Betekenisvolle Menselijke Tussenkomst

Een hulpmiddel voor het vormgeven en inrichten van betekenisvolle menselijke tussenkomst

Consultatiedocument | Maart 2025

AUTORITEIT
PERSOONSGEGEVENS

Vooraf

Geautomatiseerde besluitvorming wordt in allerlei sectoren toegepast. Voor het vormgeven van algoritmes voor geautomatiseerde individuele besluitvorming en omringende processen zijn een aantal concepten uit de Algemene verordening gegevensbescherming (AVG) en de Richtlijn voor gegevensbescherming bij rechtshandhaving (RGR) cruciaal, waaronder het besluit dat "uitsluitend op een geautomatiseerde verwerking" is gebaseerd.[1] In andere woorden: waarbij geen sprake is van betekenisvolle menselijke tussenkomst.

De European Data Protection Board (EDPB) heeft in de Guidelines inzake geautomatiseerde individuele besluitvorming en profilering voor de toepassing van Verordening (EU) 2016/679 (hierna: guidelines) verduidelijkt wat hiermee wordt bedoeld: "Uitsluitend geautomatiseerde besluitvorming is het nemen van besluiten met technologische middelen en zonder menselijke tussenkomst" en "De verwerkingsverantwoordelijke kan de bepalingen van artikel 22 niet omzeilen door menselijke tussenkomst voor te wenden."[2] Aan die menselijke tussenkomst zitten namelijk eisen: "Om daadwerkelijke menselijke tussenkomst te realiseren moet de verwerkingsverantwoordelijke ervoor zorgen dat al het toezicht op de besluitvorming zinvol is, en niet slechts een symbolische handeling is."[3]

Wat menselijke tussenkomst betekenisvol maakt, is nog niet volledig uitgekristalliseerd. Onderzoekers van de Brussels Privacy Hub bekeken de manieren waarop mensen algoritmes zouden moeten controleren, en zouden moeten ingrijpen wanneer het misgaat. Zij concludeerden dat: "Determining what could be meant precisely by meaningful is indeed an even more complicated -but necessary- task. The scarce precedents in the CJUE and national courts do not make it any easier."[4] Dit document biedt handvatten aan functionarissen gegevensbescherming (FG's), verwerkingsverantwoordelijken en andere betrokkenen om te bepalen wanneer menselijke tussenkomst betekenisvol zou kunnen zijn. De guidelines, wetenschappelijke literatuur, enkele rechterlijke uitspraken en kennis van AP-medewerkers, die zich bezig hebben gehouden met het onderwerp geautomatiseerde besluitvorming, vormen de basis voor dit document. Dit document is daarnaast besproken met andere Europese gegevensbeschermingsautoriteiten. Ook de AI-verordening (Verordening (EU) 2024/1689) biedt duiding. In artikel 14, waarin de eisen aan menselijk toezicht op AI-systemen met een hoog risico zijn geformuleerd, zien we menselijke betrokkenheid terugkomen als bescherming tegen negatieve effecten van algoritmische besluitvorming: "menselijk toezicht is gericht op het voorkomen of beperken van de risico's voor de gezondheid, veiligheid of grondrechten."[5]

Scope

Dit document gaat over betekenisvolle menselijke tussenkomst waardoor geen sprake is van geautomatiseerde besluitvorming zoals bedoeld in artikel 22 lid 1 AVG en artikel 11 lid 1 RGR. Deze artikelen hebben betrekking "uitsluitend op een geautomatiseerde verwerking gebaseerd besluit". Daaronder verstaan we een besluit dat volledig op een geautomatiseerde verwerking is gebaseerd en rechtsgevolgen heeft voor betrokkenen of hen op een andere manier in aanmerkelijke mate treft. Als daarbij betekenisvolle menselijke tussenkomst plaatsvindt, betekent dit dat er geen sprake is van een uitsluitend op een geautomatiseerde verwerking gebaseerd besluit.

Daarnaast is 'menselijke tussenkomst' op grond van artikel 22 lid 3 AVG ook een passende maatregel om de rechten, vrijheden en gerechtvaardigde belangen van de betrokkene te beschermen. Het gaat hier om het geval dat een uitsluitend op geautomatiseerde verwerking gebaseerd besluit wel mag worden genomen, omdat er sprake is van een uitzondering als genoemd in artikel 22 lid 2, onder a of c, AVG. De onderdelen die in dit document worden besproken, zijn ook relevant voor het regelen van menselijke tussenkomst wanneer een betrokkene hier recht op heeft nadat er een geautomatiseerd besluit is genomen. Maar het document is niet met het oog daarop geschreven.

Dit document is bedoeld als een hulpmiddel voor degenen die binnen een organisatie die menselijke tussenkomst vormgeven en inrichten. In dit document heten zij de 'inrichters'. Het is ook voor degenen die dit uitvoeren. Zij heten in dit document de 'beoordelaars'. Dit document is geen afvinklijst: niet alle vragen en onderdelen zullen of kunnen in elk proces van toepassing zijn. De context en individuele omstandigheden van het geval zijn hierbij uiteraard relevant en doorslaggevend.

Algoritmes

De term algoritmes komt in de AVG niet voor. Toch gebruiken we hierna die term. Er wordt in dit document gesproken van algoritmes wanneer het gaat om geautomatiseerde verwerkingen die leiden tot een bepaalde uitkomst in het kader van een te nemen besluit. Het is verder van belang om onderscheid te maken tussen zogenaamde op regels gebaseerde algoritmes en ***machine learning***-algoritmes. Op regels gebaseerde algoritmes volgen een relatief eenvoudige beslisboom (als X, dan Y), formule, of stappenplan. Bij dit type algoritme is betekenisvolle menselijke tussenkomst in het algemeen goed in te richten. Bij een ***machine learning***-algoritme wordt de exacte koppeling tussen input en output door een computer bepaald. ***Machine learning*** is onderdeel van artificiële intelligentie (AI). De complexiteit van ***machine learning*** heeft gevolgen voor de transparantie en uitlegbaarheid van de totstandkoming van de uitkomsten van deze algoritmes. Dat kan een probleem zijn als betrokkenen er gevolgen van ondervinden.[6]

We willen vooraf een paar opmerkingen maken over de inzet van algoritmes. Zo gaan we in dit document verder niet in op de vraag of de inzet van een algoritme in een bepaald proces überhaupt passend is, of dat de gegevens die door een algoritme worden verwerkt, geschikt zijn om door een algoritme te worden beoordeeld. Denk aan bijvoorbeeld een beoordeling van de rijvaardigheid. Daarnaast is de manier waarop een mens tot een besluit komt niet altijd beter of transparant. Dat gezegd hebbende: hoe meer het verantwoord nemen van een besluit een beroep doet op menselijk inzicht, ervaring, maatwerk of intuïtie, hoe minder passend het kan zijn om uitsluitend een algoritme een rol te laten spelen in de beslissing.[7]

Ook vinden we het belangrijk om te waarschuwen voor een tunnelvisie op het proces van menselijke tussenkomst. De vele factoren en vragen die hierboven zijn genoemd, kunnen een groter probleem verbloemen, namelijk dat het besluit dat genomen wordt van zichzelf onethisch is. Of dat de inzet van een algoritme moreel problematisch is. Daarom is het goed om ook na te denken over de aard van het te nemen besluit, los van de vraag of menselijke tussenkomst wel of niet betekenisvol is.

Voorbeeld

De inzet van een algoritme kan het proces van een besluit volledig transformeren. Voorheen zou een ambtenaar met een sociaalmaatschappelijke achtergrond in een gesprek bij mensen thuis een inschatting maken van de hele situatie bij een gezin met financiële problemen. Nu moeten beide partijen het soms doen met een online formulier. Maar een situatie kan te complex zijn voor het formulier; in dat geval is het verstandig dat de beoordelaar de ruimte krijgt om maatwerk te leveren.

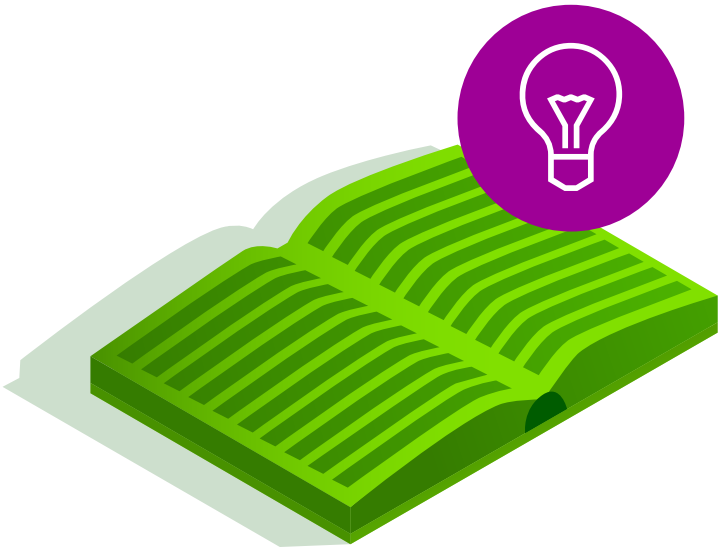
Leeswijzer

De onderdelen die menselijke tussenkomst betekenisvol maken hebben we opgedeeld in vier hoofdstukken: mens, technologie en ontwerp, proces, en governance. Elk onderdeel bestaat uit verschillende subonderdelen. Bij ieder onderdeel staan vragen die kunnen helpen bij het nadenken over de inrichting van menselijke tussenkomst op een betekenisvolle manier.

