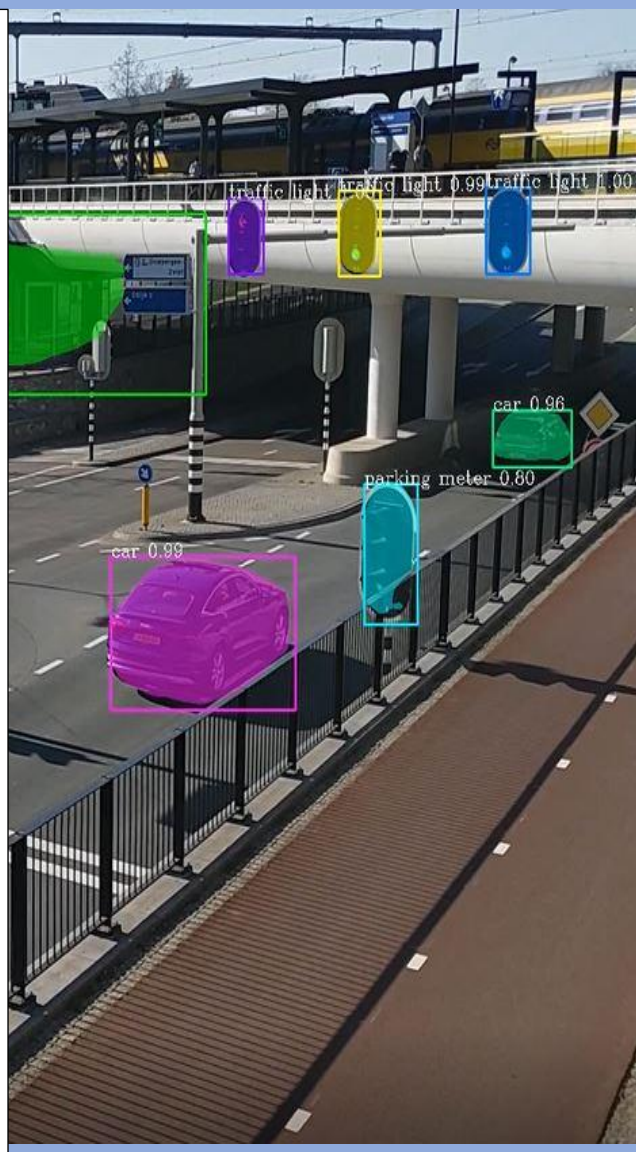
Kwaliteit en Testen

van

Artificial Intelligence



September 2021

Sander Mol
Peter Collewijn
Hannie van Kooten

Werkgroep Testen en AI

TEST NET

Versie: September 2021

Geschreven door Sander Mol en Peter Collewijn, onder redactie van Hannie van Kooten, met bijdragen van Rik Marselis en Mariëlle van der Sluys en medewerking van alle leden van de werkgroep Testen en AI.

Commentaar en verbeterpunten op dit document worden op prijs gesteld. Stuur deze reacties naar de werkgroep via mail. Zie de laatste bladzijde voor de contactgegevens.

De werkgroep wil graag in contact komen met andere groepen of organisaties die zich bezighouden met dit onderwerp. Ken je zo'n groep, of maakt je deel uit van zo'n groep en zoek je ook deze samenwerking, neem dan ook contact via email. Zie de laatste bladzijde voor de contactgegevens.

Introductie

Werkgroep Testen en AI

Deze whitepaper is geschreven door de TestNet werkgroep Testen en AI. Deze werkgroep is opgericht in januari 2018 en heeft zich in de eerste periode voornamelijk gericht op het testen met behulp van kunstmatige intelligentie ofwel Artificial Intelligence (AI). Na de publicatie van de whitepaper over dit onderwerp in 2019 zijn we verder gegaan met het onderwerp testen van AI. Met enige regelmaat geeft de werkgroep presentaties over het onderwerp testen met of van AI bij TestNet of elders.

Door het schrijven aan de whitepaper hebben we geprobeerd om op een gestructureerde manier vorm te geven aan ideeën die in de werkgroep leefden over kwaliteit en testen van AI. Bij een onderwerp dat nog in de kinderschoenen staat zijn er nog geen gebaande paden of best practices. Gaandeweg zijn wij gekomen tot de huidige indeling en het is ook voor ons een ontdekkingstocht geweest.

Doel van deze whitepaper

Er zijn twee belangrijke doelen van deze paper. De eerste is om collega-testers genoeg informatie en vertrouwen geven om zelf de risico's bij een AI-implementatie te herkennen, waarbij zij met eigen kennis en kunde het AI-testtraject vorm kunnen geven.

Het tweede doel is om een impuls te geven aan verdere ontwikkelingen op het gebied van kwaliteit en testen van AI, in het algemeen. Daarbij denken we aan zaken als het op langer termijn ontwikkelen van best practices, maar ook aan nieuwe vormen van testen en methodes die mogelijk gaan ontstaan. Wij willen de lezers dan ook uitnodigen om ervaringen te delen en te reageren op wat wij geschreven hebben. Contactgegevens staan op de laatste pagina.

Bronnen

Over de risico's en het testen van AI zijn de laatste jaren enkele boeken geschreven en zijn er enkele cursussen ontwikkeld. Ook zijn er richtlijnen opgesteld of in ontwikkeling vanuit diverse overheden, brancheorganisaties en andere samenwerkingsverbanden. De bronnenlijst is te vinden in de bijlage. Deze bronnen hebben we besproken binnen de werkgroep en vervolgens vertaald naar deze whitepaper.

We beseffen dat deze publicatie een momentopname is. Toch zijn we ervan overtuigd dat deze whitepaper een goed startpunt is voor het testen van AI, met inzichten die nog lange tijd bruikbaar zijn in het testvak.

Samenvatting en leeswijzer

Deze whitepaper is opgedeeld in hoofdstukken over onderwerpen zoals bijvoorbeeld risico's of ethische richtlijnen. In het tweede deel bevinden zich bijlagen over zaken als bronnen, verdiepingsmateriaal en een woordenlijst. In de woordenlijst wordt verwezen naar begrippen in de tekst, naar externe bronnen of er wordt een korte omschrijving gegeven.

In hoofdstuk 1 gaat het over algoritmes. Algoritmes om in 'gewone' software de prijs van een huis in te berekenen en over Machine Learning algoritmes. Wat maakt dat ze zo verschillend zijn en wat is Machine Learning nou eigenlijk?

In hoofdstuk 2 komen de risico's van Machine Learning aan de orde. Er is voor gekozen om, in onze ogen, de belangrijkste vijf te benoemen, die we tot algemene risico's hebben benoemd. Er zijn er veel meer, maar we hebben ons beperkt tot een hanteerbare hoeveelheid, zoals afhankelijkheid van data en beperkte uitlegbaarheid.

In hoofdstuk 3 wordt ingegaan op categorieën van Machine Learning, die we verschijningsvormen hebben genoemd. Daarbij kan gedacht worden aan beeldherkenning of spraakgeneratie. Per vorm worden er een aantal risico's besproken, die het meest tot de verbeelding spreken of verduidelijkend zijn. De meeste van de algemene risico's gelden voor alle vormen van toepassing. We hebben ook hier geen uitputtende lijst opgenomen, begrijpen vinden we belangrijker dan volledig zijn. De risico's blijken over het algemeen een sterkere relatie te hebben met de applicatie dan met de verschijningsvorm, merkten we na het schrijven van het hoofdstuk.

In hoofdstuk 4 hebben we de mate van autonomie van de AI-toepassing als uitgangspunt genomen. De mate van zelfstandigheid van een ML applicatie kan variëren van een soort collega waar je voorbewerkte informatie aangeeft om het te laten bewerken waarna je het resultaat van deze bewerking zelf nog controleert. Tot het andere uiterste, het volledig zelfstandig verkrijgen van informatie om daarna zelfstandig besluiten te nemen, denk aan een zelfrijdende auto.

In hoofdstuk 5 staat ethiek centraal. Een interessant onderwerp dat direct verband houdt met het testen van de AI toepassing. Welke regelgeving wordt er in de EU ontwikkelt om tot een robuuste veilige AI te komen? Welke aspecten van belang zijn zoals rechtvaardigheid, privacy, eerlijkheid en transparantie komen hier aan de orde.

In hoofdstuk 6 wordt dieper ingegaan op de niet functionele aspecten van een Machine Learning applicatie. In de ISO 25010 norm staat deze beschreven voor standaard software. Deze blijken te beperkt en er wordt ingegaan op een aantal initiatieven om specifiek voor ML nieuwe kenmerken en attributen toe te voegen.

In hoofdstuk 7 komen we aan bij het testen van AI. Hoe komen we tot een kwalitatief goede AI en welke tools kunnen we daarvoor inzetten? In dit hoofdstuk staan veel bekende testmethoden zoals A/B testen naast minder bekende zoals Metamorphic testing.

In hoofdstuk 8 wordt gekeken naar testen in de praktijk. Welke rollen zijn er binnen een AI-project en wie kijken er naar kwaliteit. De vaardigheden en kennisgebieden waarin testers die zich verder willen gaan ontwikkelen in de wereld van AI worden opgesomd.

Tot slot herhalen we onze oproep om kennis te delen en samen aan best practices te werken en de positionering van testers in een AI-team uit te werken en te verstevigen.

Inhoudsopgave

1 Artificial Intelligence

De begrippen Artificial Intelligence, Machine Learning en Deep Learning worden vaak door elkaar gebruikt. Ook deze whitepaper gebruikt voor de herkenbaarheid meestal de term ‘AI’, terwijl we ons vooral op de subcategorieën Machine Learning en Deep Learning richten. Dit hoofdstuk geeft de verschillen, maar ook de verwevenheid van de drie begrippen aan.

1.1 Wat is Artificial Intelligence (AI)

AI ofwel kunstmatige intelligentie is een containerbegrip voor alles wat we intelligent vinden en voortkomt uit een computer of machine in plaats van uit een mens. Artificial Intelligence is de laatste tien jaar een modewoord geworden. Begrippen als Machine Learning (ML), Unsupervised en Supervised Learning, Deep Learning worden regelmatig gebruikt en in het nieuws wordt er vaak gesproken over ‘algoritmes’.

Artificial Intelligence krijgt in ons dagelijks leven een steeds grotere rol. Met name het AI-onderdeel Machine Learning heeft een vlucht genomen, doordat de rekenkracht van computers en de hoeveelheid beschikbare data sterk zijn toegenomen, terwijl de kosten voor deze rekenkracht en het ontsluiten van data sterk zijn gedaald. De techniek is nu voor iedere geïnteresseerde binnen handbereik en het aantal toepassingen zal de komende jaren sterk toenemen.

1.2 Wat is Machine Learning (ML)

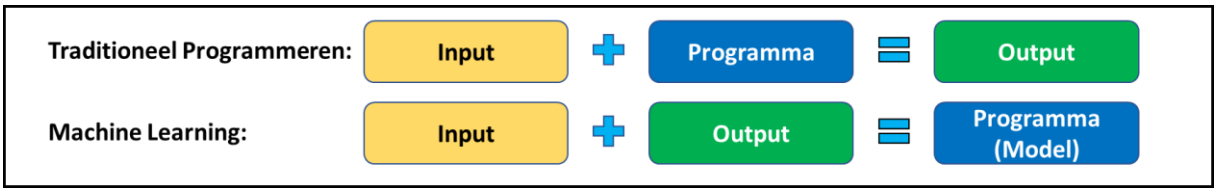
In traditioneel geprogrammeerde software vinden we intelligentie terug in de vorm van geprogrammeerde regels die wij mensen logisch vinden. We hanteren hiervoor begrippen zoals beslisregels, formules of algoritmes. Een voorbeeld van dit soort regels is het berekenen van een voorspelde huizenprijs:

De voorspelde huizenprijs in euro’s = 700 * woonoppervlak +
500 * perceeloppervlak +
8.000 * aantal slaapkamers +
score op basis van de postcode

Vervolgens kunnen we van een nieuw huis eenvoudig de voorspelde prijs berekenen, zolang we de inputvariabelen van dit huis maar weten.

Een belangrijk uitgangspunt bij geprogrammeerde regels is dat ze vooraf goed zijn doordacht, in dit geval door mensen die de huizenmarkt erg goed kennen. We weten dus ook dat, als we twijfels hebben of deze regel (nog) juist is, we bij deze experts terecht kunnen. Er is hierdoor een vaste manier om te bepalen of een test daadwerkelijk geslaagd is, ook wel een ‘test oracle’ genoemd.

Bij ML-algoritmes ligt dit anders. Het vertrekpunt is geen goed doordachte regel, maar een set aan verzamelde voorbeelden. Die voorbeelden bestaan uit invoervariabelen en een bijbehorende uitvoervariabele, die samen leiden tot een algoritme ofwel een model. In een afbeelding ziet dit er als volgt uit.



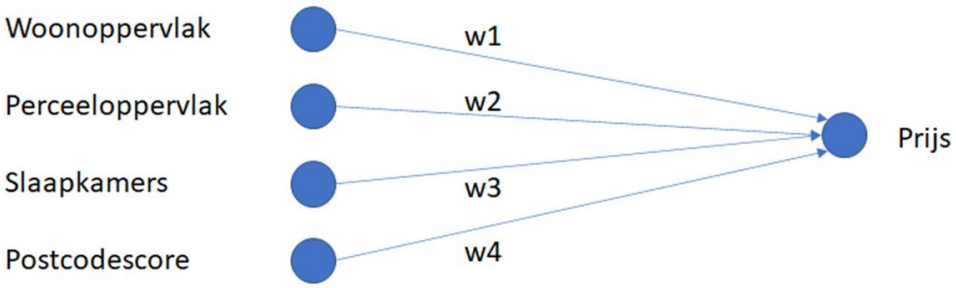
Afbeelding 1: Traditioneel programmeren - Machine Learning

Bij huizenprijzen zouden dit de eerste drie voorbeelden kunnen zijn:

Woonoppervlak	Perceeloppervlak	Slaapkamers	Postcode	Prijs
110	173	4	3311 AA	400.000
104	145	6	3351 ES	375.000
122	211	5	3352 VA	450.000
...	...	...	...	...

Tabel 1: Kenmerken van huizen - Huizenprijs

Vervolgens krijgt de ML-tool de opdracht om een zo goed mogelijk algoritme te bedenken voor het voorspellen van de prijs, ook van nieuwe huizen. Opnieuw zijn daarbij de invoervariabelen de basis voor de prijsvoorspelling. In een afbeelding ziet dat er als volgt uit. De mate waarin de invoer-variabelen meewegen, zijn nog onbekend. We tonen ze als variabelen, w (weight) 1 tot en met 4.



