HOE MAAK JE DE BESLISSINGEN VAN AI **VERKLAARBAAR**?

Hoe kan je weten of de beslissing die een AI-systeem voor jou maakt wel correct is, eerlijk is en dat er geen fout in het systeem zit? In deze brAInfood gaan we dieper in op 'verklaarbaarheid' (explainability) van AI-systemen. Wat betekent verklaarbaarheid? Hoe kunnen AI-systemen verklaarbaar worden gemaakt? Welke verplichtingen heb je als AI-ontwikkelaar naar je gebruikers toe wanneer zij inzage vragen in de verklaarbaarheid van je AI-systeem? Waar wordt al sterk ingezet op verklaarbare AI-systemen? Heel wat informatie, daarom bevat deze brAInfood twee pagina's. Lees dus zeker beide pagina's.

Kenniscentrum Data & Maatschappij (2020). Hoe maak je de beslissingen van AI verklaarbaar? brAInfood van het Kenniscentrum Data & Maatschappij. Brussel: Kenniscentrum Data & Maatschappij.

brAInfood van het Kenniscentrum
Data & Maatschappij

Wat is 'verklaarbaarheid' en waarom is het belangrijk?

Of we het doorhebben of niet, we komen allemaal wel eens in aanraking met een AI-systeem. Soms wordt een AI-systeem gebruikt om jou een leuke, onschuldige aanbeveling te geven, in Spotify bijvoorbeeld. Maar soms wordt een AI-systeem ingezet om beslissingen te maken die jouw leven sterk kunnen beïnvloeden, bijvoorbeeld of je in aanmerking komt voor een lening of niet. Als een bankmedewerker deze beslissing voor jou maakt, dan verwachten we dat hij/zij kan uitleggen **waarom** je bent geweigerd voor een lening en wat je moet veranderen om toch in aanmerking te komen. Als een AI-systeem zo'n beslissing voor jou maakt, hoe weet je dan dat die beslissing correct is, eerlijk is en dat er geen fout in het systeem zit? Dit zijn vragen waar '**Explainable AI**' een antwoord op probeert te geven.

De 'explainability' of verklaarbaarheid van een AI-systeem zegt iets over in hoeverre een mens kan begrijpen (1) **hoe** het AI-systeem op een globale manier werkt en/of (2) **op basis waarvan** een individuele conclusie tot stand is gekomen. Vanuit een 'explainability' perspectief zijn de verschillende soorten AI-systemen grofweg in **drie categorieën** op te delen (white box, black box, grey box).

White box, black box, grey box

Een **white box** AI-systeem is een **simpel systeem**, waarvan de **globale werking te begrijpen** is voor mensen. Dit kan bijvoorbeeld een eenvoudige beslissingsboom zijn, of een aantal regels waar het AI-systeem aan voldoet. Om een white box systeem verklaarbaar te maken, hoeven deze regels alleen maar naar **mensentaal** te worden vertaald. Vanuit de globale werking is het dan gemakkelijk om de individuele uitkomsten van het systeem te verklaren. De eenvoud van een white box systeem maakt echter ook dat deze alleen maar **simpele vraagstukken** (accuraat) kunnen oplossen en niet zo vaak meer worden gebruikt.

Aan de andere kant van het spectrum zijn de **black box** AI-systemen. Black box AI-systemen zijn zeer **complexe systemen** waarvan de **globale werking vaak verborgen** is, omdat deze bijvoorbeeld zichzelf patronen kunnen aanleren. Dit maakt dat black box systemen veel **ingewikkeldere vraagstukken** (accuraat) kunnen oplossen, maar ook dat weinig mensen, zelfs de makers vaak niet, weten hoe het systeem werkt of op welke regels de individuele conclusies zijn gebaseerd.

Om een black box systeem verklaarbaar te maken is het vaak nodig om een **extra module** te gebruiken die de black box kan nabootsen, bijvoorbeeld een white box module. Dit geeft dan een **versimpelde weergave van de globale werking** van het systeem. Een andere optie is om met hulp van een white box module **regels** op te stellen voor de individuele conclusies van het black box systeem die wel begrijpelijk zijn voor mensen. Ook kan het in sommige gevallen nuttig zijn om het black box AI-systeem in een extra module te vragen **welke input data het meest belangrijk** waren om tot een bepaalde individuele conclusie te komen.