Consultatie

De Autoriteit Persoonsgegevens (AP) nodigt u uit om te reageren op dit document via een e-mail naar ppa@autoriteitpersoonsgegevens.nl. Dat kan tot en met 6 april 2025. De reacties die we ontvangen, vatten we samen in één document. We noemen daarbij geen namen, organisaties of contactgegevens. De samenvatting publiceren we op de website van de AP en gebruiken we om dit document te verbeteren. Het verbeterde document publiceren we later in 2025.

Alle feedback is welkom. We zijn met name benieuwd naar inzichten uit de praktijk, bijvoorbeeld: Wat maakt het werk makkelijker voor degenen die de menselijke tussenkomst uitvoeren? Wat staat hen juist in de weg? Hoe werken die personen samen met het algoritme?



Inhoudsopgave

1. Mens

Wat mist een algoritme dat een mens wél heeft?

De AVG staat niet toe dat mensen rechtsgevolgen ondervinden of op een andere manier aanmerkelijk getroffen worden door de uitkomst van een algoritme (uitzonderingen daargelaten). Een mens mag dat besluit wel nemen.

Welke menselijke eigenschappen mist een algoritme als het gaat over het nemen van een besluit dat rekening houdt met menselijke waardigheid? En wat kan een verantwoordelijke doen om die eigenschappen tot uiting te laten komen?

Onderdelen

Alle relevante factoren
Volgens de guidelines moeten beoordelaars alle gegevens die relevant zijn voor het besluit in hun analyse betrekken.[8] Dat kan ook om meer gaan dan de gegevens die door het algoritme worden verwerkt. Bijvoorbeeld eventuele aanvullende informatie die de betrokkene kan verstrekken. Menselijke beoordelaars kunnen afwegen of het relevant is om gebruik te maken van dezelfde gegevens die door het algoritme worden gebruikt, of dat beoordelaars rekening moeten houden met andere aanvullende factoren of informatie. Als er geen ruimte is om zo'n afweging te maken, betekent dit dat hun beoordeling mogelijk niet voldoende betekenis heeft en dat de beslissing uiteindelijk toch als 'uitsluitend geautomatiseerd' kan worden beschouwd.

Mensen kunnen ook gegevens meewegen die moeilijk of onmogelijk in een algoritme te vatten zijn. Er wordt een samenwerking tussen mens en algoritme geïmpliceerd in artikel 22 AVG, waarin de menselijke beoordelaars letten op de individuele omstandigheden van de zaak en algoritmes algemene patronen weergeven.[9] Een mens kan daarbij dienstdoen als een bescherming tegen machinematige fouten en ***digital rigidity:***[10] de onverschilligheid van algoritmes tegenover bepaalde typen relevante informatie en het platslaan van (complexe) situaties in een spreadsheet. Algoritmes kunnen bijvoorbeeld door een fout geplaatste komma een uitzonderlijk hoog bedrag signaleren als indicatie voor fraude, terwijl een menselijke beoordelaar sneller ziet dat het om een typefout gaat. Ook bij bepaalde uitzonderingen kunnen algoritmes verkeerde conclusies trekken, zoals een leeg adresveld vanwege een beschermd woonadres van een betrokkene.

Het is wenselijk dat de verantwoordelijke samen met de beoordelaars nadenkt over de aspecten die het algoritme in overweging moet nemen. En over de - eventuele aanvullende - aspecten waar de menselijke beoordelaars rekening mee kunnen houden voor zij tot hun uiteindelijke besluit komen. Bepaalde factoren zijn meer geschikt om door een algoritme beoordeeld te worden, zoals een verplicht minimaal aantal jaren werkervaring van een sollicitant. Andere factoren, zoals de kwaliteit van het schrijven in een sollicitatiebrief, vragen over het algemeen eerder een menselijke blik. Een menselijke beoordelaar zal uit sommige data ook meer informatie kunnen halen dan een algoritme dat kan.

Voorbeeld: klachtenafdeling

Bij de klachtenafdeling van een bezorgdienst wordt elke klacht door een mens en een algoritme gezamenlijk beoordeeld. De menselijke beoordelaar en het algoritme zien in eerste instantie beiden dezelfde klacht van een klant over een bezorger. Het algoritme selecteert op basis van het cijfer dat de klant aan de bezorger heeft gegeven of deze klacht nader onderzocht zou kunnen worden. De beoordelaar kijkt of de melding specifiek genoeg en inhoudelijk substantieel is, omdat een vage melding een signaal kan zijn van een klacht die alleen is ingediend om de bezorger te 'pesten'. Pas daarna wordt een besluit over de klacht gemaakt.

Dit alles wil niet zeggen dat het altijd beter is om een mens (meer) variabelen in overweging te laten nemen. Menselijke beoordelaars kijken met hun eigen vooroordelen naar een besluit, en het weghalen van deze (voor)oordelen kan in sommige gevallen juist een argument zijn voor het inzetten van een algoritme.

De factoren die de verantwoordelijke bepaalt zijn niet uitputtend. Het is wenselijk dat beoordelaars de ruimte krijgen om andere factoren die zij relevant vinden toe te voegen. Zo kan het soms nodig zijn om contact op te nemen met een betrokkene om inzicht te krijgen in een afwijkende variabele, zoals een verkeerd ingevuld veld, of informatie die met een goede reden ontbreekt, zoals een beschermd woonadres.

Voorbeeld: magazijnmedewerker

Een algoritme merkt een werknemer in een magazijn aan als onderpresteerder omdat deze werknemer later dan zijn collega's in heeft geklokt. De beoordelaar zou misschien al snel opmerken dat de medewerker die dag naar de tandarts moest.

De ruimte die de beoordelaar krijgt hoeft niet onbeperkt te zijn. In sommige gevallen is het bijvoorbeeld niet wenselijk dat een beoordelaar informatie toevoegt. Naast het algemene uitgangspunt van dataminimalisatie, gelden ook specifieke beperkingen voor bijzondere persoonsgegevens. De betrokkene kan het ook als negatief ervaren als deze wordt gevraagd om meer informatie. Het kan zelfs zo zijn dat zo'n nader onderzoek een betrokkene in aanmerkelijke mate treft. De training van beoordelaars kan aandacht besteden aan deze negatieve aspecten van het opvragen van informatie: de impact op de betrokkene, alternatieven voor contact opnemen, situaties waarin extra onderzoek nuttig is.

Vragen om bij de inrichting te stellen:

- → **Betrekken beoordelaars alle relevante informatie in hun beoordeling?**
  - Kunnen beoordelaars specifieke omstandigheden meenemen in hun beoordeling die een algoritme niet meeweegt?
- → **Hebben beoordelaars meer informatie tot hun beschikking dan het algoritme heeft?**
  - Zo niet, kunnen beoordelaars daar toegang toe krijgen?
- → **Op basis van welke informatie dienen beoordelaars een besluit te beoordelen, of tegen het algoritme in te gaan?**
- → **Op welke manier(en) kunnen beoordelaars tegen het algoritme ingaan?**
- → **Kunnen beoordelaars gegevens in het algoritme naast zich neerleggen, aanvullen of corrigeren?**
  - Bijvoorbeeld wanneer er informatie ontbreekt, of het duidelijk is dat iets fout is geformuleerd.
- → **Is het duidelijk wat van beoordelaars wordt verwacht?**

Menselijk inzicht

Wat maakt menselijke tussenkomst ***menselijk***? Deze vraag raakt aan de essentie van menselijke tussenkomt. En is tegelijkertijd misschien wel de vraag met het minst duidelijke antwoord. Naast het feit dat een mens en een algoritme verschillende gegevens kunnen beoordelen, beoordelen mensen en algoritmes gegevens verschillend. De menselijke manier is minder tastbaar dan de algoritmische manier. In de literatuur worden verschillende pogingen gedaan om die te formuleren: ethiek, een menselijke maat, empathie, het vatten van onvoorspelbare factoren, (emotionele) intelligentie, moreel vermogen, geweten, interpretatie, ervaring, creativiteit of niet-lineaire redenatie. Dit idee wordt het vaakst omschreven als '***discretion***' (in dit document vertaald als 'menselijk inzicht').[11] In het proces is dan ruimte voor dat menselijke inzicht, om zo de menselijke waardigheid te beschermen.