Afbeelding 2: Gewichten van kenmerken van huizen voor huizenprijs bepaling

Dan komen we bij de 'magie' van ML; het zelfstandig zoeken naar de juiste wegingen. Bij ML worden algoritmes gebruikt die de grondslag hebben op statistische technieken om deze wegingen te bepalen. Er ontstaat wederom een soort formule (ofwel algoritme of model) zoals getoond in het begin van dit hoofdstuk. Nu is het echter belangrijk om te beseffen dat de factoren (ofwel wegingen) niet op basis van menselijke expertise tot stand zijn gekomen. Ook is het mogelijk dat door het toepassen van een ander algoritme er een ander resultaat ontstaat. Van het resultaat is echter nooit volledig vast te stellen of dit helemaal juist is; het 'test oracle' ontbreekt.

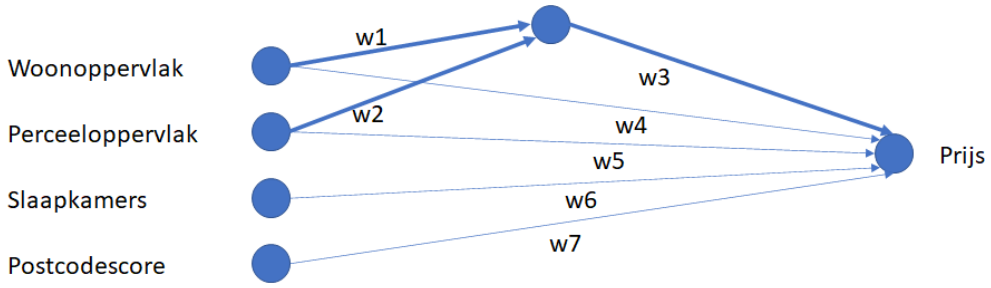
Toch biedt ML vaak voldoende zekerheid om in de praktijk te worden toegepast. Het gemak om data te gebruiken om van te leren en uiteindelijk tot een productiewaardig model te komen is erg aantrekkelijk. Bovendien is het met ML mogelijk om toepassingen te ontwikkelen die met geprogrammeerde software niet mogelijk zou zijn. Denk bijvoorbeeld aan het herkennen van beelden of spraak. Daarnaast is het mogelijk om getrainde modellen naar eigen behoefte aan te passen of te detailleren. Zo is het mogelijk om een model, dat is getraind om auto's te herkennen,

aan te passen om bepaalde merken of klassen te herkennen. Dit beperkt de benodigde tijd om het model te trainen aanzienlijk.

1.3 Wat is Deep Learning (DL)

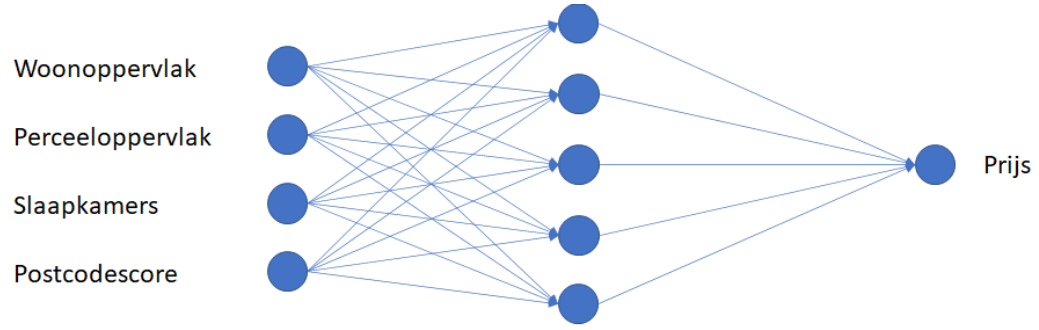
Vaak is de voorspellende waarde van ML-modellen te beperkt als er alleen wordt gekeken naar de directe relatie tussen een invoervariabele en de uitvoervariabele. De verbanden tussen twee invoervariabelen zouden ook iets kunnen zeggen over de uitvoervariabele. Zo zou een combinatie van een klein woonoppervlak en een groot perceeloppervlak kunnen duiden op een luxe, vrijstaande villa en dus een hogere prijs.

We hebben daarom een uitbreiding nodig op de eerder gebruikte afbeelding om de zojuist genoemde verbanden te kunnen vinden. Specifiek voor woonoppervlak en perceeloppervlak ziet dit er als volgt uit.



Afbeelding 3: Gewichten van kenmerken van huizen voor huizenprijs-bepaling geavanceerd

In plaats van een algoritme te gebruiken, wordt er nu gestart met het willekeurig toekennen van waarden aan de gewichten. De voorspelde prijs op basis van deze gewichten zal vermoedelijk totaal verkeerd zijn en de fout is met de werkelijke uitkomst is groot. Het doel is om deze fout zo klein mogelijk te maken. De DL-tool past daarom de gewichten een klein beetje aan en bepaalt of de voorspelde prijs een klein beetje beter of een klein beetje slechter is geworden. Bij een verbetering worden de nieuwe wegingen als uitgangspunt gebruikt voor de volgende kleine aanpassing van deze wegingen. Dit iteratieve proces wordt net zo lang herhaald totdat het model een nauwkeurig voorspellende waarde heeft of er geen verbetering meer optreedt.



Afbeelding 4: Gewichten van kenmerken van huizen voor huizenprijs-bepaling met tussenlaag

Om al deze combinaties te kunnen maken via bovenstaande afbeelding, is voor alle 25 lijntjes een weging nodig. Deze wegingen krijgen vooraf een willekeurige waarde en worden bij iedere trainingsronde een klein beetje verbeterd.

Het is belangrijk om te beseffen dat het model hierboven een extra ‘laag’ heeft gekregen van 5 nodes, waarbij deze nodes geen vaste betekenis hebben terwijl de invoervariabelen dit wel hebben. We hadden in plaats van 5 ook bijvoorbeeld 3 of 20 nodes in deze extra laag kunnen kiezen.

Ook bij 20 nodes zou het model nog bijzonder klein zijn. In de praktijk zijn tientallen invoervariabelen, tientallen nodes per laag en tientallen lagen geen uitzondering. Hierdoor wordt dit model vaak een neuraal netwerk genoemd, omdat het doet denken aan de opbouw van hersenen via neuronen. Dit verhoogt het aantal wegingen dat kan worden aangepast om te komen tot een goede voorspelling. De toevoeging van al deze nodes zonder vaste betekenis zorgt tevens voor minder voorspelbaarheid van de uitvoervariabele.

1.4 AI en risicogebaseerd testen

Bij het testen van software kijken we als testers al decennia naar de productrisico’s, om zo te bepalen hoeveel tijd we nodig hebben en waar we deze tijd aan willen besteden. Risico’s bestaan uit de impact van een bepaalde negatieve gebeurtenis en de kans op deze gebeurtenis.

Dit geldt natuurlijk ook bij software die de regels heeft geleerd via training op basis van data. Als op het vliegveld je gezicht op de camera niet kan worden gematcht met de foto op je paspoort, dan kun je na een handmatige verificatie alsnog vliegen. Dit is een lage impact. Als je echter wordt herkend als crimineel, dan heeft dat mogelijk een hoge impact. Dat hangt af van hoe men met dit resultaat omgaat; misschien moet je langs security en mis je je vlucht. In het ergste geval wordt je geboeid weggebracht en zit je misschien de nacht in een cel.

Dit laat meteen zien dat het risico van AI-voorspellingen twee kanten op kan; een onterechte herkenning (false positive) of onterecht iets *niet* herkennen (false negative). In beide gevallen is het goed om vooraf na te denken over de impact.

Risicogebaseerd testen blijft dus van belang. Wel hebben we gemerkt dat de risico’s van AI-ondersteunde software er in het algemeen anders uitzien dan bij geprogrammeerde software. Het volgende hoofdstuk over gaat over deze risico’s.

2 De algemene risico's van AI

Iedere softwaretoepassing heeft eigen, specifieke risico's. Het zal dan ook altijd nodig zijn om bij het ontwikkelen van software na te denken over deze risico's. Wel denken we als TestNet werkgroep dat er in het algemeen risico's te benoemen zijn die uitsluitend of relatief vaak gelden voor AI-ondersteunde software. Dit kan helpen om sneller en completer te zijn bij de risicoanalyse van een specifieke AI-ondersteunde softwaretoepassing.

In dit hoofdstuk werken we de belangrijkste algemene risico's verder uit. Daar waar nodig vatten we ook de belangrijkste technische concepten samen.

2.1 Onzekere uitkomsten

Bij het berekenen van huizenprijzen was al te zien, dat het relatief makkelijk is om de verwachte uitkomsten te bepalen met regel gestuurde, geprogrammeerde algoritmen. Het gebruik van ML leidt echter tot een applicatie waarvan we de exacte regels niet vooraf kennen. Hier zijn nog enkele voorbeelden:

- Op basis van de tekst die de klant in de chat heeft getypt, voorspelt het algoritme dat hij een schade aan zijn auto wil melden.
- Op basis van de pixels van een bepaalde afbeelding, voorspelt het algoritme dat er een hond op deze afbeelding staat.

In al deze gevallen geeft het algoritme geen absoluut zeker antwoord, maar een antwoord met een bepaalde zekerheids- of nauwkeurigheidspercentage. Bij de afbeelding met de hond voorspelt het algoritme bijvoorbeeld met 97% zekerheid een hond, maar ook met 82% zekerheid een wolf en met 56% zekerheid een kat.

De vraag die voor iedere softwaretoepassing moet worden gesteld, is in hoeverre je wilt bouwen op deze onzekere resultaten. Is het bij de afbeelding van de hond voldoende om deze met 90% zekerheid te herkennen? En wat als zowel de hond als de wolf met 95% zekerheid herkend worden in één afbeelding? En als we in totaal 100 dieren willen herkennen, hoe weten we dat we alle mogelijke combinaties van juiste en onjuiste voorspellingen correct afhandelen?

Dit zijn allemaal onzekerheden die bijdragen aan het risico. Bij het voorspellen van de juiste dieren is dit nog vrij onschuldig, maar bij het voorspellen van ziekten heeft dit al snel een hoge impact, zowel bij false-positives (zorgkosten en onjuiste behandeling) als false-negatives (uitblijven van behandeling). Zo zijn er nog veel meer voorbeelden waarbij het risico van onzekere uitkomsten erg groot is.

Omdat dit onderwerp zo belangrijk is en misschien complex lijkt, hebben we in bijlage C een uitgebreide uitleg opgenomen over nauwkeurigheidspercentages.

2.2 Afhankelijkheid van data

Zoals in het vorige hoofdstuk is aangegeven, is ML het onderdeel van AI dat leert van verzamelde data om te komen tot voorspellingen bij nieuwe data. De kwaliteit van deze data is van zeer groot belang en er is een veelvoud aan risico's bij het verzamelen en gebruiken van deze data:

Hieronder volgen een aantal punten die van invloed zijn op de kwaliteit van de data en dus uiteindelijk op de kwaliteit van het ML model.

- **Gebruikte bronnen voor de data**. Deze bronnen moeten representatief zijn en in een juiste verhouding staan met de werkelijkheid. Welke bron gebruikt men voor het bepalen van een huis; het kadaster, de gemeente met de WOZ-waarde, of de prijs van een huis dat net in dezelfde straat is verkocht? Of een combinatie van deze bronnen?
- **Selectie van de data**. Het is onmogelijk om alle beschikbare data te gebruiken en er zal dus filtering plaatsvinden. Hierbij ontstaat het risico dat de dataset te breed of te smal gekozen wordt. Voor het bepalen van een huizenprijs is de oppervlakte en het aantal kamers zeker van belang, maar of de tuin van gras is voorzien of van een stenen terras wordt al twijfelachtig. Maar het zou best kunnen zijn dat dit soort gegevens juist positief bijdraagt in de juistheid van het model.
- **Het doel van de data**. Data kan verzameld zijn met een ander doel dan voor het trainen van ML-modellen. Er kunnen gegevens ontbreken, omdat men ze voor het oorspronkelijke doel niet nodig had of omdat ze tijdens het verzamelen niet gecontroleerd zijn. Er kan ook onnauwkeurigheid zijn opgetreden door het weglaten van een gegeven als eenheid.
- **Volledigheid van de data**. De wijze waarop de gegevens worden verzameld. Zijn alle velden op de juiste identieke wijze gevuld, komen deze velden ook voor in alle gegevensbronnen? Bij een webformulier bestaat de mogelijkheid om velden verplicht te maken of door middel van selectievelden van vaste waarden te voorzien. Bij mail, chat en telefoongesprekken niet. Bedenk dat de verzamelde data slechts een subset is van de werkelijkheid. Er is een zeer reële kans dat de verzamelde data de werkelijkheid niet goed representeert en zelfs dat sommige praktijkgevallen niet beschikbaar zijn om een model te trainen.
- **Eenduidigheid van de data**. Het aanvullen van gegevens. Het kan voorkomen dat velden geen waarde of meerdere waarden bevatten die hetzelfde betekenen. Bijvoorbeeld Geslacht: Man en M of Vrouw en V. Er zal gekozen worden om deze velden van een eenduidige inhoud te voorzien. Dus het veld geslacht alleen de waarde M en V. Ontbrekende waarde worden voorzien van een default waarde of bijvoorbeeld een gemiddelde.
- **Persoonsafhankelijkheid van data.** Data is gebonden aan de persoon die het verzamelt, of die de verzameling heeft ingericht. De vraagstelling en antwoordmogelijkheden kunnen sturend zijn. De vastlegging kan selectief zijn, zowel bij het invoeren van registraties in een database als bij het creëren en verzamelen van beelden, teksten, enzovoort.
- **Tijd van de dataverzameling**. Het moment van verzamelen. Ook de datum en tijdstip zijn van belang. Vindt dit alleen plaats op een bepaald tijdstip van de dag, of worden alleen werktijden meegenomen? Het kan zijn dat er een verband is met het moment van de week of na een gebeurtenis. Bijvoorbeeld een piek van meldingen na het weekend of na het lanceren van een software update. Dit kan een ander soort informatie zijn. Tot hoever gaat men terug in de tijd voor de huizenprijs; een maand, een jaar of vijf jaar?
- **Formaat en bron van de data.** De integratie van de gegevensbronnen. Is het formaat van de brondata hetzelfde? Een datumveld kan in het ene systeem dag-maand-jaar zijn en in een ander systeem jaar-maand-dag. Ook bestaat de kans dat in cijfers de notaties van de punt en de komma afwijkt. Daarnaast kunnen gegevens in verhouding of in eenheid afwijken. Stel het ene huizensysteem werkt met lengte en breedte in centimeters en een ander met meters,

dat is een factor 100 verschil. Ook de frequentie of tijdsspanne kan een verschil geven. Het ene systeem rapporteert het aantal verkopen per uur en het andere systeem het aantal verkopen per dag. Bij integratie van deze gegevens moet men hier rekening mee houden.