Een **grey box** systeem zit, zoals de naam al doet vermoeden, tussen een black box en white box systeem in. Deze systemen proberen de **hoge nauwkeurigheid van een black box** en de **simpele verklaarbaarheid van een white box** te combineren. Bijvoorbeeld door verklaarbaarheid gedurende het gehele ontwikkelingsproces mee te nemen en in te bouwen in het black box model (**explainability-by-design**), waardoor er geen aparte module meer nodig is om de werking na te bootsen of individuele uitkomsten te verklaren.

Hoe kan je de genomen beslissingen achterhalen?

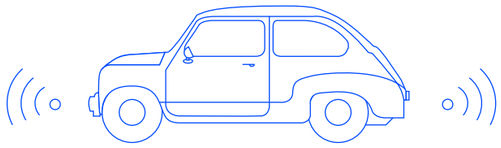
Als gebruiker van een dienst die een AI-component bevat is het niet altijd eenvoudig om te achterhalen waarom de dienst je bepaalde aanbevelingen presenteert en om de genomen beslissingen te controleren/verifiëren. Het gebruik van een white box of grey box kan hier een verklaring geven. Toegang tot deze boxes is vaak niet mogelijk doordat bedrijven geen inzage willen verschaffen in de interne werking van hun algoritme omdat deze bv. bedrijfsgeheimen of intellectuele eigendommen van de organisatie bevatten. Wanneer **inzage** toch wordt verleend, zijn de verklaringen vaak **op een technische manier verwoord**, gekoppeld aan specifieke 'AI-terminologie'. Een doorsnee gebruiker wordt als gevolg **niet altijd op een (aan)gepaste manier geïnformeerd**. Om hier een (wettelijke) oplossing voor te bieden is er bij het opstellen van de **AVG-reglementering** expliciet rekening gehouden met deze uitdaging.

Samengevat stelt de AVG dat een organisatie die persoonsgegevens verzamelt en verwerkt de **gebruiker** hiervan **op de hoogte** moet **stellen**. De gebruiker heeft het **recht om toegang** te krijgen tot zijn of haar persoonlijke data en, in het geval van geautomatiseerde besluitvorming, heeft hij/zij het recht toegang te krijgen tot zinvolle **informatie over de betrokken logica**, alsook het **belang** en de **beoogde gevolgen van de verwerking**. Deze informatie moet "in een beknopte, transparante, begrijpelijke en gemakkelijk toegankelijke vorm, gebruik makend van **duidelijke en eenvoudige taal**", gecommuniceerd worden.

Elke gebruiker heeft dus het recht om een verzoek in te dienen om toegang te krijgen tot zijn/haar persoonlijke data. Deze data moet op een toegankelijke manier aangeboden worden, samen met informatie rond de verwerking en analyse die de organisatie op/met deze data uitvoerde. Een organisatie die gebruik maakt van AI-systemen om een dienst aan te bieden moet de nodige verklaringen verstrekken, aangepast aan de individuele gebruikers(groepen) of stakeholders.

Een **gebruikersonderzoek** waarbij gepeild wordt bij verschillende gebruikers(groepen) en stakeholders naar de **verwachtingen rond** onder andere "**zinvolle informatie**", "**toegankelijke vorm**" en "**duidelijke en eenvoudige taal**" kan een concrete invulling geven aan deze begrippen. Op deze manier zijn organisaties niet enkel een voorbeeld voor de implementatie van de AVG-regelgeving, maar kunnen zij ook het vertrouwen bij gebruikers en de transparantie naar hen toe verhogen.

In **welke domeinen** wordt sterk ingezet op verklaarbare AI-systemen?



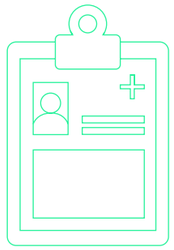
Autonome voertuigen

Reeds lang wordt voorspeld dat AI de 'motor' zal worden van de mobiliteit van de toekomst. AI zorgt namelijk voor de aansturing van autonome voertuigen, die op hun beurt voor een betere en efficiëntere verkeersdoorstroming moeten zorgen en vooral het verkeer veel veiliger moeten maken. Het waarborgen van veiligheid is ook één van de hoofdoorzaken waarom autonome voertuigen nog geen grote doorbraak kennen.

Het AI-systeem in autonome voertuigen neemt beslissingen op basis van de enorme **gegevensstroom over de weg- en rijomstandigheden** die het interpreteert via de **sensoren en camera's** van het voertuig. De gegevens worden verwerkt door AI-algoritmes die een inschatting maken van een situatie. Op basis van die inschatting neemt het AI-systeem een beslissing.