Het menselijk inzicht hoeft niet ongelimiteerd te zijn. We kunnen de manier waarop mensen tot hun besluit komen minder hard maken dan de manier waarop algoritmes tot een uitkomst komen. Bij mensen is het soms niet meer dan een gevoel. En bij algoritmes is het een precies uitziende numerieke score. Daarom kan het moeilijk zijn om menselijk inzicht als basis te gebruiken voor het verwerpen van de uitkomst van een algoritme. Dit geldt zeker voor keuzes die beter geschikt zijn voor een menselijk besluit dan voor gevolgen op basis van de uitkomst van een algoritme.

Voorbeeld: sollicitatie

Er zijn algoritmes in ontwikkeling die de leiderschapskwaliteiten van een sollicitant zouden kunnen afleiden uit diens schrijfstijl. Deze algoritmes kunnen de sollicitant daarvoor een score geven. Een beoordelaar die na een gesprek met de kandidaat meent dat deze onvoldoende leiderschapskwaliteiten heeft op basis van ervaring, raakt misschien aan het twijfelen wanneer een algoritme de kandidaat met een 9 heeft beoordeeld op basis van de sollicitatiebrief. Het oordeel van de beoordelaar is minder 'hard' te maken dan het cijfer van het algoritme.

Soms gaat nuance verloren bij de ontwikkeling van een algoritme. De beoordelaar kan deze nuance terugbrengen in het besluit. Bij de vertaling van beleid voor een bepaald besluit naar de code voor het algoritme, maken de ontwikkelaars en/of de verantwoordelijken keuzes: welke factoren wegen we wel en niet mee? Met welke individuele situaties en uitzonderingen houden we bij voorbaat al rekening? De beoordelaar kan waar nodig afwijken van de rigide blik van het algoritme op de situatie van een betrokkene.

In het besluit wegen verschillende belangen mee. Het is belangrijk dat de beoordelaar de belangen van de betrokkene meeweegt. Belangen zoals risico's voor de gezondheid, veiligheid en grondrechten uit de AI-verordening. Vaak wegen ook organisatorische belangen mee bij het toepassen van een algoritme. Dit zou het proces namelijk sneller, efficiënter, en daarmee goedkoper maken. Soms heeft zo'n belang verregaande invloed op het besluit en staat die in de weg van een betekenisvolle, menselijke tussenkomst.

Voorbeeld: verzekeringsclaim

Een tussenpersoon beoordeelt namens verzekeraars of een claim aan een verzekerde moet worden uitbetaald. Het is voor de verzekeraar financieel gunstiger als er minder claims worden uitbetaald. De tussenpersoon geeft aan de ontwerpers van het algoritme en aan de beoordelaars mee dat besluiten niet te vaak in het voordeel van de verzekerde mogen uitvallen.

Het is belangrijk om hierbij opnieuw te noemen dat het toevoegen van menselijk inzicht in het proces ook nadelige gevolgen kan hebben. Mensen kunnen vooroordelen hebben die een algoritme niet heeft. Of kunnen mogelijk minder goed zijn in het beoordelen van bepaalde variabelen die een algoritme wél goed kan beoordelen. Mensen kunnen ook minder goed zijn in het tegelijkertijd beoordelen van alle factoren die relevant zijn voor een besluit.

Vragen om bij de inrichting te stellen:

- → **Passen beoordelaars bij elke casus menselijk inzicht toe door bijvoorbeeld op hun eigen manier naar aspecten van het te nemen besluit te kijken?**
- → **In welke mate vereist de beslissing een inschatting van individuele factoren?**
- → **Welke eisen worden er aan beoordelaars gesteld?**
- → **Welke andere belangen dan die van de betrokkene weegt de beoordelaar mee?**
- → **Om te testen of er genoeg ruimte is voor het menselijk inzicht kan een beoordelaar gevraagd worden om enkele besluiten te nemen voordat een algoritme dat doet, of voordat een algoritme hiervoor input heeft geleverd. Vervolgens worden de besluiten van de beoordelaar vergeleken met de output van het algoritme. Dit wordt *cognitive forcing* genoemd.**

Bekwaamheid

De beoordelaar wordt geacht bekwaam te zijn.[12] Dat houdt in dat diegene beschikt over de kennis en vaardigheden die nodig zijn om het besluit te nemen. Daarnaast heeft diegene kennis over het algoritme. Een beoordelaar begrijpt in ieder geval op hoofdlijnen hoe een algoritme tot een resultaat komt. Een beoordelaar die geen inzicht heeft in het algoritme zou bij het afwegen van een resultaat in feite het hele proces opnieuw doorlopen. De verwerkingsverantwoordelijke zorgt voor informatie en training, zodat de beoordelaar voldoende inzicht heeft in het algoritme.

Het kan voorkomen dat de vaardigheden voor het beoordelen van het resultaat van een algoritme verschillen van de vaardigheden voor het nemen van een het besluit (zonder algoritme). Het beoordelen van een algoritme is soms ingewikkelder dan het nemen van het besluit zelf. Het besluit zelf kan ook zulke specialistische kennis vereisen, bijvoorbeeld in een medische context. In sommige gevallen kan de beoordelaar zich dan niet ook nog voldoende specialiseren in de technische details van het algoritme. Dan kan het nodig zijn om een ***team-in-the-loop*** in te stellen in plaats van een enkel persoon. De beoordelaars, die inhoudelijk het besluit nemen over de betrokkenen, werken dan samen met collega's die het algoritme kunnen beoordelen.

Vragen om bij de inrichting te stellen:

- → **Begrijpen beoordelaars hoe en op basis van welke gegevens een algoritme tot een resultaat komt?**
- → **Beschikken beoordelaars over basiskennis rondom statistiek?**
- → **Weten beoordelaars met welke factoren het algoritme rekening houdt?**
  - Kunnen beoordelaars de data die het algoritme heeft gebruikt inzien?
- → **Zouden beoordelaars het besluit ook zonder algoritme kunnen maken?**
- → **Kunnen beoordelaars het besluit alleen maken, of is er een team nodig waarin verschillende specialisaties vertegenwoordigd zijn?**
  - Hoe werkt dat team samen?
- → **Is er voldoende inzicht in het algoritme dat wordt gebruikt?**
- → **Is het duidelijk wat van een beoordelaar verwacht wordt als het gaat om menselijkheid en organisatorisch belang?**

2. Technologie en ontwerp

Hoe beïnvloedt het algoritme de menselijke beoordelaar?

Bij menselijke tussenkomst is het van belang om naar de technologie te kijken. Technologie is nooit neutraal en kan invloed hebben op de mate waarin menselijke tussenkomst betekenisvol is. De vormgeving van de interface van een algoritme en de data waarop een uitkomst is gebaseerd kan bijvoorbeeld invloed hebben op beoordelaars. Onthoud wel dat dit menselijke keuzes zijn. De mens ontwerpt immers de technologie. In het algemeen geldt dat dat hoe meer een mens zich aanpast aan een algoritme, hoe meer geautomatiseerd is. Door het algoritme kan een beoordelaar bijvoorbeeld beperkt worden in handelingsopties (zoals een binaire ja/nee-optie), terwijl die beoordelaar vóór de komst van het algoritme misschien tot een maatwerkoplossing was gekomen. Als het algoritme voorschrijft hoe een beoordelaar moet handelen, kan dit dus ten koste gaan van diens autonomie. Menselijke tussenkomst kan meer betekenis krijgen door passende technische maatregelen.

Relevante concepten

Automation bias

Automation bias verwijst naar het idee dat mensen het prestatievermogen en de nauwkeurigheid van algoritmes vaak overschatten. We hebben te veel vertrouwen in de werking van algoritmes, zelfs wanneer deze fouten maken. Met andere woorden: mensen nemen de output van algoritmes vaak snel als waarheid aan. Dit kan ertoe leiden dat mensen eigen kennis of waarnemingen negeren. Uit Brits onderzoek bleek bijvoorbeeld dat politieagenten in Londen de betrouwbaarheid van real-time gezichtsherkenningstechnologie enorm overschatten. Zij oordeelden drie keer zo vaak dat het algoritme goed zat dan daadwerkelijk het geval was.[13] Algoritmes worden vaak geframed als betrouwbaarder of secuurder dan mensen.