- **Labelen van data.** Het geven van een label aan een object lijkt simpel, maar is in praktijk vaak toch een uitdaging. Zie het onderstaande voorbeeld over het labelen van katten. Verkeerd gelabelde gegevens zorgen voor verkeerde informatie. Het labelen is een subjectief proces. De staat van een huis kan bijvoorbeeld slecht, matig, goed of uitstekend zijn. Maar waar ligt het verschil tussen goed en uitstekend? Dit zal van persoon tot persoon verschillen.
- **Uitschieters in de data en bandbreedte.** Het detecteren van 'outliers' ofwel uitschieters. Dit zijn waarden die buiten de gangbare bandbreedte liggen en niet binnen het verwachte patroon van een gegeven dataset volgen. Bijvoorbeeld een huis met 12 kamers, terwijl alle overige huizen 2, 3, 4 of 5 kamers heeft. Als deze outliers niet regelmatig voorkomen en niet bijdragen aan het resultaat van het model, kunnen ze worden verwijderd. In de praktijk is het goed om naar de oorzaak van deze uitschieters in de data te kijken omdat ze het gevolg kunnen zijn van bijzondere zaken als fraude, hackaanvallen of storingen. Als de oorzaak niet wordt weggenomen, dan kan herhaling van de outliers in de toekomst ook weer voorkomen. Het is goed om aan de input kant van een model controle in te bouwen om deze uitschieters te voorkomen. Het gevolg kan zijn dat een data variabele een bepaalde bandbreedte krijgt, zowel aan de kant van invoer- als uitvoervariabelen. Als er met bandbreedte gewerkt gaat worden bij het samenstellen van de dataset voor de training van het model, dan is deze bandbreedte ook van toepassing bij het gebruik van het model in de productieomgeving.

Ondanks de ruime opsomming is deze lijst is nog niet volledig. Dit geeft wederom de afhankelijkheid van data en de bijbehorende risico's op het uiteindelijk model weer.

Hoe selectief de verzameling kan zijn, wordt mooi geïllustreerd door de kattenselectie van Cassy Kozyrkov, waarbij mensen per voorbeeld mogen aangeven of ze het een kat of geen kat vinden.

Afbeelding 5: Selectie van katten

De eerste vijf afbeeldingen zijn prima te beschrijven, maar bij het zesde voorbeeld moeten we onze definitie van 'kat' toch even aanscherpen. Gaat het om katachtigen, of gaat het alleen om huiskatten? Dit is een aanscherping die, bij minder duidelijke verschillen in de praktijk, niet altijd gedaan wordt.

Naast volledigheid en juistheid van data, wordt de kwaliteit ook beïnvloedt door de gemaakte selectie en interpretatie van die data. Lage datakwaliteit zorgt ervoor dat we wisselende of ronduit

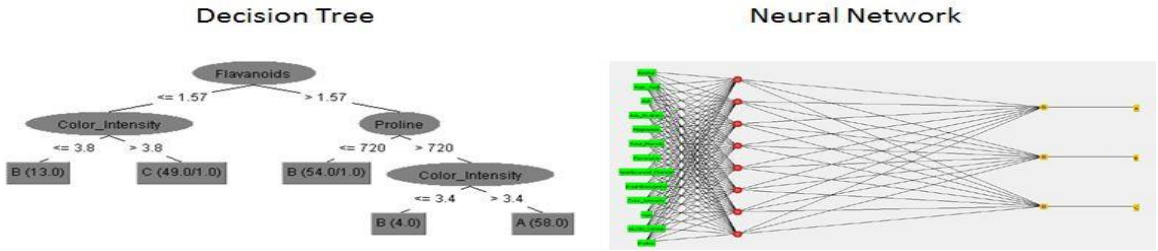
verkeerde conclusies trekken via ML. Dit zorgt voor een beperkter resultaat tijdens het live gebruiken van het ML-model.

2.3 Beperkte uitlegbaarheid

Er zijn verschillende manieren om een machine te laten leren. Een bekende variant is een beslisboom (decision tree), zie de linkerkant van de afbeelding hieronder. De beslisboom kan van boven (de stam) voor ieder voorbeeld steeds verder naar beneden worden gevolgd, terwijl bij iedere stap naar beneden een keuze wordt gemaakt op basis van de kenmerken van het voorbeeld. Bijvoorbeeld: mensen boven de 36 jaar nemen het linkerpad, mensen van 36 jaar of jonger nemen het rechterpad. Zo probeert het AI model op basis van de beschikbare inputvariabelen een onderscheid te maken tussen twee of meer uitkomsten die we willen voorspellen. Deze boom kan met ML worden opgesteld. Ook nu zal het model leren door steeds een kleine aanpassing te maken en te beoordelen of dit een betere of slechtere voorspelling geeft. Het model dat na alle kleine aanpassingen het beste resultaat geeft, wordt gebruikt om voor nieuwe situaties een voorspelling te doen.

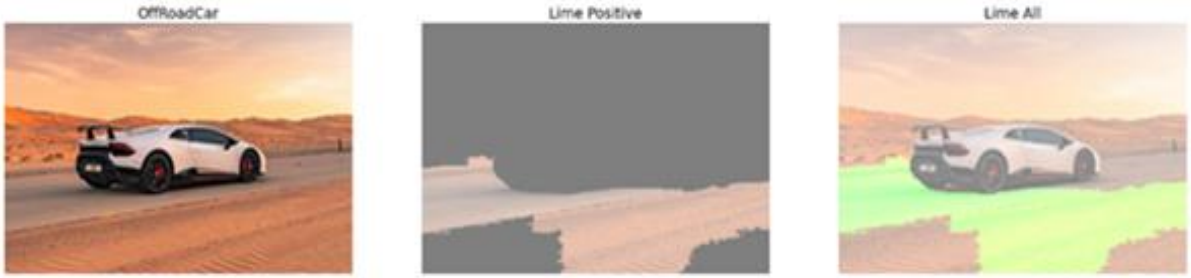
Het voordeel van deze beslisboom is dat we bij alle getrainde en nieuwe situaties kunnen aangeven *waarom* een bepaalde voorspelling wordt gedaan. Je kunt immers de beslisboom van boven naar beneden volgen tot je bij de voorspelling komt.

Zie de rechterkant van de afbeelding hieronder voor de uitingsvorm van het neurale netwerk, die het meest wordt gebruikt. Deze modellen zijn bijzonder complex. En hoewel ze zeer nauwkeurige voorspellingen kunnen geven, kunnen we als mens moeilijk of niet meer uitleggen hoe de voorspelling tot stand komt.



Afbeelding 6: Decision Tree - Neuraal Netwerk

Eén van de technieken waar een voorspelling via DL soms nog wél verklaard kan worden, is bij het herkennen van afbeeldingen. Onderstaand neuraal netwerk was getraind om standaard personenauto's van offroad rally auto's die rijden in de Parijs - Dakar race van elkaar te onderscheiden. Ook al had het model een nauwkeurigheid van 95% op de test afbeeldingen, bij het analyseren bleek dat het model zich vooral baseerde op de achtergrond; zand uit de woestijn of geen zand.



Afbeelding 7: Foutieve herkenning van een auto op basis van de omgeving

Een mens zou voor zijn of haar voorspelling gekeken hebben naar de auto zelf, bijvoorbeeld naar de wielen en niet naar de achtergrond zoals in bovenstaand voorbeeld. We denken dus al snel dat een machine dezelfde manier van redeneren gebruikt, met een aanzienlijk risico dat deze aanname onjuist is. Het is daarom van belang dat de wijze waarop een ML model tot een voorspelling komt transparant is om deze 'redenatie' te kunnen beoordelen.

2.4 Veranderende werkelijkheid of behoefte

Een stuk software is nooit af. Er zijn altijd nieuwe eisen en wensen of nieuwe inzichten waarop moet worden ingespeeld. IT-oplossingen met een AI-component vormen daar geen uitzondering op. Zeker bij ML-toepassingen, waar de data de bron is, is regelmatige aanpassing van belang.

Denk opnieuw aan het voorbeeld van de huizenprijzen; deze veranderen continu en soms zeer sterk. De kans dat het model een dag nadat het is getraind nog goed bruikbaar is, is vrij groot. De kans dat dit na een maand nog steeds zo is, is al een stuk kleiner en na een jaar is de kans vrijwel nihil. Ergens zal er dus een moment moeten worden gekozen om een nieuw model te trainen.

Verder moet er een bewuste afweging worden gemaakt in de hoeveelheid historie die wordt gebruikt om het model te trainen. Het heeft geen zin om het model voor huizenprijzen opnieuw te trainen met de data van een jaar geleden. Het liefst gebruik je alleen de meest actuele data, bijvoorbeeld die van de afgelopen week. Maar dan is er een reële kans dat er onvoldoende of alleen beperkt representatieve data beschikbaar is.

Stel echter dat je de historie van de afgelopen maand wilt gebruiken, dan bestaat de kans dat niet alle data op dezelfde manier is verzameld. Misschien is er vorige week afgesproken dat kelders juist wel, of juist niet meer meenemen met het aantal slaapkamers. Hoe ga je om met die verschillen? En dit is slechts een voorbeeld dat nog redelijk eenduidig kan worden opgelost, mits bekend is welke huizen een kelder hebben. Het is ook mogelijk dat de manier van data verzamelen grote impact heeft op de bruikbaarheid van historische data, of dat het zelfs helemaal niet bekend is dat er een verschil is in hoe de data is verzameld. Om daarvan een voorbeeld te geven; misschien hebben we de postcode-score recent volledig aangepast, of heeft men zonder medeweten van het projectteam de kelders in de telling verwijderd of toegevoegd.

Het vraagt waakzaamheid, inspanning en creativiteit om met een AI-model te blijven aansluiten bij de werkelijkheid. Er is zeker een risico dat dit geen of te beperkte aandacht krijgt.

2.5 Algemene angst AI

Nieuwe technologische ontwikkelingen worden meestal met het nodige wantrouwen ontvangen. Dat was zo bij de komst van computers, de komst van internet, enzovoort. Vaak is dat erg begrijpelijk. Wie had immers kunnen voorspellen wat de impact van computers en internet op de samenleving zou worden? Ook de mogelijke impact van ontwikkelingen zoals blockchain, internet of things, quantum computing en natuurlijk AI is nog niet te overzien.

De algemene angst komt deels doordat mensen niet weten wat AI is en wat het kan. Verhalen over superintelligentie, waarin computers slimmer zijn dan mensen, helpen daar niet bij. De huidige ontwikkeling van ML staat echter volledig los van zelfdenkende robots. Daadwerkelijk doelen formuleren, vervolgens alle stappen naar dat doel bedenken en deze in de juiste context uitvoeren zijn met de huidige ML-technieken onmogelijk. Het is echter wel wat de term AI, kunstmatige intelligentie, suggereert. De vooroordelen rond AI zullen daardoor nog lang blijven bestaan, zeker omdat ook de techniek achter ML moeilijk aan iedereen uit te leggen is.

Er bestaat ook algemene angst voor AI bij mensen die juist wél weten wat de techniek doet. Doordat er nog volop met AI wordt geëxperimenteerd en het te pas en te onpas wordt toegepast, gaat deze toepassing nog regelmatig mis. Hier zijn talloze voorbeelden van; chatbots die beledigende taal aanleren, agenten die steeds naar dezelfde wijk worden gestuurd op basis van verwachte criminaliteit, mannen die vaker als manager worden geselecteerd dan vrouwen, enzovoort.

Tenslotte is er ook angst voor het onbeperkte gebruik van data om tot conclusies te komen. Persoonlijke data wordt immers overal verzameld en sommige organisaties gaan hier wellicht te voortvarend mee om. Zeker als dit overheidsorganisaties zijn, waarbij je als burger geen keuze hebt om je data af te staan, zorgt dit voor behoorlijke angst.

Als nuance; de verschijningsvorm is hier van belang! Als AI wordt gebruikt op een bekende plek, het een bekende toepassing is en je zelf kunt instemmen om het te gebruiken, is het doorgaans geen probleem. Denk bijvoorbeeld je Apple Carplay of je (gezichtsbewerkings)programma's op je tablet. De risicoafweging blijft maatwerk.

2.6 Nieuwe uitdagingen voor testers

De algemene risico's in dit hoofdstuk laten zien dat het belangrijk is om zorgvuldig te werk te gaan bij het implementeren van AI, of het gebruik ervan zelfs maar te overwegen. Testers zijn gewend om te denken vanuit risico's. Sinds de komst van de eerste computerprogramma's is het testen uitgegroeid tot een volwaardig vakgebied. Met de komst van AI zal het testvak opnieuw veranderen.

De ontwikkeling van AI-modellen gaat echter ook in hoog tempo door. Voor een belangrijk deel gaat deze ontwikkeling via toepassingen in de praktijk. Het wordt daarbij steeds makkelijker om een model te trainen. Nu al kan met 10 regels code een model getraind worden. Via technieken zoals AutoML wordt dit nog toegankelijker; het trainingsprogramma zoekt zelf het AI-model en de bijbehorende parameters die waarschijnlijk het beste bij de data passen.

Het gat tussen het uitdenken van een idee voor een AI-toepassing en de daadwerkelijke implementatie ervan is dus vele malen kleiner dan bij een regelgestuurde toepassing. Het blijft een feit dat ook voor ook AI-toepassingen alle kwaliteits- en testactiviteiten nodig zijn. Door de eerder vermelde snelle ontwikkelingen en eenvoudigere bereikbaarheid van de development mogelijkheden van AI zal dit sterk toenemen. Het testen van deze toepassing kan niet achterblijven. Niet alleen blijft er een behoefte om de functionaliteit te testen, maar ook andere aspecten zoals ethiek met uitlegbaarheid, aansprakelijkheid en niet discriminerend, zijn van belang. Indien dit niet meegroeit zal het vertrouwen in AI-toepassingen snel verdwijnen zodra de gebruikers of algemene opinie regelmatig met missers wordt geconfronteerd. Men kan stellen dat voor de verdere groei en acceptatie van AI een verdere groei in het testen van AI toepassingen noodzakelijk is.

3 Verschijningsvormen van AI

In dit hoofdstuk worden de verschillende verschijningsvormen van AI besproken, zoals beeldherkenning en spraakgeneratie. Er is een selectie gemaakt van de risico's uit het vorige hoofdstuk. Niet alle risico's worden bij elke vorm besproken, hoewel ze bijna allemaal van toepassing zijn. Sommige risico's spreken meer tot de verbeelding of zijn beter uitlegbaar bij de ene vorm dan bij de andere. Daarop is de selectie gebaseerd.

3.1 Patroonherkenning in datasets

Bedrijven en overheden gebruiken het herkennen van patronen in datasets steeds vaker. Zij hebben in de loop der jaren grote datasets opgebouwd met heel veel informatie over hun klanten of burgers. Door de komst van AI zien ze de mogelijkheid om hier meerwaarde uit te halen.