Die beslissing zal volledig transparant moeten zijn. AI-algoritmes die je kan trainen via deep learning, bijvoorbeeld op classificatie en het herkennen van patronen, zullen in een groot deel van de gevallen doen wat je verwacht, maar niet in alle scenario's. Bij het gebruik van Deep Learning zal er dus gebruik gemaakt moeten worden van een grey box model of white box module, om zo **inzicht** te krijgen **in hoe beslissingen tot stand komen**.

In een **veiligheidsgevoelige materie** als autonome voertuigen moet de uitkomst dus volledig verklaarbaar zijn. Men moet op de hoogte zijn van de morele prioriteiten die geprogrammeerd zijn. Een autofabrikant moet kunnen traceren hoe het model dat beslissingen neemt van punt A naar punt B gaat. Anders zal het voertuig niet vergund worden.

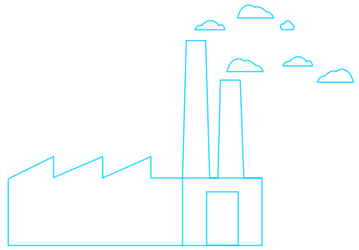


Gezondheidszorg

IBM kondigde in 2013 een toepassing aan die revolutionair zou zijn in de strijd tegen kanker. In samenwerking met het Anderson Cancer Center van de Universiteit van Texas MD zou IBM via zijn Watson for Oncology-toepassing artsen in staat stellen om waardevolle **inzichten** te ontdekken **uit de rijke patiënten- en onderzoeksdatabanken** van het kankercentrum.

Het systeem verwerkt gegevens uit medische literatuur, behandelingsrichtlijnen, medische dossiers, rapporten van laboratoria enz. om aanbevelingen voor kankerbehandeling te formuleren. Het algoritme suggereert vervolgens welke **behandelingsopties** de arts kan gebruiken bij de behandeling van de patiënt. In de programmering van het algoritme zijn tal van verschillende gegevensbronnen opgenomen die verschillend kunnen worden gewogen.

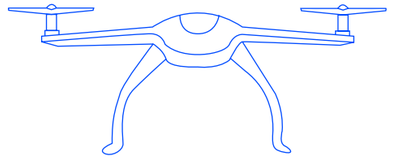
In 2017 kwam evenwel aan het licht dat het systeem foutief en ronduit gevaarlijk advies gaf voor de behandeling van kanker. Het black-box-karakter van het AI-systeem zorgde namelijk voor problemen. Hoewel het platform inzage gaf in de literatuur waaruit het zijn conclusie trok, was het **onduidelijk en onverklaarbaar** hoe het de **scoringscriteria** hanteerde om sommige studies te waarderen ten opzichte van andere. Het project werd vervolgens voor een bepaalde periode on hold gezet.



Fabrieken

Momenteel zijn er al robots die in fabrieken de werknemers assisteren in de uitvoering van hun job, bijvoorbeeld door een deel van het laswerk over te nemen. Er zijn zelfs robots die in staat zijn om hand in hand te werken met werknemers. Wanneer robots werknemers ondersteunen en assisteren, worden zij ook wel '**cobots**' genoemd.

Door de sterke samenwerking tussen robots en werknemers, die in de toekomst hoogstwaarschijnlijk verder zal toenemen, is het noodzakelijk dat er een **goede communicatie**, **vertrouwen**, **duidelijkheid** en **begrip** is tussen de machine en de mens. Explainable AI speelt een belangrijke rol bij onder meer het creëren van vertrouwen doordat het systeem aan de werknemer het onderliggende model kan uitleggen en verklaren waarom het bepaalde beslissingen maakt.



Defensie

Slimme AI drones kunnen ingezet worden als wapen in oorlogen. Op die manier is het niet langer nodig om een fysieke (face-to-face) confrontatie aan te gaan met het doelwit. Dit systeem kan al worden toegepast, maar een persoon heeft momenteel nog wel steeds de eindbeslissing. Indien dit volledig autonoom zou worden, is het cruciaal dat de **resultaten** van het systeem **zeer accuraat** zijn. Niemand wil natuurlijk een fout doelwit raken. Daarom moet het heel duidelijk zijn hoe het systeem juist werkt en hoe de beslissingen worden gemaakt. Onder meer The U.S. Defense Advanced Research Projects Agency onderzoeken dit verder.