Deze framing kan bijdragen aan ***automation bias***. Het is dus interessant om na te gaan hoe er binnen een organisatie over een algoritme wordt gesproken. Wordt deze bijvoorbeeld als heel betrouwbaar beschreven en als cruciaal onderdeel in een proces? Of wordt wellicht benadrukt dat het algoritme slechts een ondersteunende rol heeft? Het is belangrijk dat beoordelaars bekend zijn met ***automation bias***, zodat zij dit herkennen. Hier kan tijdens trainingen aandacht aan worden besteed.

Algorithmic aversion

Anderzijds kunnen mensen het prestatievermogen van een algoritme onderschatten, zelfs wanneer bekend is dat het algoritme nauwkeuriger is. Dit wordt ook wel ***algorithmic aversion*** genoemd. Dit gevoel treedt vaak op wanneer algoritmes keuzes over mensen maken. Denk bijvoorbeeld aan het verstrekken van een visum of het toekennen van een lening. Door ***algorithmic aversion*** kunnen verschillende problemen ontstaan, zoals de menselijke bias.

Deze twee concepten zijn relevant voor betekenisvolle menselijke tussenkomst. ***Automation bias*** en ***algorithmic aversion*** laten zien dat de toevoeging van een mens niet altijd tot wenselijke resultaten leidt. Enerzijds kunnen mensen geneigd zijn om de output van algoritmes snel als waarheid aan te nemen, wat de menselijke tussenkomst minder betekenisvol maakt. Anderzijds kunnen mensen onterecht minder vertrouwen hebben in algoritmes. Deze kritische houding resulteert mogelijk in onbenutte mogelijkheden die een verantwoord ingericht algoritme wel zou kunnen bieden.

Onderdelen

Interface

Ontwerp kan invloed hebben op ons gedrag. Elementen van een object kunnen bepaald gedrag stimuleren, of juist hinderen. Hetzelfde geldt voor interfaces van computers. Denk aan het gebruik van bepaalde kleuren, bijvoorbeeld bij pop-up vensters. Het ontwerp van de interface van een algoritme kan dus van invloed zijn op beoordelaars. Het liefst wordt een interface ontworpen met de uiteindelijke gebruikers in gedachten. Dit klinkt logisch, maar het gaat in de praktijk niet altijd goed.

Voorbeeld: vliegtuigradar

De USS Vincennes, een oorlogsschip van de Amerikaanse marine, schoot in 1988 een Iraans passagiersvliegtuig neer als gevolg van slecht interface ontwerp. Verkeersleiders dachten dat het vliegtuig richting het schip vloog, terwijl het in werkelijkheid van het schip wegvloog. De richting van het vliegtuig was niet duidelijk op het scherm te zien. Ook liet het scherm niet zien wat de snelheid van het vliegtuig was. Daardoor moesten medewerkers gegevens handmatig vergelijken en berekeningen in hun hoofd, op kladblokjes of op een rekenmachine maken.[14]

Ook kan een interface de neutraliteit beïnvloeden. Zo kan kleurgebruik bepaalde associaties oproepen. Denk bijvoorbeeld aan een rode risicoscore in de interface van een frauderisicoalgoritme. Een rood signaal kan het idee geven dat iemand daadwerkelijk een fraudeur is, terwijl dit misschien niet het geval is. Een neutralere score geeft een beoordelaar meer kans om zelf een objectief besluit te nemen.

Mensen communiceren op een andere manier dan computers. Goed ontworpen interfaces zorgen ervoor dat de communicatie tussen mens en computer beter verloopt. Hier kan in het ontwerp van interfaces rekening mee worden gehouden door bijvoorbeeld toelichting bij bepaalde data te geven. In de interface zou kunnen worden aangegeven hoe, en op basis waarvan, een risicoscore tot stand is gekomen. Ook kunnen elementen als kleur, lettertype of pop-ups helpen bij het begrijpelijk maken van een interface.[15] Ontwerpelementen kunnen dus worden ingezet om betekenisvolle tussenkomst te stimuleren. Een beoordelaar kan door de interface bijvoorbeeld aangespoord worden om bepaalde inputdata te controleren.

Hoeveelheid data

De hoeveelheid data die een beoordelaar te zien krijgt bij de uitkomst van een algoritme is erg belangrijk voor het proces en voor de toegevoegde waarde van menselijke tussenkomst. Je kunt je voorstellen dat het moeilijk is om tot een juist besluit te komen wanneer je over (te) weinig informatie beschikt. Tegelijkertijd kan een overvloed aan data ervoor zorgen dat het ingewikkeld wordt om tot een goed besluit te komen. Veel data kunnen immers overweldigend zijn. Algoritmes komen vaak tot een bepaalde output op basis van honderden of duizenden datapunten. Dit is voor een mens moeilijk te overzien. Dus welke data krijgt een beoordelaar dan wél te zien? Hier dient een organisatie goed over na te denken.

Data in context

Daarnaast heeft data zonder context weinig betekenis voor een mens. Zo zegt een zuurstofmeter in een kantoorruimte ons weinig als we niet weten wat het gewenste zuurstofgehalte is. Bij sommige processen worden mensenlevens gereduceerd tot data en hebben de uitkomsten van algoritmes mogelijk ingrijpende gevolgen. Het is dus belangrijk dat data in de juiste context wordt geplaatst, zodat de beoordelaar tot een goed overwogen besluit kan komen. Je zou kunnen stellen: hoe abstracter de data is, hoe minder betekenisvol de menselijke tussenkomst zal zijn. Abstracte data zijn getallen zonder duidelijke uitleg, zoals een risicoscore van 0.5. Zonder context is niet duidelijk waar deze risicoscore op is gebaseerd en hoe deze zich verhoudt tot andere scores.

Vragen om bij de inrichting te stellen:

→ **Wat is er op het interface te zien wanneer een beoordelaar de uitkomst van een algoritme beoordeelt?** [16]
- In hoeverre is het interface overzichtelijk en begrijpelijk ontworpen?
  - *Zijn beoordelaars betrokken in de ontwerpfase?*
  - *Maakt het interface de beslissing overzichtelijker, bijvoorbeeld met duiding bij cijfers en grafieken of een betrouwbaarheidsscore bij het resultaat?*
  - *Zijn er bepaalde designelementen die invloed kunnen hebben op de neutraliteit van beoordelaars?*
- In hoeverre begrijpen beoordelaars wat er op de interface te zien is?
- Hoe wordt data aan beoordelaars gepresenteerd?
  - *Wordt data op een duidelijke en begrijpelijke wijze gepresenteerd?*
  - *Is duidelijk wat er met bepaalde data wordt bedoeld?*
  - *Welke data hebben beoordelaars nodig om tot een besluit te komen?*

Volgorde van data

Ook is het van belang welke data een beoordelaar als eerste te zien krijgt. Vaak vormt de informatie die een persoon als eerste te zien krijgt het uitgangspunt voor latere beslissingen. Ons brein hecht zich namelijk aan een bepaald referentiepunt, ongeacht wat dat referentiepunt is. Dit wordt ook wel ***anchoring*** genoemd. Door bepaalde informatie als eerste te tonen, kan een vervolgbeslissing dus gestuurd worden. In de praktijk gebeurt dit bijvoorbeeld wanneer iemand als 'potentieel fraudeur of risico' wordt gemarkeerd. Dit zal invloed hebben op het oordeel van de beoordelaar.

Over het algemeen besteden mensen meer aandacht aan factoren die benadrukt worden. Een beoordelaar kan dus (te veel) beïnvloed worden door de presentatie van data of de uitkomst van een algoritme. Zo is aangetoond dat beoordelaars iets meer als 'risico' inschatten als een door het algoritme gemaakte risicoscore onderdeel is van een uitkomst.[17]

Welke data een beoordelaar als eerste te zien krijgt, beïnvloedt dus het oordeel, maar ook de volgorde waarop data gepresenteerd worden, heeft daar invloed op. Zo kan het sorteren van data op alfabetische volgorde ongewenste effecten hebben. Wanneer personen bijvoorbeeld alfabetisch worden gesorteerd op woonplaats, kan het voorkomen dat iemand uit Amsterdam een grotere kans heeft nader gecontroleerd te worden dan iemand uit Zoetermeer.

Tot slot kan voor een besluit nodig zijn dat een beoordelaar nagaat of de informatie die het algoritme heeft gebruikt juist is. Dit geldt vooral wanneer het om uitzonderlijke data gaat. Zo kan er in een interface een optie zijn om data te controleren, corrigeren, aanvullende data op te vragen of toe te voegen.

Vragen om bij de inrichting te stellen:

→ **Welke data krijgen beoordelaars te zien tijdens het beoordelen van een uitkomst?**
- Wat voor effect heeft dit?