Het meest zichtbare resultaat van patroonherkenning zijn interacties met klanten waarbij voorspeld wordt welk product de klant graag wil hebben op basis van eerdere aankopen en interesses van deze klant. Ook wordt AI ingezet om social media posts te vinden waar de organisatie graag op wil reageren, of om te herkennen welke klant de grootste kans heeft om weg te gaan. Daarnaast zijn er nog meer interne processen waar AI voor wordt gebruikt. Denk bijvoorbeeld aan het beoordelen van een kredietaanvraag, het constateren van fraude, het voorspellen van benodigd onderhoud aan voertuigen of simpelweg het categoriseren van emails op basis van inhoud.

Het herkennen van patronen in grote datasets is dermate complex dat dit menselijkerwijs niet meer mogelijk is. Met behulp van ML is het wel mogelijk deze patronen zichtbaar te maken en een resultaat of categorie te voorspellen. Dit leidt tot nieuwe bruikbare informatie, maar heeft een keerzijde, want het is moeilijk uitlegbaar hoe men aan deze voorspelling is gekomen.

Deze beperkte uitlegbaarheid heeft ook gevolgen. Indien een persoon van bijvoorbeeld fraude wordt beschuldigd, maar de reden daarvoor niet is uit te leggen, mag dat wettelijk niet worden geaccepteerd.

> *Voorbeeld*
>
> Dit was ook een reden voor de rechtbank om het Nederlandse SyRI systeem te verbieden. Het SyRI systeem was een initiatief van het ministerie van Sociale Zaken en werd gebruikt in Rotterdam, Eindhoven, Haarlem en Capelle aan den IJssel. Het had tot doel om fraude in belastingen, toeslagen en uitkeringen tegen te gaan. Er werden grote hoeveelheid gegevens gedeeld en geanalyseerd om vervolgens risicomeldingen doen. Dit zijn meldingen waarmee iemand in verband wordt gebracht met mogelijke fraude.
>
> De rechtbank vond dat de staat nieuwe technologische mogelijkheden moet benutten om fraude te voorkomen en te bestrijden, maar dat er wel een juiste balans moet zijn tussen de voordelen en het recht op respect voor het privéleven. Hoe gegevens worden verwerkt en geanalyseerd is niet transparant, zei de rechtbank. Hiermee ontstond er een risico op "onbedoeld discriminerende en/of stigmatiserende effecten".

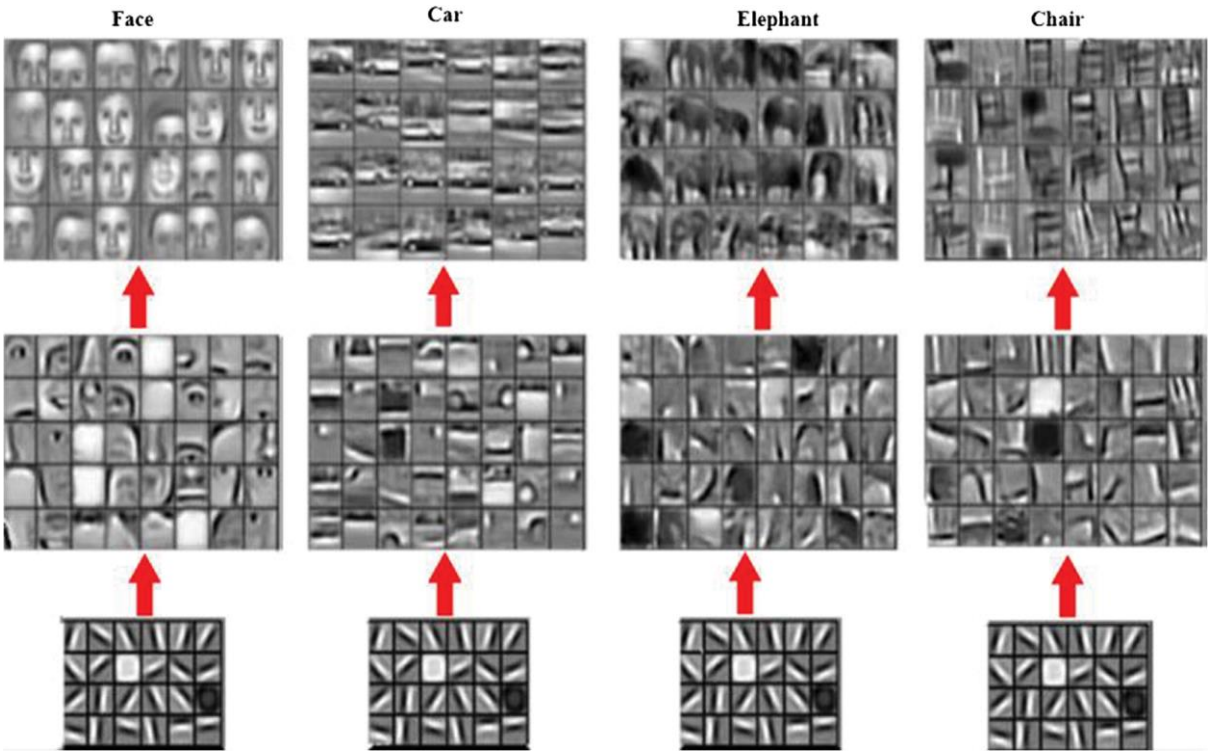
De algemene risico's uit het vorige hoofdstuk spelen een rol bij patroonherkenning in datasets. Door de omvang van de datasets is er een grote afhankelijkheid van deze data, een grote kans dat niet alle data onder dezelfde omstandigheden zijn verzameld (met daardoor last van een veranderende werkelijkheid) en is er beperkte uitlegbaarheid. Bovendien wordt deze data gebruikt voor het trainen

van AI-modellen, terwijl het niet met dat doel verzameld is. Klanten zullen dat bezwaarlijk vinden, zeker als hun data voor een voorspelling wordt gebruikt die voor hen nadelig is.

3.2 Beeldherkenning

Beeldherkenning is een verschijningsvorm die het meest tot de verbeelding spreekt. Hierdoor ontstaat een grote mate van creativiteit in het bedenken van toepassingen. Iedereen kent inmiddels de gebruikelijke voorbeelden zoals het herkennen van je gezicht om de telefoon te ontgrendelen, het herkennen van objecten door zelfrijdende auto's en het herkennen van afwijkingen en ziektebeelden op medische foto's. De creativiteit reikt echter veel verder; het herkennen van de hoeveelheid afvalzakken langs de weg (voor efficiënter ophalen), het in kaart brengen van het aantal gaten in de weg en de omvang en diepte ervan, het bepalen van de conditie van fruit, het signaleren van gevaarlijke situaties op een brug en het beoordelen van schade aan auto's, huizen en broeikassen.

Beeldherkenning kan beschouwd worden als een verbijzondering van patroonherkenning, het gaat immers om een patroon van de pixels waaruit het beeld bestaat. De laatste jaren zijn er specifieke technieken ontwikkeld voor beeldherkenning. Eén van deze technieken heet Convolutional Neural Network (CNN), de werking hiervan is te zien in onderstaande afbeelding.



Afbeelding 8: De lagen van een model voor beeldherkenning

Het CNN zoekt eerst basisvormen en basiskleuren en combineert deze basisvormen tot steeds complexere vormen, totdat de complexe vormen uiteindelijk samen een resultaat zoals een kat of een auto voorspellen.

De afbeelding laat goed zien dat de meest eenvoudige vormen slechts één keer hoeven te worden getraind en daarna steeds kunnen worden hergebruikt. Modellen voor het herkennen van algemene objecten zoals mens, hond, kat, fiets en auto zijn daarom makkelijk te vinden, gratis te gebruiken en hebben meestal een nauwkeurigheid van boven de 90%.

Voor de meeste toepassingen van beeldherkenning is een algemeen model met algemene objecten niet voldoende. Er moeten eerst specifieke voorbeelden worden verzameld om het gewenste doel te bereiken. Hierdoor kunnen de herkenningspercentages een stuk lager uitvallen, zeker als net wordt gestart met trainen.

Bovendien moet er worden geïnvesteerd om iedere voorbeeldafbeelding aan de juiste uitkomst te koppelen. Dit kan gevolgen hebben voor het uiteindelijke gebruik van het model in de praktijk, afhankelijk van de impact die een verkeerde beoordeling heeft. Een vuilniszak te veel of te weinig herkennen is hooguit vervelend, een ziektebeeld te veel of te weinig herkennen kan fataal zijn. Daarom is men in de medische sector terughoudend in het gebruik van AI. Tijdens de training bleken de getrainde modellen een hoge nauwkeurigheid te hebben, maar in de praktijk worden ze niet vaak toegepast omdat het risico van falen te hoog is.

Risico's kunnen in de loop van de tijd ook door veranderd inzicht ontstaan.

> *Voorbeeld*
>
> Als voorbeeld kan hierbij de auto's van Tesla worden genoemd. Deze auto's beschikken over geavanceerde ML systemen om zelf rijden mogelijk te maken. Deze auto's zijn daarvoor uitgerust met camera's. Om de ML modellen te verbeteren heeft Tesla de mogelijkheid om via deze auto's extra informatie te verzamelen, bijvoorbeeld van specifieke verkeerssituaties of verkeersborden. Echter, dit zou in theorie ook gebruikt kunnen worden om informatie van andere objecten of personen te verzamelen. Deze mogelijkheid is voor de Chinese strijdkrachten een reden om Tesla's te weren bij hun militaire complexen.

Het risico bij deze verschijningsvorm hangt enorm af van het toepassingsgebied.

Bij het herkennen van gevaarlijke situaties op een brug is het doorgaans goed om de brugwachter te attenderen dat er 'iets' anders is dan het verwachte beeld, of het nu een persoon is of een vuilniszak. Bij een zelfrijdende auto is dit onderscheid wél weer van belang; het kan een aanleiding zijn dat de auto zelfstandig afremt op een drukke weg.

Afhankelijk van de toepassing, speelt de onzekerheid bij beeldherkenning een beperkte of grotere rol. De onzekerheid wordt vergroot doordat ook wij mensen niet altijd een eenduidige verwachting hebben, zoals we zagen bij de vraag of een tijger een kat is. AI kan worden gebruikt om dit soort onduidelijkheden aan het licht te brengen tijdens de trainingsfase. In de praktijk wil je echter de grootste twijfelgevallen geadresseerd hebben en dat zal een bewuste inspanning vragen.

3.3 Sequentieherkenning

Bij sequentieherkenning gaat het om patronen waarbij volgorde van belang is. Volgorde kan diverse vormen hebben.

Een voorbeeld waar snel een voorstelling van te maken is, is het analyseren van patronen in transacties, om zo ongebruikelijke patronen en daarmee wellicht fraude te vinden. Of de volgorde waarin men video's bekijkt, om zo de best passende volgende video voor te stellen. Op iets grotere schaal kan men een patroon vinden van seizoensinvloeden op bijvoorbeeld landbouw.

Een ander voorbeeld, dat misschien minder voor de hand ligt, is het analyseren van tekst, video, spraak of andere geluiden. Tekst is een opeenvolging van letters en woorden, video is een opeenvolging van beelden, spraak en geluiden zijn opeenvolgingen van geluidsgolven.

Bij sequentieherkenning is het bepalen van het tijdwindow van belang. Hiermee wordt bedoeld het interval tussen verschillende tijdstippen om het patroon te herkennen. Dat kan variëren van seconden tot minuten of zelfs dagen. Dit hangt van het onderwerp af, bij een telefoongesprek zal dit eerder in seconden zijn dan in minuten.

Bij het zoeken naar sequentiepatronen zal altijd naar de context moeten worden gekeken. Een persoon met een andere moedertaal kan langer over het beantwoorden van een vraag van een callcenter medewerker doen. Een ijsverkoper zal meer kleine transacties hebben dan een autoverkoper. Ook kunnen patronen tijdelijk veranderen. Denk bijvoorbeeld aan de gevolgen die de Covid-pandemie op het uitgavenpatroon heeft. Veel AI-modellen zullen opnieuw zijn getraind om hier rekening mee te houden. Het is dus van belang om het datapatroon waarop het AI-model getraind is continue te monitoren en te bewaken dat het aansluit op de werkelijkheid. Dit is afhankelijk van het soort AI-model; het is onwaarschijnlijk dat beeldherkenningssoftware in auto's wekelijks nieuwe verkeersborden of nieuwe kinderwagens hoeft te leren.

3.4 Regressie

Deze verschijningsvorm kan in de testwereld verwarring opleveren aangezien het testen op wijziging tussen softwareversies ook regressietesten worden genoemd. Regressie als ML-verschijningsvorm is een heel bekende vorm, waar het voorspellen van huizenprijzen of aandelenkoersen voorbeelden van zijn.

Veel ML toepassingen doen op basis van het ontdekte patroon een voorspelling voor een zekere waarde op een zeker moment. Deze voorspelling is niet voor 100% zeker. Alleen in tegenstelling tot de vorige verschijningsvormen wordt er nu niet met de kans van 90% zekerheid voorspeld dat een aandeel bijvoorbeeld volgende week € 10,00 waard is, maar met een bandbreedte. Dus een aandeel is volgende week € 10,00 waard met een bandbreedte van bijvoorbeeld € 0,50.

Zoals eerder vermeld moet om tot een bepaalde prijs te komen het ML model getraind worden op basis van data. De risico's die verbonden zijn aan data zijn in het vorige hoofdstuk uitgebreid vermeld. Bij deze categorie komt een nieuw risico naar boven, namelijk extrapolatie. Stel het huizenprijs model is gebaseerd op woningen met 2 en 4 kamers in de trainingsdata, mag je dan een prijs voorspellen voor woning met 3 kamers? Dit zou nog kunnen op basis van een gewogen gemiddelde. Bij het voorspellen van een woning met 6 kamers neemt de onzekerheid op het juist zijn van de voorspelling toe.

De huizenprijs is een eenvoudig model. Bij modellen met meer abstracte data of veel meer variabelen is het, voor ons als mens, moeilijk te bepalen wat juiste voorspellingen zijn en in welke mate het resultaat is gebaseerd op extrapolatie.

3.5 Tekstgeneratie

Met deze verschijningsvorm komt vrijwel iedereen dagelijks in aanraking. Het kan zijn via de spellingscontrole, een chatbot of bij het vertalen van een tekst in een andere taal. Deze whitepaper is zelf ook in het Nederlands én het Engels uitgebracht met behulp van Google Translate. Zelfs de Nederlandse lezer loopt grote kans dat er zinnen zijn die door Google Translate zijn gegenereerd. Uit ervaring is namelijk gebleken dat een document vertalen van het Nederlands naar het Engels én vervolgens weer terug vaak beter leesbare zinnen oplevert.

