht mogelijk is.

Sigra

Expertisecentrum Privacy & Informatieveiligheid

Vereisten van betrouwbare AI

2. Technische robuustheid en veiligheid

Technische robuustheid is essentieel voor betrouwbare AI en hangt samen met het principe van schadepreventie. AI-systemen moeten preventief omgaan met risico's en zich betrouwbaar gedragen om onbedoelde schade te minimaliseren en onacceptabele schade te voorkomen. Dit geldt ook bij veranderingen in de omgeving of interacties met andere actoren. Bovendien moet de fysieke en geestelijke integriteit van mensen worden beschermd.

Weerbaarheid tegen aanvallen en beveiliging

AI-systemen moeten worden beschermd tegen kwetsbaarheden en aanvallen, zoals hacking, datavergiftiging en modellekken. Aanvallen kunnen zowel de gegevens als het gedrag van het systeem beïnvloeden, wat kan leiden tot onjuiste beslissingen of fysieke schade. Beveiligingsprocessen moeten mogelijke onbedoelde toepassingen en misbruik door kwaadwillenden voorkomen en beperken om AI-systemen veilig te maken.

Uitwijkplan en algemene veiligheid

AI-systemen moeten noodplannen hebben voor problemen, zoals overschakelen naar een regelsysteem of een menselijke beheerder inschakelen. Ze moeten veilig werken zonder schade aan levende wezens of het milieu te veroorzaken, en onbedoelde gevolgen en fouten moeten worden geminimaliseerd. Risico's moeten worden beoordeeld en verduidelijkt, met veiligheidsmaatregelen afhankelijk van het risico en de capaciteiten van het systeem. Bij hoge risico's moeten proactief veiligheidsmaatregelen worden ontwikkeld en getest.

Nauwkeurigheid

Nauwkeurigheid in AI-systemen verwijst naar hun vermogen om correcte beslissingen, voorspellingen of aanbevelingen te doen. Een goed ontwikkelings- en evaluatieproces kan onbedoelde risico's door onjuiste voorspellingen beperken. Als fouten onvermijdelijk zijn, moet het systeem de kans op dergelijke fouten kunnen aangeven. Hoge nauwkeurigheid is cruciaal in situaties met directe gevolgen voor mensenlevens.

Betrouwbaarheid en reproduceerbaarheid

Betrouwbaarheid en reproduceerbaarheid zijn cruciaal voor AI-systemen. Een betrouwbaar systeem functioneert goed met verschillende soorten input en in diverse situaties, wat helpt om onbedoelde schade te voorkomen. Reproduceerbaarheid houdt in dat een AI-experiment hetzelfde gedrag vertoont onder gelijke omstandigheden, wat wetenschappers en beleidsmakers helpt om nauwkeurig te beschrijven wat AI-systemen doen. Replicatiebestanden kunnen het testen en reproduceren van gedrag vergemakkelijken.

Sigra

Expertisecentrum Privacy & Informatieveiligheid

Vereisten van betrouwbare AI

3. Privacy en datagovernance

Privacy is nauw verbonden met het principe van schadepreventie en is een belangrijk grondrecht dat door AI-systemen beïnvloed wordt. Om privacyschade te voorkomen, is geschikte datagovernance nodig. Dit omvat de kwaliteit en integriteit van gegevens, de relevantie van gegevens binnen het toepassingsgebied van de AI-systemen, toegangsprotocollen, en de capaciteit om gegevens te verwerken op een manier die de privacy beschermt.

Privacy en gegevensbescherming

AI-systemen moeten privacy en gegevensbescherming waarborgen gedurende hun hele levenscyclus. Dit omvat zowel de door de gebruiker aangeleverde informatie als de informatie die tijdens interacties wordt gegenereerd. AI-systemen kunnen voorkeuren en persoonlijke kenmerken van gebruikers afleiden, zoals seksuele geaardheid en politieke standpunten. Om vertrouwen in gegevensverzameling te waarborgen, moeten de verzamelde gegevens niet worden gebruikt voor onwettige of onrechtvaardige discriminatie.

Toegang tot gegevens

Organisaties die persoonsgegevens verwerken, moeten gegevensprotocollen instellen om de toegang tot gegevens te beheren. Deze protocollen moeten specificeren wie onder welke omstandigheden toegang heeft tot de gegevens. Alleen gekwalificeerd personeel met de juiste bevoegdheid en noodzaak mag toegang krijgen.

Kwaliteit en integriteit van gegevens

De kwaliteit en integriteit van gegevens zijn cruciaal voor de prestaties van AI-systemen. Gegevens kunnen vertekeningen, onnauwkeurigheden en fouten bevatten die moeten worden verholpen voordat het systeem wordt getraind. De integriteit van de gegevens moet worden gewaarborgd om te voorkomen dat kwaadwillige gegevens het gedrag van het systeem veranderen. Alle processen en gegevenssets moeten bij elke stap worden getest en gedocumenteerd, ook voor AI-systemen die extern zijn verkregen.