→ **Hoeveel data krijgen beoordelaars te zien tijdens het maken van een besluit?**
- Krijgen beoordelaars voldoende data te zien om tot een weloverwogen beslissing te komen?
- Ontvangen beoordelaars voor elke besluit dezelfde hoeveelheid data?
  - *Zo niet, wat gebeurt er wanneer een beoordelaar over (te) weinig data beschikt?*
  - *Is er in de interface bijvoorbeeld een optie om aanvullende data op te vragen of toe te voegen?*
  - *Is er in de interface ook een optie om data te corrigeren?*
- Worden data in een bepaalde volgorde gepresenteerd?
- Welk mogelijk effect heeft dit?

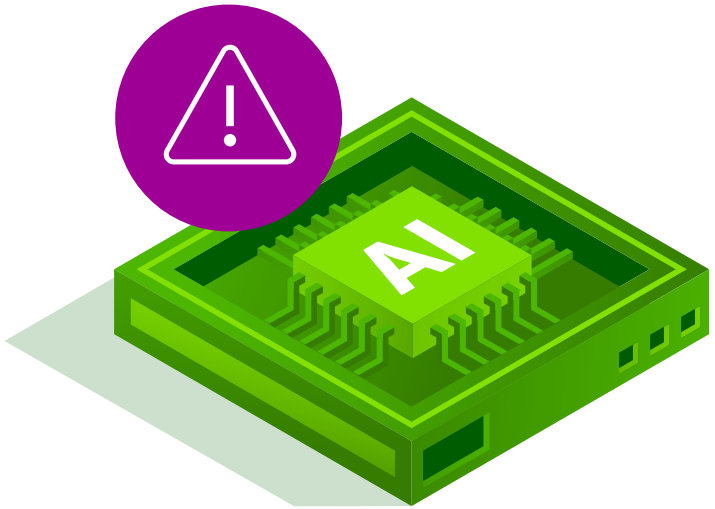
Routinematigheid

Als het werk van een beoordelaar routinematig wordt, zonder het resultaat daadwerkelijk te beïnvloeden, verliest het zijn functie.[18] In het ontwerp van het algoritme zou hier rekening mee gehouden kunnen worden. Zo kan een uitkomst op verschillende manieren worden voorgelegd en er kunnen controlevragen worden ingebouwd. Er kan ook variatie zijn in de presentatie van output. Zo kan een algoritme een risicoscore aan mensen toekennen, maar er kan ook toegelicht worden waarom een individu een risico vormt, zonder hier een risicoscore aan te koppelen. Denk hierbij bijvoorbeeld aan 'risico op diefstal', 'risico op inbraak' en 'risico op vandalisme', in plaats van een algemene risicoscore. Deze variatie in output vergt meer denkwerk vanuit de beoordelaar om tot een weloverwogen besluit te komen. Dat kan leiden tot betekenisvolle menselijke tussenkomst. Belangrijk is dan wel dat de inrichter zorgt voor gelijke behandeling van betrokkenen. Let op: houd er rekening mee dat een verschillende presentatie van een beslissing ervoor kan zorgen dat de beslissing zelf verandert.

Ook kan de beoordelaar regelmatig een besluit voorgelegd krijgen dat helemaal zonder algoritme moet worden genomen. Tot slot kunnen aan de betrokkenen, geselecteerd door het algoritme voor bijvoorbeeld nadere controle, steekproefsgewijs betrokkenen worden toegevoegd die níet zijn geselecteerd. Dit zorgt ervoor dat de beoordelaar niet blind kan vertrouwen op de algoritmische output.[19]

Vragen om bij de inrichting te stellen:

- → **Is er in het ontwerp van het algoritme nagedacht over het voorkomen van routinematigheid bij beoordelaars?**
  - Hoe wordt de uitkomst voorgelegd?
  - Zijn er controle vragen?
  - Worden data of de output in een verschillende volgorde gepresenteerd? Of varieert de volgorde waarop een deelbesluit wordt genomen?
- → **Komen alle resultaten die beoordelaars te zien krijgen vanuit het algoritme?**
- → **Moeten beoordelaars besluiten vaak handmatig maken?**
- → **Worden beoordelaars alert gehouden?**
- → **Wordt er op fouten gecontroleerd?**



3. Proces

Hoe hebben de keuzes van de organisatie invloed op de menselijke beoordelaar?

Bij het beoordelen van de betekenis van menselijke tussenkomst is het van belang om te kijken naar het proces rondom de menselijke tussenkomst. In het beste geval heeft een organisatie een duidelijk beschreven proces voor betekenisvolle menselijke tussenkomst en kunnen betrokken medewerkers van de organisatie dit goed uitleggen. Zo is er bijvoorbeeld nagedacht over ***hoe*** een beoordelaar de uitkomst van een algoritme hoort te beoordelen. Ook zijn er bepaalde eisen aan de betekenisvolle menselijke tussenkomst gesteld en is goed gemotiveerd op basis van welke data een beoordelaar een besluit hoort te nemen.

Voorbeeld: contentmoderatie

Onderzoekers bekeken werknemers die content op een groot sociaal netwerk modereerden, nadat een algoritme de content had gemarkeerd als mogelijk problematisch.[20] De werknemers "werken vaak in zeer moeilijke omstandigheden met een laag loon en krijgen meestal maar een paar seconden om over elk stukje inhoud te beslissen. Omdat de werknemers zeer slecht betaald worden, trekken deze banen geen geschoolde werknemers aan en al helemaal geen werknemers met enige juridische of technologische kwalificaties" (vertaald). Het onderzoek concludeert: "Hoewel dit zou kunnen duiden op menselijke invloed, zijn mensen zo diep ingebed in de algoritmische systemen dat het moeilijk is om hen daadwerkelijke besluitvormers te noemen. De analyse lijkt er eerder op te wijzen dat de algoritmische systemen eigenlijk zijn ingebed om een beperkte mate van menselijke invloed binnen een zeer complex systeem te suggereren. Hoewel menselijk werk op sommige gebieden noodzakelijk kan zijn, lijkt menselijke autonomie een deel van de reden voor menselijke betrokkenheid. In werkelijkheid is die menselijke autonomie er heel weinig" (vertaald).

Onderdelen

Timing

Menselijke tussenkomst kan voorafgaand aan de definitieve besluitvorming op verschillende momenten in het proces plaatsvinden. Grofweg kan menselijke tussenkomst op de volgende momenten plaatsvinden:[21]

a. Een algoritme levert de benodigde informatie voor een besluit, maar de beoordelaar neemt het besluit. De timing van de menselijke tussenkomst is hier dus in het einde van het proces, maar nog voordat het uiteindelijke besluit genomen wordt.

b. Een algoritme levert informatie voor een (deel)aspect van het besluit of wordt gebruikt om verbanden tussen data te beschrijven, een diagnose te stellen, gebeurtenissen te voorspellen. De beoordelaar neemt het besluit. De inzet van dit algoritme levert dus informatie op die na nader onderzoek kan leiden tot een besluit. In dit geval vindt menselijke tussenkomst al eerder in het besluitvormingsproces plaats.

c. Menselijke tussenkomst vindt op verschillende momenten in het proces plaats.

Je kunt je voorstellen dat menselijke tussenkomst minder betekenisvol is wanneer er alleen een controle op de output van een algoritme plaatsvindt. Vooral wanneer niet duidelijk is hoe het algoritme tot deze output is gekomen. De beslissing lijkt dan in feite al genomen. In dit soort gevallen heeft het algoritme eigenlijk een doorslaggevende rol. Maar ook dan kan een grondige analyse door een mens, voordat het besluit wordt genomen, alsnog betekenisvol zijn.

Voor betekenisvolle menselijke tussenkomst moet een beoordelaar invloed kunnen uitoefenen op de uitkomst van een algoritme. De mate van invloed hangt deels af van waar in het proces menselijke tussenkomst plaatsvindt.

Een algoritme dat een scholier labelt als 'achterblijver' verschilt in timing van een algoritme dat verschillende datapunten aanlevert, waar de beoordelaar zelf een conclusie uit trekt. Bij het eerste voorbeeld vindt menselijke tussenkomst achteraf plaats. Bij het tweede voorbeeld vindt menselijke tussenkomst eerder in het proces plaats.

Hier is sprake van een ondersteunend algoritme. Dit verschil kan ook blijken uit de reden dat het algoritme wordt gebruikt. Wordt er werk uit handen genomen van experts voor meer efficiëntie? Of wordt de expertise juist versterkt, zoals het geval is bij diagnostische tools?

Menselijke tussenkomst kan ook op meerdere momenten in het proces plaatsvinden. De input van een algoritme wordt dan bijvoorbeeld eerst door een beoordelaar gecontroleerd. Het algoritme levert vervolgens informatie voor deelaspecten van het besluit. De beoordelaar beoordeelt deze en de uiteindelijke output wordt ook door een beoordelaar getoetst. Aangezien dit verschillende mensen dit vaak doen, spreken we hier ook wel van team-in-the-loop.