> *Voorbeeld*
>
> Een actuele ontwikkeling is het “Generative Pre Trained Transformer 3” (GPT-3) model. Dit is een door de OpenAI organisatie ontwikkeld taal model en is in staat om programmacode en zelfs poëzie te genereren. Het model bestaat uit 175 miljard parameters en voor de training gebruikte men 499 miljard tokens. Er is ingeschat dat hiervoor 700 GB geheugen en 3.14E23 FLOPS aan computercapaciteit voor nodig is geweest. Kortom, een gemiddelde PC zal daar duizenden jaren voor nodig hebben. De kosten voor deze training worden geraamd op $ 4.600.000. Dit is wederom een voorbeeld van dat er maar een beperkt aantal partijen zijn die dit kunnen uitvoeren, zowel m.b.t. de investering als met betrekking tot de benodigde computerkracht. Daarnaast heeft op dit ogenblik alleen Microsoft de exclusieve licentie gekregen op het gebruikt en de broncode.

Het formuleren van grammaticaal correcte zinnen is echter maar een deel van de uitdaging. Het programma waar de gebruiker teksten mee uitwisselt, zal die gebruiker op een consistente manier moeten benaderen. Er moet een onderscheid gemaakt worden of het om een formele tekst gaat, bijvoorbeeld een tekst die volledige juridisch correct moet zijn, of om een informele tekst, bijvoorbeeld een email naar een collega. Afhankelijk van het type gesprek dat ontstaat, moet het programma informerend, kalmerend of juist activerend zijn. Kort gezegd; het programma moet een eigen persoonlijkheid en stijl hebben om de communicatie vlot te laten verlopen.

Zo mag een zeer enthousiaste klant bijvoorbeeld niet minder enthousiast gemaakt worden door koele ambtelijke antwoorden. Het is van belang dat een chatbot dit kan onderkennen en kan adviseren het gesprek door te geven aan een callcentermedewerker of het gesprek te beëindigen. Er zijn chatbots geweest die discriminerende opmerkingen maakten of zijn gaan vloeken. Dit effect is niet alleen slecht voor het imago van het desbetreffende bedrijf, maar ook voor de acceptatie van deze techniek in het algemeen.

3.6 Spraakgeneratie

Aan het genereren van spraak wordt al decennia gewerkt. Het heeft een ontwikkeling doorgemaakt van het opzeggen van woorden tot het voorlezen van zinnen waardoor een zo goed als menselijk gesprek mogelijk is. De spraak is inmiddels nagenoeg gelijk aan hoe een echt mens het zou uitspreken. Het is daarom goed om spraakgeneratie vanuit een ethisch standpunt te bekijken.

Zo is het doorgaans geen probleem als de gebruiker de bron van de gesproken tekst zelf heeft gekozen, bijvoorbeeld bij de voorleesfunctie op websites of bij een navigatiesysteem. Ook als de bron van de tekst uit een vertrouwde omgeving komt, levert dit nauwelijks bezwaren op. Denk aan Siri, Alexa of Google Home. Of aan pratende robots voor dementerende gebruikers. In al deze gevallen heeft de gebruiker bewust gekozen voor deze toepassing.

Het wordt ethisch lastiger als de gebruiker niet zelf heeft gekozen voor de toepassing van spraakgeneratie of zelfs niet op de hoogte is van deze toepassing. Een duidelijk voorbeeld is de demo van de Google Assistant die telefonisch een afspraak maakt bij de kapper. Overigens wordt hier gebruik gemaakt van een keten van AI-toepassingen, waarin spraakgeneratie slechts het eindstation is.

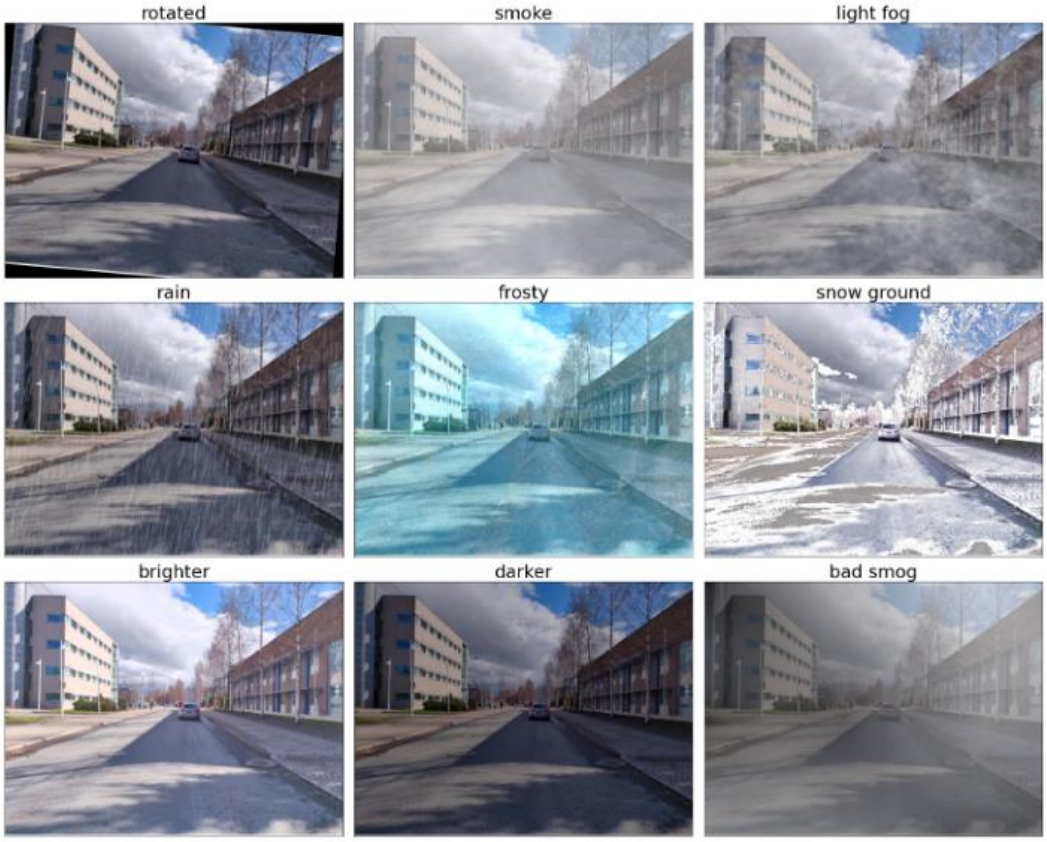
Op deze manier maakt ML een behoorlijke stap in de echte wereld en wordt het onduidelijk wanneer en hoe we met een AI te maken hebben. De ethische vraag is of dit daadwerkelijk een stap te ver is of dat wij als maatschappij aan deze techniek moeten wennen. Duidelijk is wel dat wij mensen het minimaal willen weten als er een AI met ons communiceert. Dit kan via een aankondiging en

eventueel akkoord van de menselijke gesprekspartner, maar misschien ook door de kwaliteit van de spraak weer een stuk minder menselijk te maken.

3.7 Beeldgeneratie

Met beeldgeneratie hebben we als testers nog weinig ervaring. Wel komt het regelmatig in het nieuws, vooral in de vorm van 'deep fakes' ofwel afbeeldingen en video's die levensecht lijken, maar door een AI-model zijn gegenereerd. Het kan voor twijfel zorgen over de echtheid van deze beelden en video's, of het kan juist ongemerkt worden ingezet om de publieke opinie of individuele gedachten en handelen te beïnvloeden. In het laatste geval kun je spreken over manipulatie of chantage. In die zin is er grote angst voor beeldgeneratie, juist omdat het nog zo onbekend is. Er zijn nog vele ethische, juridische en maatschappelijke risico's om te adresseren.

Toch zal ook deze techniek langzaam maar zeker een weg vinden naar reguliere, algemeen geaccepteerde toepassingen. Denk aan toepassingen voor het verbeteren of inkleuren van oude foto's of films. Ook kan bij zelfrijdende auto's een overzicht van de omgeving worden gegenereerd. Vervolgens kan een AI-model worden gebruikt om omgevingen te veranderen, bijvoorbeeld door sneeuw, schemer of mist te genereren, waar de zelfrijdende auto vervolgens op getraind wordt. Wel is het de vraag of het wenselijk is als AI-modellen van elkaar leren. In hoeverre bereid je het model immers nog voor op de échte werkelijkheid.



Afbeelding 9: Uitzicht op straat met wisselende omgevingscondities (afbeelding: Teemu Kanstrén)

Zolang dit een handige toepassing is en een controle met het origineel of de werkelijkheid mogelijk is, heeft deze toepassing een maatschappelijk toegevoegde waarde. Daarnaast is het, net als met spraakgeneratie, belangrijk dat de gebruiker wordt geïnformeerd wanneer er AI-gegenereerde beelden of video's worden getoond.

4 Gradaties in autonomie

Lang niet alle AI-modellen opereren volledig zelfstandig vanaf het moment dat ze in gebruik zijn genomen. Dit hangt in grote mate af van het vertrouwen in het model, maar ook van het risico wanneer er fouten optreden. Sommige modellen zullen nooit autonoom worden ingezet. Dit hoofdstuk zet de verschillende gradaties van autonomie op een rij.

4.1 Handmatige invoer en controle

In de meest simpele toepassingsvorm biedt iemand handmatig één of meerdere voorbeelden aan in een getraind AI model en controleren één of meerdere personen het resultaat. Dit is de meest voorzichtige variant, waarbij menselijke ogen kijken naar de kwaliteit van de invoerdata en de bruikbaarheid van de voorspelling. Deze voorzichtigheid kan komen door beide elementen waar risico's uit bestaan:

- De kans dat de data of de voorspelling niet goed is. Bijvoorbeeld omdat er nog weinig ervaring is opgedaan met de voorspelling. Ook zeer diverse invoerdata, complexiteit van de voorspelling, of meerdere voorspellingsdoelen kunnen hiervoor redenen zijn. Bijvoorbeeld een chatbot van een verzekeraar die zowel autoruitschade als autodiefstal moet afhandelen.
- De impact die het heeft als verkeerde voorspellingen worden gebruikt in de praktijk. Denk aan verlies van (gezond) leven, hoge kosten, imagoschade enzovoorts.

Handmatige invoer en controle is een 'dure' variant om AI te implementeren. Het werk dat in al deze controles gaat zitten, drijft de kosten van het gebruik van het AI-model op. Om steeds efficiënter te kunnen werken, is het gebruikelijk om op basis van ervaring een vaste set aan controles op te stellen. Dit kunnen zowel controles op de invoerwaarden als op de voorspellingen zijn.

Bij invoercontroles komt deze ervaring van experts uit de praktijk, of vanuit ervaring met het voorbewerken van de data die al eerder aan het model is aangeboden. Het optimale resultaat van deze invoercontroles is dat 'iedere' toekomstige invoer aangeboden kan worden, waarbij het model deze in alle scenario's goed afhandelt zonder menselijke tussenkomst.

Ook bij de handmatige controle op voorspellingen door het model is efficiëntie te behalen. Denk aan steekproefsgewijs controleren in plaats van een volledige controle. Er kunnen clusters worden gemaakt van samenhangende invoer- en voorspellingswaarden. In dat geval kan per cluster bekeken worden hoe groot de kans op foute voorspellingen is. Verder kan de aandacht worden verlegd naar de grensgevallen. Bijvoorbeeld: het model voorspelt met 51% zekerheid dat de ruitschade vergoed kan worden. Of juist bij de extreme zekere voorspellingen, bijvoorbeeld 99% zekerheid daar waar doorgaans 80% de maximale zekerheid van het model is.

In deze variant heeft de mens nog altijd controle over het proces en geeft de AI slechts een voorzet. Dit kan ook worden omgedraaid, vooral als men nog leert hoe de voorspelling werkt; de mens geeft de voorspelling (wel of niet verzekeren, wel of geen ziektebeeld) en de AI kijkt of dit afwijkt van de eigen voorspelling. Dit is een nuttige toepassing als je wilt voorkomen dat de mens zich laat leiden door het AI-model.

4.2 Autonoom invoeren of verwerken

Autonomie van AI is een zeer gevoelig onderwerp in de maatschappij, er zijn immers vele voorbeelden waarbij een AI-model in de praktijk niet robuust genoeg is gebleken. Bijvoorbeeld door alle data die kon worden aangeboden, of voor alle voorspellingsvarianten en -combinaties die het kon geven. Het autonoom invoeren van data in een AI-model en/of het autonoom verwerken van de voorspelling gebeurt doorgaans alleen als:

1. Er voldoende zicht is op de kwaliteit van de invoerdata én op alle variaties van invoerdata die nu en in de nabije toekomst aan de AI kunnen worden aangeboden.
2. De impact van verkeerde uitkomsten voldoende is beoordeeld, zodat we denken dat we goed kunnen omgaan met alle juiste, twijfelachtige én verkeerde voorspellingen die het model kan geven.

Het is belangrijk om na te gaan wanneer er 'voldoende' kennis is opgedaan over de impact van verkeerde voorspellingen. Bij het aanbevelen van een product in een webwinkel kan de impactbepaling, bij welke varianten het mis kan gaan, minder uitgebreid zijn dan voor een operatierobot of zelfrijdende auto. Toch zijn er ook gradaties van impact met productvoorstellen; het tonen van een 18+ product aan een kind van 6 jaar, of aanbevelingen van medische producten hebben een hogere impact dan het aanraden van tent wanneer er naar producten voor dieren wordt gekeken. In alle gevallen is het een risicobeoordeling.

Autonomie vereist een goed doordachte, goed afgewogen en goed geteste set aan controles op de invoer en de uitvoer. De steekproeven uit de vorige paragraaf zijn nog altijd mogelijk als kwaliteitscontrole achteraf, maar hebben geen risicobeperkende invloed meer op het gebruik in de praktijk. De verwerking heeft dan immers al plaatsgevonden.

Er zijn bij autonoom gebruik altijd vangnetten nodig om de belangrijkste ongewenste voorspellingen te voorkomen. Autonoom gebruik betekent namelijk dat het AI model onderdeel is van een groter geheel, een keten. In deze keten wordt minimaal de invoer ontvangen vanuit een gebruiker of ander systeem, wordt deze invoer omgezet zodat het model een voorspelling kan doen en wordt de voorspelling omgezet naar een actie. Een chatbot die op de traditionele manier geprogrammeerd is, kan redelijk veilig ingezet worden. Bij een chatbot die gemaakt is op basis van een ML model is het verstandig om vangnetten in te zetten.

Het testen van ketens is niets nieuws. Projecten hebben hier al tientallen jaren ervaring mee via standaarden zoals TMAP. Wel blijkt juist dat projecten die gestart zijn om te zien 'wat AI kan' hier nog wel eens op vastlopen, zowel op technisch vlak als op de afspraken die binnen organisaties gelden over het gebruik van data.

4.3 Een veelvoud van berekeningen

In de vorige paragraaf werd er vanuit gegaan dat een AI model een voorspelling deed en daarmee feitelijk een beslissing nam. Voor het afvangen van onlogische of weinig overtuigende voorspellingen kunnen beperkende regels ofwel controles worden ingezet.