Sigra

Expertisecentrum Privacy & Informatieveiligheid

Vereisten van betrouwbare AI

4. Transparantie

Deze vereiste is nauw verbonden met het principe van verantwoording en omvat de transparantie van relevante elementen voor een AI-systeem, zoals de gegevens, het systeem en de bedrijfsmodellen.

Traceerbaarheid

Traceerbaarheid houdt in dat de gegevenssets, processen en algoritmen die leiden tot beslissingen van een AI-systeem goed gedocumenteerd moeten worden om transparantie te vergroten. Dit geldt ook voor de beslissingen zelf, zodat onjuiste beslissingen kunnen worden geanalyseerd en toekomstige fouten voorkomen. Traceerbaarheid maakt controleerbaarheid en verklaarbaarheid mogelijk.

Communicatie

AI-systemen mogen zich niet voordoen als mensen en gebruikers moeten weten dat ze met een AI-systeem te maken hebben. AI-systemen moeten herkenbaar zijn als zodanig. Waar nodig moet de optie worden geboden om menselijke interactie te kiezen om grondrechten te waarborgen. De capaciteiten en beperkingen van het AI-systeem moeten op een passende manier worden gecommuniceerd aan AI-professionals of eindgebruikers, inclusief informatie over nauwkeurigheid en beperkingen.

Verklaarbaarheid

Verklaarbaarheid betreft het vermogen om zowel de technische processen van een AI-systeem als de gerelateerde menselijke beslissingen te verklaren. Technische verklaarbaarheid vereist dat beslissingen van een AI-systeem door mensen begrepen en getraceerd kunnen worden. Soms moeten afwegingen worden gemaakt tussen verklaarbaarheid en nauwkeurigheid. Bij significante gevolgen voor mensenlevens moet een verklaring van het besluitvormingsproces beschikbaar zijn, afgestemd op de deskundigheid van de belanghebbende. Ook moeten verklaringen beschikbaar zijn over hoe het AI-systeem het besluitvormingsproces, de ontwerpkeuzes en de motivering voor de installatie beïnvloedt, om transparantie van het bedrijfsmodel te waarborgen.

Sigra

Expertisecentrum Privacy & Informatieveiligheid

Vereisten van betrouwbare AI

5. Diversiteit, non-discriminatie en rechtvaardigheid

Om betrouwbare AI te realiseren, moeten inclusie en diversiteit gedurende de hele levenscyclus van het AI-systeem worden bevorderd. Dit houdt in dat alle belanghebbenden betrokken moeten worden en dat er gelijke toegang en behandeling moet zijn via inclusieve ontwerpprocessen. Deze vereiste is nauw verbonden met het principe van rechtvaardigheid.

Voorkomen en onrechtvaardige vertekening

AI-systemen kunnen onbedoelde historische vertekening, onvolledigheid of slechte governance bevatten, wat kan leiden tot vooroordelen en discriminatie. Dit kan onbedoelde schade veroorzaken, zoals exploitatie van vertekening of oneerlijke concurrentie. Vertekening moet zoveel mogelijk worden verwijderd tijdens de gegevensverzameling en ontwikkeling van AI-systemen. Toezichtsprocessen en diversiteit in personeel kunnen helpen om vertekening tegen te gaan en een breed scala aan meningen te waarborgen.

Toegankelijkheid en universeel ontwerp

AI-systemen moeten gebruikers centraal stellen en ontworpen worden voor gebruik door iedereen, ongeacht leeftijd, geslacht, vermogens of eigenschappen. Dit is vooral belangrijk voor mensen met een beperking. AI-systemen moeten principes van universeel ontwerp volgen om een breed scala aan gebruikers te bedienen volgens relevante toegankelijkheidsnormen, zodat iedereen gelijke toegang en actieve participatie heeft.

Participatie van belanghebbenden

Voor de ontwikkeling van betrouwbare AI-systemen is het belangrijk om belanghebbenden te raadplegen die direct of indirect met het systeem te maken hebben gedurende de hele levenscyclus. Regelmatige feedback en langetermijnmechanismen voor participatie van belanghebbenden, zoals werknemers, moeten worden ingesteld om hun betrokkenheid te waarborgen.

Sigra

Expertisecentrum Privacy & Informatieveiligheid

Vereisten van betrouwbare AI

6. Maatschappelijk en milieuwelzijn

AI moet rechtvaardig en zonder schade worden ontwikkeld, met aandacht voor samenleving, gevoelige wezens en milieu. Duurzaamheid en ecologische verantwoordelijkheid moeten worden bevorderd, en onderzoek naar AI-oplossingen voor mondiale problemen gestimuleerd. AI moet ten goede komen aan alle mensen, inclusief toekomstige generaties.