Vragen om bij de inrichting te stellen:

- → **Op welk moment in het besluitvormingsproces vindt menselijke tussenkomst plaats?**
  - Levert het algoritme een volledige uitkomst aan?
  - Of levert een algoritme informatie over een (deel)aspect van het te nemen besluit?
  - Moeten beoordelaars het besluit van een algoritme te controleren?
- → **Vindt er menselijke tussenkomst op verschillende momenten in het proces plaats?**
- → **Welke rol speelt het resultaat van het algoritme in het besluit?**
- → **Wat is de rol van het algoritme in het proces (bepalend, ondersteunend, controlerend)?**

Werkdruk

Beoordelaars nemen vaak beslissingen over mensen met mogelijk ingrijpende gevolgen. Het is belangrijk dat beoordelaars hier genoeg tijd voor hebben. Hoe minder tijd een beoordelaar heeft om tot een besluit te komen, hoe minder betekenisvol de menselijke tussenkomst zal zijn. Hierbij is het van belang om te onthouden dat er geen ideale

tijdseenheid is. Dit is afhankelijk van de context. In het beste geval wordt het gehele proces nagelopen, om zo te controleren of beoordelaars genoeg tijd hebben voor hun besluit.

Een organisatiecultuur die de nadruk legt op efficiëntie kan de menselijke tussenkomst negatief beïnvloeden. Bijvoorbeeld als er een bepaald aantal besluiten per dag beoordeeld moet worden. Het kan zelfs voorkomen dat organisaties met targets werken, denk aan commerciële partijen. Dit kan ervoor zorgen dat beoordelaars niet genoeg tijd krijgen of nemen om tot een weloverwogen besluit te komen.

Vragen om bij de inrichting te stellen:

- → **Hoeveel tijd hebben beoordelaars doorgaans voor het beoordelen van de uitkomst van een algoritme?**[22]
  - Hoe verhoudt dit zich tot de aard van het te nemen besluit?[23]
- → **Hoeveel beoordelingen maken beoordelaars gemiddeld per dag?**
- → **Is er een minimaal aantal besluiten dat beoordelaars per dag moeten beoordelen?**

Bevoegdheid

Wanneer een beoordelaar een uitkomst van een algoritme beoordeelt, is het belangrijk dat deze tegen deze uitkomst van een algoritme in ***mag*** gaan. Daarnaast is het van belang dat een beoordelaar dit ook daadwerkelijk doet. Een beoordelaar kan formeel wel bevoegd zijn om tegen het algoritme in te gaan, maar kan in de praktijk obstakels ondervinden. De obstakels kunnen menselijke tussenkomst minder betekenisvol maken.

Voorbeeld: monitoring op werk

Een organisatie zet een monitoringsalgoritme in om de prestaties van werknemers in de gaten te houden. Een manager krijgt een mail van het monitoringsalgoritme dat één van haar werknemers slecht zou presteren. De manager ziet meteen dat de melding aan de verkeerde medewerker is gekoppeld, maar heeft nooit instructie gekregen om wijzigingen in het systeem aan te brengen.

Het is goed om te controleren of een beoordelaar zich vrij voelt om tegen het algoritme in te gaan. Organisatiecultuur kan hierin ook een rol spelen. Wanneer er bijvoorbeeld een sterke hiërarchische organisatiecultuur heerst, kunnen beoordelaars zich niet vrij genoeg voelen om tegen een algoritme in te gaan. Of wanneer beoordelaars streng worden afgestraft op fouten, kunnen beoordelaars terughoudend zijn om tegen het algoritme in te gaan.

Ook de gevolgen die een beoordelaar ondervindt van eigen fouten of fouten van het algoritme wegen mee. Als een menselijke beoordelaar verantwoordelijk wordt gehouden voor het nemen van een onjuiste beslissing, kan dit ervoor zorgen dat deze minder snel tegen een algoritme ingaat. Het is daarom wenselijk dat een beoordelaar enerzijds bevoegd is om tegen het algoritme in te gaan, maar anderzijds niet wordt bestraft bij een fout. Dat betekent niet dat een beoordelaar geen gevolgen mag ondervinden voor een verkeerde beslissing. De verantwoordelijke moet de kwaliteit van de menselijke tussenkomst wel kunnen beoordelen (zie ook de paragraaf over monitoring).

Voorbeeld: uitkeringen

Bij het Poolse equivalent van de UWV ontvingen medewerkers een handboek bij de implementatie van een algoritme. In het handboek stond dat elke wijziging in het systeem (bijvoorbeeld door het corrigeren van data) werd vastgelegd.[24] Uit angst om afgestraft te worden voor fouten, gingen beoordelaars maar in enkele gevallen tegen het algoritme in.

Vragen om bij de inrichting te stellen:

- → **Zijn beoordelaars bevoegd om een uitkomst van het algoritme te overrulen?**
  - En hoe is dit proces ingericht? Is dit vastgelegd in beleid?
- → **Hoe vaak gaan beoordelaars tegen een uitkomst van een algoritme in?**
- → **Ervaren beoordelaars een drempel om tegen het algoritme in te gaan?**
  - Ondervinden beoordelaars negatieve gevolgen wanneer zij tegen het algoritme ingaan?
  - Hoe makkelijk is het voor beoordelaars om tegen het algoritme in te gaan?
  - Heeft de organisatiecultuur invloed op het werk van beoordelaars?
- → **Vinden er kwaliteitscontroles plaats op het werk van beoordelaars?**
  - Wat zijn de gevolgen van deze controles?
  - Worden deze kwaliteitscontroles teruggekoppeld?
- → **Neemt de organisatie verantwoordelijkheid voor het nemen van een onjuist besluit over een persoon?**

Ondersteuning

Het werk van een beoordelaar vergt veel aandacht. Er worden immers in relatief korte tijd vaak beslissingen genomen met ingrijpende gevolgen voor mensen. Wanneer beoordelaars hier voldoende ondersteuning bij krijgen, kan dit een indicatie zijn van betekenisvolle menselijke tussenkomst. Ondersteuning kan op verschillende manieren plaatsvinden, zoals met een vertrouwenspersoon. Maar beoordelaars kunnen ook steun vinden bij elkaar. Beoordelaars nemen doorgaans individueel besluiten. Een team-in-the-loop-aanpak kan de mentale druk onder beoordelaars verdelen. Bij een team-in-the-loop-aanpak is niet één beoordelaar verantwoordelijk voor een besluit over een individu, maar kijken er meerdere beoordelaars mee.

Vragen om bij de inrichting te stellen:

- → **Hebben beoordelaars toegang tot voldoende (mentale) ondersteuning, als nodig?**
  - Is er een vertrouwenspersoon?
  - Hebben de beoordelaars de mogelijkheid om elkaar om hulp te vragen?
  - Is er tijdens training aandacht voor weerbaarheid?

4. Governance

Hoe houdt de organisatie eindverantwoordelijkheid?

Het is belangrijk dat een organisatie de verantwoordelijkheid blijft nemen voor de inzet van een algoritme. Menselijke tussenkomst zorgt ervoor dat de uitkomst van een algoritme niet leidt tot een besluit dat uitsluitend op geautomatiseerde verwerking is gebaseerd. Die verantwoordelijkheid hoort niet alleen bij de beoordelaar te liggen. Daarvoor waarschuwen verschillende onderzoekers die zich bezighouden met geautomatiseerde besluitvorming.[25] Hoe kan een organisatie de eindverantwoordelijkheid voor de uitkomsten van het proces bij de juiste personen houden?

Onderdelen

Inrichting

Het is wenselijk dat een organisatie het beleid voor betekenisvolle menselijke tussenkomst goed vastlegt in procedures. Procedures zoals keuzes over de inrichting, waarvoor alle hierboven beschreven onderdelen als leidraad kunnen dienen. In een Data Protection Impact Assessment (DPIA) stelt de verwerkingsverantwoordelijke de mate en fasering van menselijke tussenkomst vast in het besluitvormingsproces.[26] Voor een systeem dat uitkomsten geeft die aanmerkelijke gevolgen hebben voor betrokkenen zal hoogstwaarschijnlijk een DPIA moeten worden uitgevoerd.[27]

Het is goed om de beoordelaars te betrekken bij de inrichting van het proces en de ontwikkeling van het algoritme. Managers en ontwikkelaars die verantwoordelijk zijn voor de keuzes staan namelijk mogelijk verder van het proces en hebben mogelijk niet de juiste kennis.