Als er meerdere AI-voorspellingen en/of geprogrammeerde regels aan elkaar worden gekoppeld, maakt dit het nog een stuk complexer. Denk aan een systeem dat gesproken vragen beantwoordt met gesproken tekst. Dit bevat diverse subsystemen zoals spraak-naar-tekst, interpretatie-van-tekst, genereren-van-het-antwoord en tekst-naar-spraak. Of iets complexer: vraag aan Siri of Google

Assistant naar de dichtstbijzijnde supermarkt, waarbij ook geolocatie en een kaartapplicatie moeten worden meegenomen.

Hoe complexer de beslissing, hoe groter het belang om alle losse componenten goed te testen. Met andere woorden; bedenk eerst welke varianten er per component mogelijk zijn en stel vast dat alle controles goed werken, voordat het onderdeel geactiveerd wordt en een resultaat oplevert voor de volgende stap in de keten. Nadat alle componenten zijn getest, is het tijd om de keten als geheel te testen. Het nadeel van een groot aantal componenten is dat de keten op een groot aantal manieren kan 'falen' en de afhandeling daarvan op een juiste manier moet gebeuren. Gelukkig kan bij het testen van componenten en ketens gebruik gemaakt worden van de in de vorige paragraaf genoemde teststandaarden.

4.4 Aansturing van machines

Bij het toevoegen van fysieke machines in een keten van algoritmes treden andere, deels nieuwe overwegingen op. Het geheel van algoritmes en machines wordt aangeduid als een robot. Dit kan uit meerdere lagen bestaan, bijvoorbeeld de AI die de beslissingen neemt, tussenliggende systemen die de beslissingen vertalen naar acties en de fysieke hardware die de feitelijke acties uitvoert. Bij het testen moet met elk van deze lagen rekening gehouden worden.

Een machine kan vaak een groot aantal handelingen uitvoeren. Van al deze handelingen voorspelt een AI-model of dit de juiste is, waarbij de hoogste score doorgaans 'wint'. Ook hier zullen we echter meerdere scenario's moeten naspelen. Stel dat twee handelingen een nagenoeg gelijke waarde hebben, welke moet dan winnen? Moet in dat geval wel gekozen worden voor één van deze handelingen? En wat als de hoogst voorspelde handeling slechts 35% zekerheid scoort?

Dit soort overwegingen komen ook terug bij AI-modellen en bij ketens die geen fysieke machines aansturen. Toch is er een belangrijk verschil; bij een stuk software is er doorgaans een 'plan B' beschikbaar; als je geen goede film kunt aanbevelen, kun je altijd nog de meest bekeken film aanbevelen. Als je niet weet of een schade kan worden behandeld, kun je aangeven dat een medewerker ernaar gaat kijken. Dit soort veilige opties zijn vaak lastig bij fysieke machines. De machines zijn soms gedwongen een keuze te maken. Als deze keuze gevolgen heeft voor een of meer mensen kan dit ethische consequenties hebben. Hierbij speelt mee dat de keuzes en het komen tot een uiteindelijk resultaat bij het gebruik van ML vaak niet transparant of uitlegbaar is. Daarom worden dit soort ethische keuzes niet graag aan een AI-algoritme overgelaten. Ethiek is dan ook een zeer belangrijk onderwerp binnen kwaliteit en testen van AI.

5 Ethische richtlijnen

De relatie tussen AI en ethiek is al eerder ter sprake gekomen. De risico's van beperkte uitlegbaarheid en angst voor AI hebben voor een gedeelte een ethische achtergrond. Bij het ontwikkelen en testen van AI applicaties kan het onderwerp ethiek nieuwe gezichtspunten opleveren.

Er is een aantal organisaties uit zowel de publieke als private sector die de laatste jaren ethische richtlijnen hebben opgesteld. In dit hoofdstuk worden de ethische richtlijnen en de concept-regelgeving van de EU als uitgangspunt gehanteerd. Dit hoofdstuk moet gezien worden als een samenvatting van deze regelgeving en de onderliggende uitgangspunten, zonder verdere interpretatie.

Aangezien ethiek voor velen een nieuw terrein is, wordt gestart met een korte paragraaf over AI en ethiek. Na het lezen van deze paragraaf zijn de vervolgparagrafen beter te duiden.

5.1 AI en ethiek

Deze paragraaf is gebaseerd op de paper: 'AI heeft geen stekker meer' van Rudy van Belkom. Er is voor dit document gekozen omdat het de relatie tussen AI en ethiek bespreekt op basis van een studie van een groot aantal bronnen, en toch nog makkelijk leesbaar is.

5.1.1 Ethiek

Wat 'het goede' is en hoe 'juist te handelen' zijn vragen die de mens al eeuwen bezighoudt. Ethiek is een tak binnen de filosofie die zich bezighoudt met dit vraagstuk. Wanneer het over ethische vraagstukken gaat dan vinden we er allemaal wel 'iets' van. Daarnaast vinden mensen verschillende dingen belangrijk en is dit ook vaak afhankelijk van de context.

Binnen de ethiek bestaan er dan ook verschillende stromingen en opvattingen die soms lijnrecht tegenover elkaar staan. In sommige gevallen ligt de nadruk op de handeling zelf (beginselethiek) en in andere gevallen meer op de consequenties die het handelen heeft (gevolgenethiek). En soms ligt de nadruk juist op de intentie van degene die de handeling uitvoert (deugdethiek).

5.1.2 Ethiek in relatie met AI

Voorheen had ethiek voornamelijk betrekking op het menselijk handelen. Met de komst van AI is er een nieuwe speler in het spel, namelijk de zelflerende technologie. Naarmate de technologie autonomer wordt, ontstaan er nieuwe ethische vraagstukken. Denk aan de vraagstukken met betrekking tot uitlegbaarheid ('explainability'), privacy en vooringenomenheid ('biases'). De AI-technologie draagt bij aan beslissingen en de beslisregels van deze technologie zijn lastig te herleiden. Dit is nieuw terrein voor de mens en zorgt logischerwijs voor gevoelens van angst.

AI is op sommige vlakken al beter dan de mens, met name wanneer het gaat om zeer specialistische toepassingen. Zo blijken algoritmes nu al beter in staat om vormen van kanker te herkennen op longfoto's dan artsen. Er kan de vraag gesteld worden of het in deze situaties nog wel verantwoord is om bepaalde taken door mensen uit te laten voeren. En tevens of in deze situaties mensen de keuze zouden moeten krijgen om een AI te verkiezen boven de mens.

5.1.3 Transparantie en rechtvaardigheid

In alle discussies over ethiek in relatie met AI zijn er twee punten die steeds naar boven komen, namelijk transparantie en rechtvaardigheid.

- Naarmate AI steeds autonomere beslissingen kan gaan nemen, is het van belang dat het inzichtelijk is op welke wijze de beslissingen van dergelijke systemen tot stand zijn gekomen.
- Wanneer het over rechtvaardigheid gaat, dan gaat het in de basis over de gelijkwaardigheid van mensen. Mensen dienen gelijk behandeld te worden én gelijke kansen te krijgen. Dit betekent niet dat er geen verschillen tussen mensen kunnen of mogen zijn. Maar wanneer mensen verschillend behandeld worden moet hier een duidelijk aanwijsbare reden voor zijn om het verschil te kunnen rechtvaardigen.

AI-toepassingen worden in verschillende domeinen gebruikt, zoals bedrijfsleven, overheid, veiligheid en de medische wereld. Hierdoor is het niet eenvoudig om duidelijke en overkoepelende ethische kaders op te stellen. Wat ‘goed’ of ‘juist’ is hangt steeds af van de specifieke context. Daarbij maken de verschillen in ethische stromingen en opvattingen het er ook niet eenvoudiger op. Het blijft een democratisch vraagstuk om tot consensus te komen over wat de beste beslissing is voor de samenleving is. De discussie om tot deze consensus te komen, wordt vaak gevoerd op basis van praktische voorbeelden, zoals onderstaand voorbeeld van Uber.

> *Voorbeeld van transparantie in de algoritmes*
>
> Uber laat zijn algoritmes bepalen welke chauffeurs moeten worden ontslagen. Daarvan beschuldigen een aantal chauffeurs het bedrijf. De chauffeurs zouden zijn ontslagen nadat de Uber-algoritmes hadden uitgewezen dat ze zich schuldig maakten aan 'frauduleuze activiteiten', terwijl ze deze activiteiten zelf ontkennen.
>
> Geautomatiseerd mensen ontslaan mag niet, stellen de chauffeurs. Volgens de Britse en de Europese privacywet moet een mens zo'n ingrijpend besluit nemen en mag dat niet afhangen van de uitkomst van een algoritme. Uber ontkent dat er geautomatiseerd is besloten welke chauffeurs moeten vertrekken, maar volgens de chauffeurs is er geen sprake van 'betekenisvolle menselijke tussenkomst'. Wat wel vereist is: iemand van vlees en bloed moet meer doen dan een advies van een algoritme volgen.
>
> In februari 2021 is Uber bij verstek door de rechtbank van Amsterdam bevolen om een chauffeur weer toe te laten op het platform. Doordat het gaat om een verstekvonnis - wat betekent dat Uber niet aanwezig is geweest bij de zitting - is er door de rechter niet inhoudelijk naar de zaak gekeken.
>
> In een andere zaak wilden chauffeurs weten welk oordeel de algoritmes van Uber over hen vellen en welk effect dat heeft op de manier waarop ritten tussen chauffeurs worden verdeeld. Op basis van de Europese privacywetgeving hebben ze recht om dit soort informatie te krijgen. De privacywet AVG laat inwoners van EU-landen opvragen wat bedrijven en instanties van hen weten, maar ook wat er vervolgens met die gegevens gebeurt. Zo moeten ze kunnen zien hoe hun gegevens worden behandeld en of er op basis van die gegevens geautomatiseerd besluiten worden genomen. Dit voorbeeld toont aan dat transparant en uitlegbaar gebruik van AI een belangrijke voorwaarde is om de juiste discussies over ethiek te voeren.

5.2 Ethische richtlijnen van de EU

Binnen Europa houdt de “High-Level Expert Group on Artificial Intelligence” zich bezig met ethiek. Hun document “Ethics Guidelines for Trustworthy AI” stelt het volgende:

> *“Bij AI-systemen moet de mens centraal staan en moet ernaar worden gestreefd deze systemen te gebruiken in dienst van de mensheid en het algemeen belang, met als doel de verbetering van het welzijn en de vrijheid”*

Hieruit blijkt dat de EU de mogelijkheden van AI-systemen onderkent en ook wil stimuleren. Men heeft als doel om de voordelen van het gebruik van deze systemen te optimaliseren en tegelijkertijd de risico’s van het gebruik te minimaliseren. Vertrouwen en betrouwbaarheid zijn randvoorwaarden om dit doel te bereiken en de expertgroep heeft een kader samengesteld waar gedurende de volledige levenscyclus van het systeem aan moet worden voldaan. Dit kader bestaat uit de volgende drie componenten:

5.2.1 Robuuste AI

Robuustheid is een bekend kwaliteitskenmerk dat ook is opgenomen in de ISO 25010 norm. In het kader van AI geeft men een bredere invulling aan dit kenmerk. Men stelt dat een AI-systeem robuust moet zijn en niet enkel vanuit technisch oogpunt, maar ook vanuit een sociaal oogpunt. Dit is zo beschreven omdat AI-systemen ongewild schade kunnen aanrichten, zelfs al zijn de bedoelingen goed. Om deze reden moeten er voorzorgsmaatregelen worden getroffen om onbedoelde negatieve gevolgen te voorkomen.

5.2.2 Wettige AI

Het AI-systeem moet wettig zijn, door te zorgen dat aan alle toepasselijke wet- en regelgeving wordt voldaan. De wetgeving bevat zowel positieve als negatieve verplichtingen. Dat betekent dat deze niet alleen moet worden uitgelegd in verband met wat er niet mag, maar ook in verband met wat er moet. Men baseert zich mede op een aantal grondrechten, zoals:

- Respect voor de menselijke waardigheid
- Vrijheid van het individu
- Respect voor democratie, justitie en de rechtsstaat
- Gelijkheid, non-discriminatie en solidariteit

Dit component wordt in de toekomst ondersteund door aanvullende wetgeving.

5.2.3 Ethische AI

Het AI-systeem moet ethisch zijn, door te zorgen dat de ethische beginselen en waarden worden nageleefd. Deze ethische beginselen komen voort uit de bovengenoemde grondrechten en zijn als volgt benoemd:

Het beginsel van respect voor de menselijke autonomie

Mensen die met AI-systemen werken, moeten hun volledige en effectieve zelfbeschikking kunnen behouden. AI-systemen mogen mensen niet onterecht onderwerpen, dwingen of misleiden.

Het beginsel van preventie van schade

AI-systemen mogen geen schade veroorzaken of vergroten of anderszins negatieve gevolgen hebben voor mensen. Ze moeten technisch robuust zijn en er moet voor worden gezorgd dat ze geen ruimte bieden voor kwaadwillig gebruik.

Het beginsel van rechtvaardigheid

Personen en groepen moeten vrij zijn van onrechtvaardige vertekening, discriminatie en stigmatisering. Daarnaast mag het gebruik van AI-systemen nooit tot gevolg hebben dat de gebruikers worden misleid of worden beperkt in hun keuzevrijheid. Ook valt onder deze dimensie de mogelijkheid om beslissing die genomen worden door AI-systemen aan te vechten.

Het beginsel van verantwoording

Verantwoording is cruciaal voor het scheppen en behouden van het vertrouwen van gebruikers in AI-systemen. Dat betekent dat processen transparant moeten zijn, het doel van het systeem kenbaar moeten worden gemaakt en dat beslissingen, voor zover mogelijk, verklaarbaar moeten zijn aan degenen die er direct of indirect de gevolgen van ondervinden.

5.3 De concept-regelgeving van de EU met betrekking tot AI

Veel AI-systemen hebben een beperkt tot verwaarloosbaar risico en kunnen worden gebruikt om veel maatschappelijke uitdagingen op te lossen. Echter creëren bepaalde AI-systemen risico's die moeten worden aangepakt om ongewenste resultaten te voorkomen. Zo kan de ondoorzichtigheid van algoritmen tot onzekerheid leiden en een obstakel vormen voor de effectieve handhaving van de bestaande wetgeving over veiligheid en grondrechten.

Voor bedrijven kan dit tot rechtsonzekerheid leiden en als gevolg van het gebrek aan vertrouwen de invoering van AI-technologieën vertragen.

De Europese Commissie heeft in april 2021 voorstellen gepresenteerd voor wettelijke regelgeving om misbruik van AI tegen te gaan. Dit is een vervolg op de Ethische richtlijnen van de EU. Deze concept regelgeving komt ook met een uitgebreid assessment om de risico’s vast te stellen en de vereisten te onderkennen.