Duurzame en milieuvriendelijke AI

AI-systemen moeten milieuvriendelijk worden ontwikkeld, geïnstalleerd en gebruikt. Dit omvat het controleren van de volledige toeleveringsketen, het kritisch onderzoeken van hulpbronnen en energieverbruik tijdens de training, en het kiezen van minder schadelijke opties. Maatregelen om de milieuvriendelijkheid van de hele toeleveringsketen te waarborgen, moeten worden aangemoedigd.

Sociale gevolgen

Sociale AI-systemen kunnen onze opvatting van sociale controle veranderen en invloed hebben op onze sociale relaties en hechting. Hoewel ze sociale vaardigheden kunnen verbeteren, kunnen ze ook bijdragen aan de verslechtering ervan, wat gevolgen kan hebben voor het fysieke en geestelijke welzijn. Daarom moeten de effecten van deze systemen zorgvuldig worden gemonitord en afgewogen.

Samenleving en democratie

Naast de impact op individuen moeten de maatschappelijke gevolgen van AI-systemen worden beoordeeld, inclusief hun effect op instellingen, democratie en de samenleving als geheel. Het gebruik van AI-systemen moet zorgvuldig worden afgewogen, vooral in verband met het democratische proces en politieke besluitvorming, inclusief verkiezingen.

Sigra

Expertisecentrum Privacy & Informatieveiligheid

Vereisten van betrouwbare AI

7. Verantwoording

De vereiste van verantwoording ondersteunt de andere vereisten en is nauw verbonden met het principe van rechtvaardigheid. Mechanismen moeten worden ingesteld om verantwoordelijkheid en verantwoording voor AI-systemen en hun resultaten te waarborgen, zowel voor als na de toepassing.

Controleerbaarheid

Controleerbaarheid betekent dat algoritmen, gegevens en ontwerpprocessen gecontroleerd kunnen worden, zonder dat bedrijfsmodellen en intellectuele eigendom altijd openbaar moeten zijn. Evaluaties door interne en externe controleurs en de beschikbaarheid van evaluatieverslagen kunnen de betrouwbaarheid van de technologie vergroten. AI-systemen die invloed hebben op grondrechten, inclusief veiligheidskritieke toepassingen, moeten onafhankelijk kunnen worden gecontroleerd.

Afwegingen

Bij het uitvoeren van de vereisten voor AI-systemen kunnen spanningen en onvermijdelijke afwegingen ontstaan. Deze moeten rationeel en methodologisch worden benaderd, waarbij relevante belangen en waarden worden geëvalueerd. Als er geen ethisch acceptabele compromissen kunnen worden gevonden, mag de ontwikkeling of het gebruik van het AI-systeem niet doorgaan. Alle beslissingen over afwegingen moeten goed worden gedocumenteerd en de beslisser moet aansprakelijk zijn en de gepastheid van de beslissingen voortdurend controleren.

Beroep

Er moeten toegankelijke mechanismen zijn om beroep in te stellen bij negatieve effecten van AI-systemen. Dit is essentieel voor het vertrouwen. Er moet speciale aandacht zijn voor kwetsbare personen of groepen.

Minimalisering en verslaglegging van negatieve gevolgen

Het is belangrijk om verslag te doen van handelingen of beslissingen die bijdragen aan de resultaten van een AI-systeem en om op de gevolgen te kunnen reageren. Het vaststellen, beoordelen, rapporteren en minimaliseren van negatieve effecten is cruciaal, vooral voor degenen die de gevolgen ondervinden. Er moet bescherming zijn voor klokkenluiders en andere entiteiten die legitieme zorgen melden. Effectbeoordelingen, zoals red teaming, kunnen helpen om negatieve gevolgen te minimaliseren en moeten in verhouding staan tot het risico van het AI-systeem.

Sigra

Bron: Onafhankelijke Deskundigengroep op Hoog Niveau inzake Artificiële Intelligentie (AI HLEG)

Sigra

Expertisecentrum Privacy & Informatieveiligheid

Meer informatie

Sigra is een regionaal samenwerkingsverband van organisaties in zorg en welzijn in Noord-Holland.

Over het expertisecentrum
Vanuit het Expertisecentrum Privacy & Informatieveiligheid helpen we leden om de privacy en informatieveiligheid in de organisatie goed te organiseren. Je kunt hier terecht voor ondersteuning en advies en om ervaringen met andere Sigra-leden uit te wisselen. Bekijk Sigra.nl voor meer informatie.

Over de campagne
Jaarlijks organiseert Sigra in oktober een Privacy & Informatieveiligheid Bewustwordingscampagne. Bekijk Sigra.nl voor meer informatie.

Contact
Het Expertisecentrum Privacy & Informatieveiligheid is bereikbaar via: pi@sigra.nl.

Sigra