Beoordelaars kunnen bijvoorbeeld adviseren welke aspecten door een algoritme beoordeeld kunnen worden en welke beter door een mens. Ook kunnen zij aangeven of de interface duidelijk genoeg is.

Voordat een algoritme wordt gebruikt, is het goed om deze met de menselijke tussenkomst te testen. De resultaten hiervan en eventuele daaruit volgende aanpassingen kunnen worden opgenomen in een DPIA. In deze testfase kan onder andere getoetst worden hoeveel tijd een beoordelaar nodig heeft voor de tussenkomst, welke kennis nog mist voor een goede beoordeling en welke variabelen volgens beoordelaars missen in het algoritmische deel.

Vragen om bij de inrichting te stellen:

- → **Is er beleid voor betekenisvolle menselijke tussenkomst vastgelegd?**
- → **Wie zijn betrokken geweest bij het opstellen van dit beleid?**
  - Is het perspectief van beoordelaars hierin meegenomen?
- → **Wie zijn betrokken geweest bij het ontwerp van (de interface van) het algoritme?**
  - Is het perspectief van beoordelaars hierin meegenomen?
  - Is het algoritme met de beoordelaars getest?
- → **Is er een DPIA op het proces uitgevoerd?**
  - Is het perspectief van beoordelaars en betrokkenen hierin meegenomen?

Training

Om betekenisvolle menselijke tussenkomst te ontwikkelen, hebben beoordelaars training en informatie nodig.[28] Veel van de hierboven genoemde onderdelen vragen instructie aan de beoordelaar.

De aspecten die bij die training van belang kunnen zijn:[29]

→ Alle relevante factoren:
- Beoordelaars begrijpen hoe hun expertise het algoritme aanvult, en weten welke factoren meegewogen moeten worden in het besluit.
  - De verantwoordelijke kan een lijst opstellen met factoren die van belang zijn.
  - Naast IT-kennis ook de benodigde (sociale) kennis om de beslissing te maken bijgebracht.
- Beoordelaars weten wanneer en op elke manier extra informatie kan worden opgevraagd, bijvoorbeeld bij de betrokkene.
  - Beoordelaars zijn op de hoogte van de impact van het opvragen van extra informatie bij de betrokken, als dat nodig is.

→ Menselijk inzicht:
- Beoordelaars weten welke mogelijkheden er zijn om maatwerk te leveren voor een specifieke situatie.
- Beoordelaars worden in de training niet aangemoedigd of geïnstrueerd om hun besluit aan te passen aan de organisatorische belangen, zoals het behoud van klanten.
- In de training is aandacht voor menselijke vooroordelen.
- Het algoritme wordt in de training niet gepresenteerd als onfeilbaar of beter dan de mens.

→ Bekwaamheid:
- Beoordelaars begrijpen hoe het algoritme tot een uitkomst komt.[30] Enkele voorbeeldvragen die dit soort kennis kunnen testen:
  - Hoe verandert de output van het algoritme als bepaalde variabelen veranderen?
  - Wat zijn de belangrijkste inputs voor de uitkomst van het algoritme?
  - Welke keuzes zijn gemaakt in de ***fairness metrics*** van het algoritme? Oftewel, welke groepen worden mogelijk vaker aangewezen door het algoritme dan anderen?
  - Welke regels volgt het algoritme?
  - Wat wordt er precies beoordeeld? De verantwoordelijke moet ook benadrukken dat het besluit impact heeft op een persoon en toelichten wat die impact is.
- Beoordelaars kennen de beperkingen van het algoritme.
  - Kunnen beoordelaars herkennen wanneer de output van het algoritme (waarschijnlijk/mogelijk) verkeerd is?
  - Weten beoordelaars hoe vaak het algoritme fouten maakt?
- Beoordelaars kunnen een stevige onderbouwing geven voor het accepteren of afkeuren van de uitkomst van het algoritme.
  - De verantwoordelijke kan richtlijnen geven voor de gevallen dat het corrigeren van een algoritme nodig is.

→ Technologie en ontwerp:
- In de training is aandacht voor ***automation bias*** en ***algorithmic aversion***.
- Beoordelaars weten hoe ze met de interface van het algoritme om moeten gaan en kennen de betekenis van de informatie die wordt afgebeeld.
- In de training wordt gewaarschuwd voor het routinematig beoordelen van de uitkomsten.

→ Proces
- Beoordelaars weten hoe ze een uitkomst van het algoritme kunnen afkeuren.
- Beoordelaars zijn zeker van het feit dat de uitkomst van het algoritme afgekeurd kan worden, zonder dat dat negatieve gevolgen heeft voor de beoordelaar. Hier is in de training aandacht voor.[31]
- Beoordelaars weten waar ze terecht kunnen met zorgen, vragen en feedback.
- De training besteedt aandacht aan weerbaarheid.

Vragen om bij de inrichting te stellen:

- → **Welke van bovenstaande onderdelen zijn relevant voor de training?**
- → **Hoe houdt de organisatie de kennis actueel, bijvoorbeeld door opfristrainingen?**

Toetsing en monitoring

Ook na de inrichtingsfase is het wenselijk dat het proces kan worden aangepast op basis van signalen zoals hierboven beschreven. Zeker nu er nog volop onderzoek is naar de menselijke rol bij geautomatiseerde besluitvorming en de vormgeving daarvan. Daarom is het belangrijk dat de verantwoordelijke zicht houdt op het functioneren van het proces. Dat kan door te toetsen en te monitoren.

Om te toetsen of menselijke tussenkomst daadwerkelijk betekenisvol is, kan de organisatie verschillende metrics bijhouden. Een simpele methode is inzicht in het aantal keer dat een beoordelaar de uitkomst van een algoritme afkeurt (of van een 'ja' naar een 'nee' verandert en andersom). Dit kan vervolgens aanleiding zijn voor verder onderzoek. Beoordelaars kunnen ook feedback krijgen over hun eigen accuraatheid in het omgaan met uitkomsten van het algoritme.

Een verantwoordelijke kan ook terugkerend ***mystery shopping***-acties uitvoeren, waarbij expres misleidende gegevens of algoritmische output aan de beoordelaar worden gegeven. De beoordelaar moet het daarmee oneens zijn en controleren of de misleidende gegevens worden opgepikt door de beoordelaars. Ook kan de beoordelaar regelmatig een besluit voorgelegd krijgen dat helemaal zonder algoritme moet worden genomen. Een andere indicatie om het proces opnieuw onder de loep te nemen, kan van de betrokkenen komen. De verantwoordelijke kan daarvoor bijhouden hoeveel klachten, bezwaren of verzoeken voor menselijke tussenkomst binnenkomen en de opvolging daarvan goed in de gaten houden.

In de evaluatie is de koppeling tussen individuele besluiten en het hele proces ook van belang. Zo is het wenselijk dat het proces en het algoritme aanpasbaar zijn na feedback van beoordelaars. De beoordelaar kan namelijk ook een rol spelen als degene die het dichtst op de besluiten zit. De beoordelaar moet dan in de gaten houden of het hele proces behoorlijk blijft. Wanneer een beoordelaar bijvoorbeeld ziet dat een groep mensen met een bepaald opleidingsniveau opvallend vaak door het algoritme wordt uitgeworpen, kan dat reden zijn om het algoritme te testen op bias. Belangrijk is wel dat de verwerkingsverantwoordelijke het toezicht op het hele proces hiermee niet afschuift op de beoordelaar. Het spreekt voor zich dat de verantwoordelijke het proces zo nodig ook aanpast na het toetsen en monitoren.

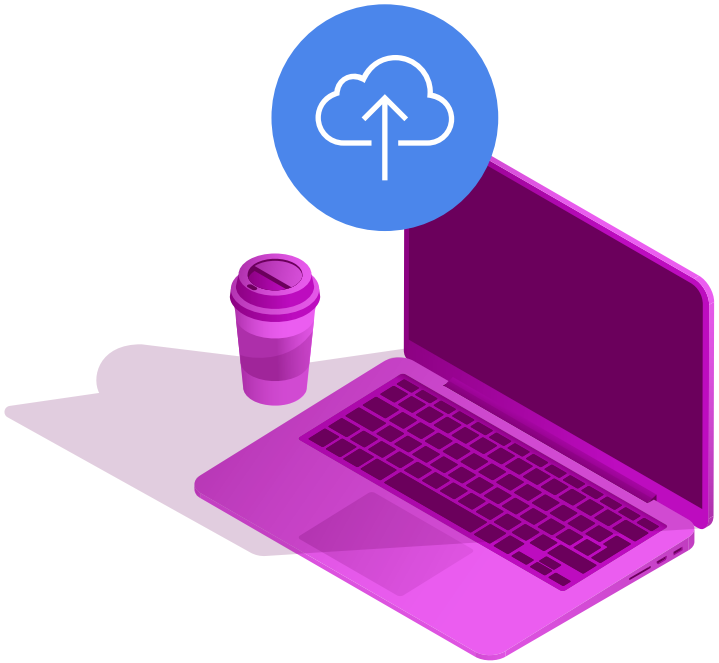
Het monitoren van de beoordelaars kan ook een ***chilling effect*** hebben. Daardoor zullen beoordelaars minder geneigd zijn om een algoritme (openlijk) in twijfel te trekken. De manier van monitoren en de communicatie daarvoor helpen dit effect te voorkomen.