De Commissie stelt een risico gebaseerde aanpak voor met vier risiconiveaus, namelijk:

5.3.1 Onaanvaardbaar risico (Unacceptable risk):

Hieronder vallen alle AI-systemen die in strijd zijn met de EU-waarden omdat ze de grondrechten schenden. Systemen die worden beschouwd als een duidelijke bedreiging voor de veiligheid, het levensonderhoud en de rechten van mensen, zullen worden verboden.

5.3.2 Hoog risico (High-risk):

AI-systemen die als hoog risico worden aangemerkt, omvatten AI-technologie die wordt gebruikt in:

- Kritieke infrastructuren (bijvoorbeeld vervoer) die het leven en de gezondheid van burgers in gevaar kunnen brengen

- Onderwijs- of beroepsopleiding, die de toegang tot onderwijs en de professionele loop van iemands leven kan bepalen (bijvoorbeeld het scoren van examens)
- Veiligheidscomponenten van producten (bijv. AI-toepassing in chirurgie ondersteund door robots)
- Werk, personeelsbeheer en toegang tot zelfstandig ondernemerschap (bijvoorbeeld CV-sorteersoftware voor wervingsprocedures)
- Essentiële particuliere en openbare diensten (bijvoorbeeld een kredietscore waardoor burgers de mogelijkheid worden ontzegd om een lening te krijgen);
- Wetshandhaving die inbreuk kan maken op de grondrechten van mensen (bijvoorbeeld evaluatie van de betrouwbaarheid van bewijs)
- Migratie-, asiel- en grenscontrole beheer (bijvoorbeeld verificatie van authenticiteit van reisdocumenten)
- Rechtsbedeling en democratische processen (bijvoorbeeld toepassing van de wet op een concrete reeks feiten)

Voor de AI-systemen die in deze categorie vallen worden bindende eisen voorgesteld voordat ze op de markt kunnen worden gebracht. Die eisen hebben betrekking op de kwaliteit van:

- gebruikte datasets
- documentatie en gegevensregistratie
- transparantie en informatieverstrekking aan gebruikers
- menselijk toezicht
- robuustheid
- nauwkeurigheid
- cyberbeveiliging

5.3.3 Beperkt risico (Limited risk):

De AI-systemen die in deze categorie vallen hebben een verplichting op transparantie. Bij het gebruik van AI-systemen zoals chatbots moeten gebruikers zich ervan bewust zijn dat ze interactie hebben met een machine.

5.3.4 Minimaal risico (Minimal risk):

AI-systemen die in deze categorie vallen, kunnen zonder bijkomende wettelijke verplichtingen conform de bestaande wetgeving worden ontwikkeld en gebruikt.

De grote meerderheid van de AI-systemen die op dit moment in de EU worden gebruikt, zoals op AI-gebaseerde videogames of spamfilters, behoren tot de laatste categorie met een minimaal risico. Aanbieders van die systemen kunnen ervoor kiezen de vereisten voor betrouwbare AI toe te passen en niet-bindende gedragscodes na te leven. De commissie zal brancheorganisaties en andere representatieve organisaties aanmoedigen om vrijwillige gedragscodes vast te stellen.

5.4 Verwezenlijking van betrouwbare AI

Voor de controle op betrouwbare AI-systemen is een checklist en een online assessment-tool beschikbaar. In bijlage A is een verwijzing naar deze hulpmiddelen opgenomen. Deze kunnen helpen bij het identificeren van mogelijke risico's van AI-systemen en bij het bepalen of, en wat voor soort maatregelen moeten worden genomen om die risico's te verkleinen. Eveneens wordt een aantal methoden waaronder het testen en valideren beschreven.

De expertgroep van de Europese Commissie stelt dat het testen en valideren van een systeem zo vroeg mogelijk moet gebeuren om te zorgen dat het systeem zich gedurende de volledige levenscyclus en met name na de installatie gedraagt zoals bedoeld. Vanwege het karakter van AI-systemen is traditioneel testen niet voldoende. Om de verwerking van gegevens te controleren en te valideren moet het onderliggende model daarom tijdens zowel de training als de installatie zorgvuldig worden gemonitord op stabiliteit, robuustheid en werking binnen heldere en voorspelbare grenzen. Er moet worden gezorgd dat het resultaat van het planningsproces consistent is met de invoer en dat de beslissingen op dusdanige wijze worden genomen dat het onderliggende proces kan worden gevalideerd.

Alle componenten van een AI-systeem moeten erin worden opgenomen, met inbegrip van gegevens, vooraf getrainde modellen, omgevingen en het gedrag van het systeem als geheel. Het systeem moet worden ontworpen en uitgevoerd door een zo divers mogelijke groep mensen, zodat tijdens het trainen voldoende aandacht is voor de verschillende mogelijke invalshoeken en achtergronden.

Er moeten meerdere maatstaven worden ontwikkeld voor de categorieën die vanuit verschillende oogpunten worden getest. Hier gaat het bijvoorbeeld om het inzetten van ethische hackers. Dit testen vanuit andere invalshoeken staat bekend als adversarial testen. Tot slot moet er worden gezorgd dat de resultaten of handelingen overeenkomstig zijn met de resultaten van de voorgaande processen, door ze te vergelijken met het eerder vastgestelde beleid, zodat dit niet wordt geschonden.

Bovenstaande beschrijving van de methode testen en valideren is overgenomen uit de richtlijnen van de EU. Zo wordt duidelijk dat ook vanuit de EU het belang wordt gezien om grondig, rekening houdend met de ethische aspecten en vanuit verschillende achtergronden en invalshoeken, te testen. De richtlijnen geven een goed podium aan kwaliteit en testen van AI.

6 Kwaliteitsattributen

Kwaliteitsattributen (ook bekend als kwaliteitskenmerken) kunnen gebruikers en belanghebbenden helpen met het bepalen welke kenmerken van kwaliteit van belang zijn voor softwaresystemen en dus ook voor AI-systemen.

De ISO 25010 is een bekende standaard voor kwaliteitsattributen. De attributen van deze standaard kunnen ook gebruikt worden voor AI-systemen, maar zijn nog niet volledig dekkend voor AI. Verschillende partijen hebben hun gedachten laten gaan, of zijn dit nog aan het doen, om (sub-) attributen toe te voegen om zo een standaard te krijgen die ook dekkend is voor AI-systemen.

6.1 De huidige ISO 25010 standaard

Voor de juiste werking van software is niet alleen de functionaliteit van belang. Andere aspecten zoals performance, gebruiksvriendelijkheid, stabiliteit en onderhoudbaarheid zijn ook belangrijk.

De ISO 25010 standaard heeft een aantal van deze aspecten, de zogenaamde kwaliteitsattributen in een framewerk samengebracht.

Met betrekking tot productkwaliteit onderkent deze standaard de volgende attributen:

Attribuut	Omschrijving
Functionality (Functionaliteit)	De mate waarin voldaan wordt aan de uitgesproken en veronderstelde behoeften.
Performance (Prestatie)	De prestaties in verhouding tot de hoeveelheid middelen gebruikt onder genoemde condities.
Compatibility (Uitwisselbaarheid)	De mate waarin een systeem informatie uit kan wisselen of uitvoeren op andere omgevingen.
Usability (Bruikbaarheid)	De mate waarin een systeem gebruikt kan worden om gespecificeerde doelen te bereiken.
Reliability (Betrouwbaarheid)	De mate waarin een systeem functies uitvoert gedurende langere tijd en verschillende situaties.
Security (Beveiligbaarheid)	De mate waarin een product of systeem informatie en gegevens beschermt.
Maintainability (Onderhoudbaarheid)	De mate waarin een systeem gewijzigd kan worden.

Tabel 2: De ISO 25010 standaard

Ieder attribuut bevat ook een aantal sub-attributen.

Zonder verder dieper in te gaan op deze aspecten, daar zijn genoeg andere bronnen voor, kan wel gesteld worden dat deze kwaliteitsattributen, zowel voor traditionele software systemen, als voor AI-systemen toegepast kunnen worden, om inzicht te geven in de kwaliteit van het onderhavige systeem. Voor een aantal sub-attributen is het voor ML-systemen niet eenvoudig om hier een invulling aan te geven. Denk maar eens na op welke wijze je de volgende zaken kan aantonen:

Functionality (Functionaliteit)		Opmerking
Completeness (Compleetheid)	De mate waarin de set van functionaliteiten alle gespecificeerde taken en doelen voor gebruikers ondersteunen.	Compleetheid zal nooit voor 100% zijn.
Correctness (Correctheid)	De mate waarin een softwareproduct of computersysteem de juiste resultaten met de benodigde nauwkeurigheid beschikbaar stelt.	Welke nauwkeurigheid is acceptabel?
Maintainability (Onderhoudbaarheid)		
Modularity (Modulariteit)	De mate waarin een systeem of computerprogramma opgebouwd is in losstaande componenten zodat wijzigingen van een component minimale impact heeft op andere componenten.	Bij regressie vaak 100% opnieuw trainen.
Reusability (Herbruikbaarheid)	De mate waarin een bestaand onderdeel gebruikt kan worden in meer dan één systeem of bij het bouwen van een nieuw onderdeel.	Dit is alleen mogelijk als exact dezelfde data structuur en verhouding wordt gebruikt.
Analysability (Analyseerbaarheid)	De mate waarin het mogelijk is om effectief en efficiënt de impact, van een geplande verandering van één of meer onderdelen, op een product of systeem te beoordelen, om afwijkingen en/of foutoorzaken van een product vast te stellen of om onderdelen te identificeren die gewijzigd moeten worden.	Dit is niet eenvoudig, zeker bij de ML variant Deep Learning.
Modifiability (Wijzigbaarheid)	De mate waarin een product of systeem effectief en efficiënt gewijzigd kan worden zonder fouten of kwaliteitsvermindering tot gevolg.	Zie wederom het punt van regressie.

Testability (Testbaarheid)	De mate waarin effectief en efficiënt testcriteria vastgesteld kunnen worden voor een systeem, product of component en waarin tests uitgevoerd kunnen worden om vast te stellen of aan die criteria is voldaan.	Zie de genoemde risico's in hoofdstuk 3 en 4 en de inleiding van hoofdstuk 7.

Tabel 3: Uitdagingen voor ISO 25010 attributen in relatie met AI

Voor AI-systemen en meer specifiek ML-systemen zijn een aantal van de huidige kwaliteitsattributen soms niet eenvoudig te definiëren of te verifiëren. Dit maakt het testen van een ML-systeem ook anders dan een traditioneel geprogrammeerd softwareproduct. Daarnaast blijkt dat een aantal van de attributen niet alle kenmerken van AI-systemen dekken.

6.2 Aanvullende kwaliteitsattributen

In de komende paragrafen geven we een overzicht van nieuwe, aanvullende kwaliteitsattributen. Deze kunnen specifiek gebruikt worden om de kwaliteit van een AI-systeem te beschrijven en te beoordelen. Dit overzicht is samengesteld uit meerdere bronnen, waarbij er ook een zekere overlap met elkaar is. Deze bronnen bevatten ook aan ethiek gerelateerde attributen.

6.2.1 Bron 1: Testing in the digital age

In het boek "Testing in the digital age - AI makes the difference" wordt een uitbreiding van de kwaliteitsattributen van de ISO 25010 standaard beschreven. Er worden 3 nieuwe hoofdgroepen toegevoegd.

Attribuut - sub-attribuut	Omschrijving
Intelligent behavior (Intelligent gedrag)	Intelligent gedrag is het vermogen om te begrijpen. Het is in feite een combinatie van redeneren, herinnering, verbeelding en oordeel. Elk van deze faculteiten is op elkaar aangewezen.
Ability to learn (Vermogen om te leren)	Het vermogen om te leren is het vermogen om te begrijpen en te profiteren van ervaring.
Improvisation (Improvisatie)	Improvisatie is de kracht van het intelligente systeem om in nieuwe situaties de juiste beslissingen te nemen.
Transparency of choices (Transparantie van keuzes)	Het begrijpen hoe een systeem tot een beslissing komt.
Collaboration (Samenwerking)	De mate waarin het systeem inspeelt op de verandering in het gedrag van mensen.
Natural interaction (Interactie)	Dit soort interactie is belangrijk in verbale en non-verbale communicatie. Vooral bij sociale

	robots is het belangrijk dat de manier waarop mensen met een robot omgaan natuurlijk is, zoals ze met mensen omgaan. Maar ook bij een zoekmachine, bijvoorbeeld Moet ik morgen een paraplu meenemen?
Morality (Moraliteit)	Zie het vorige hoofdstuk.
Ethics (Ethiek)	Zie het vorige hoofdstuk.
Privacy (Privacy)	Zie het vorige hoofdstuk.
Human friendliness (Mensvriendelijkheid)	Dit heeft te maken met veiligheid en beveiliging.
Personality (Persoonlijkheid)	Een persoonlijkheid is de combinatie van kenmerken of kwaliteiten die het onderscheidende karakter van een individu vormen.
Mood (Stemming)	Een tijdelijke gemoedstoestand of gevoel.
Empathy (Empathie)	Het vermogen om de gevoelens van een andere persoon te begrijpen en te delen.
Humor (Humor)	De kwaliteit van grappig of komisch zijn.
Charisma (Charisma)	De meeslepende aantrekkelijkheid of charme die anderen kan inspireren tot toewijding.

Tabel 4: Aanvullende kwaliteitsattributen beschreven in boek Testing in the Digital Age

6.2.2 Bron 2: ISO/CEN 5059 / ISO/IEC WO 5059

Deze nieuwe ISO standaard is nog in voorbereiding. In een aantal webinars zijn de volgende attributen genoemd:

Attribuut – subattribuut	Omschrijving
Ability to learn (Het vermogen om te leren)	Het vermogen van het systeem om te leren van het gebruik van het systeem zelf, of de gegevens en gebeurtenissen waaraan het systeem wordt blootgesteld.
Ability to generalize (Mogelijkheid om te generaliseren)	De mogelijkheid van het systeem om met succes toegepast te worden op verschillende en nooit eerder aan het systeem vertoonde scenario's.

Trustworthiness (Betrouwbaarheid)	De mate waarin het systeem kan worden vertrouwd door belanghebbenden.
Robustness (Robuustheid)	De mate waarin het systeem of een applicatie gevoelig is voor storingen van buitenaf.
Controllability (Beheersbaarheid)	De mate waarin een externe agent kan ingrijpen in het functioneren van het AI-systeem.
Explainability (Uitlegbaarheid)	De mate waarin belangrijke factoren, die van invloed zijn op het AI-system (bijv. algoritmen, model) uitgedrukt kunnen worden op een manier die mensen kunnen begrijpen.