Vragen om bij de inrichting te stellen:

- → **Wordt het algoritme aangepast na feedback van beoordelaars, betrokkenen, of monitoring?**
  - Hebben de mensen die de feedback beoordelen hiervoor de juiste vaardigheden?
- → **Voelen beoordelaars zich vrij om kritiek te leveren op het algoritme en het proces daaromheen?**
- → **Monitort de organisatie de mate van betekenisvolle menselijke tussenkomst?**
  - Denk hierbij bijvoorbeeld aan ***mystery shopping***. Evalueer hoe vaak betrokkenen bezwaar maken en hoe vaak beoordelaars tegen het algoritme ingaan.
- → **Hoe gaat de organisatie om met fouten van het systeem?**
  - Hoe gaat de organisatie om met foutpositeven en foutnegatieven?[32]

5. Conclusie

Met dit document is het begrip 'betekenisvolle menselijke tussenkomst' tastbaarder gemaakt voor verwerkingsverantwoordelijken die betekenisvolle menselijke tussenkomst willen inrichten. We vullen dit document aan als dat nodig is.



Bronnen

[1] Artikel 22 lid 1 AVG.

[2] Richtsnoeren inzake geautomatiseerde individuele besluitvorming en profilering voor de toepassing van Verordening (EU) 2016/679, p. 9-24.

[3] Richtsnoeren p. 24.

[4] Lazcoz, G., & de Hert, P. (2023). Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities. *Computer Law & Security Review, 50*, p. 18.

[5] Artikel 14 lid 2 AI verordening.

[6] Zie o.a. Robbins, S. (2017). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines,* 29; Grant, D. G., Behrends, J., & Basl, J. (2023). What we owe to decision-subjects: beyond transparency and explanation in automated decision-making. *Philosophical Studies*, *182*. Dit wil niet zeggen dat betekenisvolle menselijke tussenkomst voorgoed onmogelijk is bij case-based algoritmes: er zijn technieken in ontwikkeling die mogelijk wel inzicht kunnen verschaffen, zie hiervoor o.a. Molnar, C. (2024). Chapter 10 Neural Network Interpretation, *Interpretable Machine Learning*: https://christophm.github.io/interpretable-ml-book/neural-networks.html.

[7] Frankrijk verbiedt bijvoorbeeld het gebruik van algoritmische besluiten in het juridische domein onder artikel 47 van de *Loi Informatique et Libertés*.

[8] Richtsnoeren, p.24

[9] Binns, R. (2020). Human Judgment in Algorithmic Loops: Individual Justice and Automated Decision-Making, *Regulation & Governance, 11* (1).

[10] Zie bijvoorbeeld Sztandar-Sztanderska, K., & Zielenska, M. (2022). When a Human Says "No" to a Computer: Frontline Oversight of the Profiling Algorithm in Public Employment Services in Poland, *Sozialer Fortschritt* 71 (6-7), p. 2.

[11] Zie hiervoor o.a. Solove, D. J., & Matsumi, H. (2024). AI, Algorithms and Awful Humans, *Fordham Law Review, 92* (5); Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet*, 11(1); Lazcoz, G., & de Hert, P. (2023). Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities. *Computer Law & Security Review, 50*; Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review, 45*; Palmiotto, F. (2024). When Is a Decision Automated? A Taxonomy for a Fundamental Rights Analysis. *German Law Journal, 25* (2).

[12] Richtsnoeren, p.24.

[13] Fussey, P., & Murray, D. (2024). Policing Uses of Live Facial Recognition in the United Kingdom. *AI Now Institute*. https://ainowinstitute.org/wp-content/uploads/2023/09/regulatingbiometrics-fussey-murray.pdf

[14] Cummings, M. L. (2006). Automation and Accountability in Decision Support System Interface Design. *The Journal of Technology Studies, 32* (1).

[15] Dergelijke ontwerp onderdelen worden vaak ingezet bij dark patterns: trucs die worden gebruikt in het ontwerp van websites en apps om mensen dingen laten doen die ze niet van plan waren (zoals het accepteren van cookies). Bij dark patterns worden design onderdelen op een negatieve manier ingezet en niet in het voordeel van gebruikers. Maar ontwerp onderdelen kunnen ook in het voordeel van gebruikers worden ingezet.

[16] Loop (samen met een beoordelaar) het beoordelingsproces langs, om zo te zien wat een beoordelaar te zien krijgt bij het maken van een besluit. Let hierbij op welke data te zien is, maar ook op *hoe* de interface is vormgegeven. Is de interface bijvoorbeeld begrijpelijk en overzichtelijk?

[17] Green, B., & Chen, Y. (2021), Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts, Proceedings of the ACM on Human-Computer Interaction, 5.

[18] Zie ook de Richtsnoeren, p. 24.

[19] Zie hiervoor ook Rapportage AI- en Algoritmerisico's Nederland, editie 3, Autoriteit Persoonsgegevens, p.4., https://www.autoriteitpersoonsgegevens.nl/documenten/rapportage-ai-algoritmerisicos-nederland-ran-voorjaar-2024.

[20] Zie Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet*, 11(1).

[21] Zie ook Raad van State. (2021). Digitalisering: Wetgeving en bestuursrechtspraak. https://www.raadvanstate.nl/@125918/publicatie-digitalisering/.

[22] Loop als dit mogelijk is het proces een aantal keer na met beoordelaars, om zo een inschatting te krijgen van hoeveel tijd beoordelaars doorgaans besteden aan een besluit. Let op: hierbij kan het nodig zijn om met synthetische datasets te werken.

[23] Een besluit met minder ingrijpende gevolgen op een individu vereist wellicht minder tijd.

[24] Sztandar-Sztanderska, K., & Zielenska, M. (2022). When a Human Says "No" to a Computer: Frontline Oversight of the Profiling, 2022
Algorithm in Public Employment Services in Poland, *Sozialer Fortschritt,* 71 (6-7).

[25] Besproken door Green, B., & Wagner, B. tijdens het *IPEN event on "Human oversight of automated decision-making,"* 3 september 2024. Beoordelaars kunnen als schuldige worden aangewezen voor schadelijke effecten die veroorzaakt worden door de keuze om een algoritme in te zetten, of door verkeerd ontwerp. Menselijke tussenkomst fungeert dan als een pleister op een proces dat in de basis al kapot is.

[26] Richtsnoeren inzake geautomatiseerde individuele besluitvorming en profilering voor de toepassing van Verordening (EU) 2016/679, p. 25.

[27] Uit Lazcoz, G., & de Hert, P. (2023): "Article 35(3)(a) is however crystal clear by requiring a DPIA in all cases of systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person."

[28] In een Oostenrijkse zaak bepaalde de federale administratieve rechtbank (BVwG) dat de verwerkingsverantwoordelijke beoordelaars van training en instructie moet voorzien zodat zij de uitkomsten van het algoritme niet kritiekloos overnemen (ECLI:AT:VWG-H:2023:RO2021040010.J09).

[29] Deels afkomstig uit Information Commisioner's Office (2020). Guidance on the AI auditing framework. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.

[30] Uit Lazcoz, G., & de Hert, P. (2023): "The French Conseil Constitutionnel (Conseil Constitutionnel, Décision n° 2018-765 DC du 12 juin 2018, §71) declares human intervention a fundamental safeguard in the design and development of AI algorithms, see Malgieri (n 16) 15. And further recognises the link between this safeguard and the ability to explain, in detail and in an intelligible form, how the processing has been carried out to data subjects."

[31] In Sztandar-Sztanderska, K., & Zielenska, M. (2022). When a Human Says "No" to a Computer: Frontline Oversight of the Profiling werd een laag aantal correcties onder andere veroorzaakt door informatie die beoordelaards van de trainer hadden gekregen: "As one of the client advisors from the pilot study recalled, PES staff were warned by the trainer (who took part in designing the profiling tool) not to correct profiles, "because the Ministry will check it" and the employees will "have to explain themselves"."

[32] Denk bijvoorbeeld aan het fout markeren van iemand als 'fraudeur' in een fraude-opsporingsalgoritme. Dan spreken we van 'false positive'.

AUTORITEIT
PERSOONSGEGEVENS