Tabel 5: Concept aanvullende kwaliteitsattributen ISO/CEN 5059 / ISO/IEC WO 5059

Daarnaast zijn er nog diverse ethische attributen:

● Explicability and accountability ● Respect for democracy, justice and the rule of law ● Responsibility ● Privacy	● Fairness and non-disicrimintation ● Transparency ● Reinforcement of existing bias ● Consistency ● Free from bias

Tabel 6: Concept aanvullende ethische kwaliteitsattributen ISO/CEN 5059 / ISO/IEC WO 5059

Zoals eerder is aangegeven zijn veel van deze kenmerken in enige vorm ook genoemd in de vorige bron of bij het hoofdstuk ethiek. Om die reden werken we deze ethische attributen hier niet verder uit. Tezamen met de overige kenmerken is dit een bevestiging dat deze kenmerken van belang zijn.

6.2.3 Bron 3: DIN SPEC 92001-1 AI, Life Cycle Processes and Quality Requirements

Het Duitse norminstituut DIN heeft een specificatie ontwikkeld volgens de PAS-procedure (Publicly Available Specification) en is vrij te downloaden, zie Bijlage A voor de verwijzing. Een DIN SPEC kan gebruikt worden als basis voor een toekomstige standaard. In dit document wordt een aanpak voorgesteld om AI-gerelateerde softwarekwaliteitsaspecten te analyseren.

Het waarborgen van hoge kwaliteit van bepaalde AI-modules is een moeilijke taak. Vooral bij ML vanwege de onvoorspelbare reactie op onvoorziene input en bij DL vanwege het gebrek aan transparantie. Het op een gestructureerde manier aanpakken van deze uitdagingen is een basis voor de succesvolle ontwikkeling en integratie van robuuste, veilige en betrouwbare AI-modules.

Om dit mogelijk te maken beschrijft men in dit document een model met de volgende kwaliteitskenmerken:

Kwaliteitskenmerken	Omschrijving
Functionality & performance (Functionaliteit en prestaties)	Deze kenmerken geven de mate weer waarin een AI-module in staat is om onder gestelde voorwaarden zijn beoogde taak te vervullen.

Robustness, (Robuustheid)	Robuustheid geeft aan dat een AI-module in staat is om met foutieve, vervuilde, onbekende en vijandige invoergegevens om te gaan. Vanwege de complexiteit van de omgeving van de AI-module is robuustheid een belangrijk AI-kwaliteitsattribuut.
Comprehensibility (Begrijpelijkheid)	Machine Learning modellen kunnen ondoorzichtig zijn, waardoor het in kaart brengen van input naar output grotendeels onbegrijpelijk is voor belanghebbenden. Dit betekent dat de AI-component transparant en interpreteerbaar moet zijn. Wetgeving is een mogelijke externe beperking.

Tabel 7: Kwaliteitsattributen DIN SPEC 92001-1

Merk op dat deze kenmerken niet geheel los van elkaar staan, maar deze indeling vergemakkelijkt een gestructureerde opsomming van specifieke kwaliteitseisen.

6.3 Tot besluit

Een aanvulling op de huidige kwaliteitsattributen is wenselijk om goed aan te kunnen sluiten bij de komst van AI. Er zijn meerdere initiatieven gestart om die aansluiting te vinden, vanuit verschillende partijen en met verschillende invalshoeken. Als de initiatieven naast elkaar gezet worden, zijn er verschillen, maar wat meer opvalt zijn de overeenkomsten.

Er is in het algemeen aandacht voor autonomie en bijkomende verwachtingen rond consistentie, robuustheid en zelfredzaamheid. Er is ook aandacht voor de intelligentie en menselijkheid van niet-menselijke systemen; denk aan lerend en generaliserend vermogen, aan transparantie en vertrouwen, en aan het toepassen van sociale eigenschappen in interactie.

Het lijkt voor de hand te liggen dat er consensus zal komen over de vorming van nieuwe of aangepaste kwaliteitsattributen. Hierdoor ontstaat er een goed en breed gedragen set aan kwaliteitsattributen om de kwaliteit van AI-oplossingen te specificeren en te testen.

7 Testen van AI

De voorgaande hoofdstukken hebben laten zien vanuit welke invalshoeken je de toepassing van AI en de bijbehorende risico's kunt bekijken. Na het lezen van deze hoofdstukken zal duidelijk zijn, dat de risico's en aandachtspunten voor een groot deel anders zijn dan bij traditioneel geprogrammeerde oplossingen. Dit hoofdstuk werkt alle mogelijke risico's en aandachtspunten uit naar testactiviteiten.

De testactiviteiten volgen uit de risico's zoals in hoofdstuk 2 staan beschreven. De expliciete link tussen risico en testactiviteit leggen we in bijlage D.

7.1 Statische testen

Als ML-projecten niet succesvol zijn, blijkt achteraf vaak dat een aantal kwaliteitsproblemen al voor de start van het ontwikkelen van het model bekend hadden kunnen zijn. Dit kan worden voorkomen door statische testen. Denk hierbij aan checklists, reviews en assessments. Dit soort testen kunnen niet alleen bij aanvang van een project worden uitgevoerd, maar ook gedurende het project als er een bepaalde mijlpaal is bereikt.

7.1.1 Checklists

Er kunnen checklists gebruikt worden die gangbaar zijn in de markt, of voor het project gemaakt is in samenspraak met de stakeholders. Punten die in zo'n checklist worden opgenomen zijn bijvoorbeeld:

- Is het doel duidelijk?
- Is de tijdlijn duidelijk?
- Is de verwachting van de stakeholders haalbaar?
- Is er data beschikbaar?
- Is er voldoende data en data-verhouding in balans?
- Is de oplossing toekomst vast?
- Is er een vergelijkbaar project geweest of academische studie naar de oplossing?
- Zijn er ethisch issues te verwachten?

In ieder geval moet het succes worden gedefinieerd en welke minimumstandaard moet worden bereikt om de AI-oplossing in gebruik te nemen. Dit is zeker van belang doordat er niet met een expliciet behaald eindresultaat kan worden gewerkt. Bij gebrek aan gedetailleerde testgevallen kunnen deze checklist-test richtlijnen houvast en een zekere mate van consistentie bieden.

7.1.2 Reviews

Er zijn verschillende vormen van reviews mogelijk, van formele inspecties tot collegiale toetsing. Naast de toetsing van een bepaald aspect is ook de verkregen feedback een waardevol onderdeel van deze wijze van testen. Reviews kunnen plaatsvinden aan het begin, tijdens en na afloop van het ontwikkeltraject. Een aantal punten die zich lenen voor een review aan het begin van het project zijn het onderzoek naar de beschikbare data, het voorgestelde ML-algoritme, het doel en het middel om dit doel te bereiken, inclusief ethische afwegingen en de wijze waarop het succes van het model gemeten wordt. Tijdens of na afloop van de bouw van het model kan bijvoorbeeld ook nog gedacht worden aan een technisch review op de code of parameterinstelling en een proces-review met eventueel geleerde lessen en aanbevelingen voor verbeteringen.

De informatie verkregen uit deze statische testen is zeer geschikt voor het maken van testcases en testscenario's.

7.2 Testen van data

Bij de voorafgaande hoofdstukken is het belang van goede data voor de werking van het ML-model al benadrukt. Hieruit volgt ook de conclusie dat het testen van deze data een belangrijke stap in het ontwikkelproces is. ML leert op basis van de beschikbare data. Is deze data niet volledig of niet correct, dan zal het opgeleverde model ook niet naar behoren presteren. De afhankelijkheid van data is om deze reden één van de belangrijkste risico's die vermeld staan in hoofdstuk 2. Bij de toelichting van dit risico zijn een aantal punten genoemd die van invloed zijn op de kwaliteit van een model en kunnen als basis voor een test dienen. Bij data kan onderscheid gemaakt worden tussen de data die gebruikt wordt als invoerwaarde, bijvoorbeeld de oppervlakte van een huis, of het aantal kamers, maar ook wat deze verzameling van data uiteindelijk voorstelt, bijvoorbeeld de verkoopprijs van een huis. Zoals aangegeven bij het voorbeeld over het herkennen van een kat bij (afhankelijkheid van data), is dit vaak subjectief.

Het verkrijgen van de data en het geschikt maken voor verdere verwerking ten behoeve van het leren van het model, het zogenaamde preprocessing van de data, is een specialisme dat Data Engineering wordt genoemd. Hierbij moet gedacht worden aan het aanvullen van ontbrekende gegevens of het transformeren van de data naar een ander formaat. Deze activiteit kan zowel handmatig als geautomatiseerd worden uitgevoerd. Om de kwaliteit van dit proces te bewaken is het van belang om de stappen van verkrijgen van de data en de preprocessing stappen te beschrijven waardoor dit proces te testen is.

Bij het verkrijgen van de data moet ook getoetst worden in welke mate deze data overeenkomt met de werkelijkheid. Zo kan getoetst worden of de man/vrouw verhouding juist is en of de leeftijdsverdeling overeenkomt met de doelgroep.

Als de data uit meerdere bronnen afkomstig is, kan dit ook effect hebben op de werking van het model. Vaak worden er aan de telefoon andere vragen gesteld dan de vragen die op een website van een bedrijf staan. Daarnaast is het aannemelijk dat er een ander type persoon contact zoekt via social media dan de personen die contact zoeken via de telefoon. Om deze reden kunnen deze gegevens niet zonder controle bij elkaar gevoegd worden.

Ook bij het gebruik van het model in productie moet getest worden op data. Deels is dit een invoercontrole die getest moet worden, zoals:

- Is het formaat juist,
- Valt de data binnen de ranges die gebruikt zijn bij het leren.

Verder is het van belang het patroon van de invoerdata te vergelijken met het patroon van de data waarmee het model is getraind. Als er een verschuiving in dit patroon heeft plaatsgevonden, dan kan dit een signaal zijn dat het model niet meer optimaal werkt. Stel dat bij het maken van een model voor huizenprijzen 80% van de huizen 3 of 4 kamers heeft en na verloop van tijd blijkt dit aandeel gedaald is tot 50%. Dit kan invloed hebben op het model. Er moet om deze reden onderzocht worden in hoeverre het model nog overeenkomt met de werkelijkheid. Zie hier het belang om de verdeling van de data goed vast te leggen en te monitoren.

7.3 Testen van het model

Voor het maken van een model is een beperkte hoeveelheid code nodig, of kan soms met alleen een parameter configuratie (het zogenaamde Auto-ML) plaatsvinden. Ondanks deze beperkte inrichting, kan een relatief beperkte aanpassing van een parameter een grote impact hebben op het resultaat. Het testen van deze configuratie van het model is project specifiek, maar onderwerpen die aandacht verdienen zijn:

- Is het juiste algoritme gebruikt,
- Zijn de parameters van dit algoritme goed ingevuld
- Klopt het invoer en uitvoer formaat
- Hoelang duurt het trainen van het model

Als men in de praktijk spreekt over het testen van een model, bedoelt men niet de bovenstaande configuratietest, men heeft het over het evalueren van het model. Met andere woorden, hoe goed zijn de voorspellingen van het model?

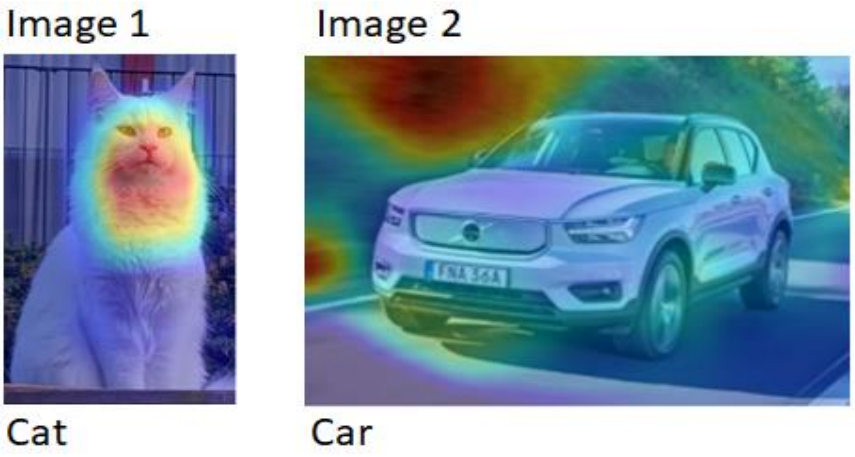
Dit is een standaard stap in het ontwikkelproces en ook een iteratief proces, de uitkomst van deze evaluatie wordt gebruikt om de parameters te tunen om een beter resultaat te krijgen. Dit gaat door totdat er geen verbetering meer is. De prestatie van een model wordt bekeken aan de hand van verschillende indicatoren, zoals nauwkeurigheid. Testen komt neer op het vergelijken van de voorspelde waarden met de werkelijke waarden. Hiervoor wordt in principe andere data gebruikt dan de data die gebruikt is voor het trainen van het model.

Het is gebruikelijk om de beschikbare data te verdelen in een gedeelte om het model mee te trainen en een gedeelte om het model mee te testen. Omdat deze testdata niet gebruikt is om het model te trainen en nog niet eerder door het model gezien is, is deze data geschikt om het model te beoordelen. Deze laatste situatie komt namelijk het meest overeen met de toekomstige data in de productieomgeving. Dit geeft ook aan dat de verdeling zorgvuldig moet gebeuren want beide datasets moeten dezelfde kenmerken hebben, bijvoorbeeld hetzelfde gemiddelde, spreiding en aantal groepen. Dit moet voordat men gaat starten met het testen van een model geanalyseerd worden omdat er anders geen oordeel over de werking van het model gegeven kan worden.

Naast het kunnen testen van een model op het verkregen resultaat, d.w.z. hoe goed is de verkregen voorspelling, kun je een model ook testen door te onderzoeken hoe het model tot deze voorspelling is gekomen. Het uitleggen van een model staat bekend als Explainable AI. Hierbij moet direct een opmerking geplaatst worden, ML-modellen die gebaseerd zijn op DL technieken zijn in de praktijk moeilijk uitlegbaar. Het interpreteren van het gedrag is een betere omschrijving. Interpretatie is de mate waarin een mens de reden van een beslissing kan begrijpen. Er kan bijvoorbeeld onderzocht worden welke invoervariabelen het meeste bijdragen aan het uiteindelijk resultaat, of nagaan hoe een individueel resultaat tot stand is gekomen.

Bij Beeldherkenning is het mogelijk om na te gaan welke pixels het meeste bijdragen aan het resultaat van de voorspelling. Dit kan zichtbaar gemaakt worden met een zogenaamde heatmap, hoe meer een pixel bijdraagt aan het resultaat hoe warmer, dus roder deze pixel op de oorspronkelijke foto wordt geprojecteerd.

Zie als voorbeeld onderstaande foto’s:



Afbeelding 10 Heatmap van Kat en Auto

Bij foto 1, de kat, zijn de meeste rode pixels van de heatmap in het gezicht van de kat. Dit is inderdaad een typisch kenmerk van een kat en een goede indicatie dat de beslissing om deze foto als kat te classificeren op basis van de juiste grond is genomen.

Bij foto 2, de auto, bevindt het meeste rode gebied van de heatmap zich buiten de auto. Dit houdt in dat het meest belangrijk gedeelte van deze foto om de voorstelling als auto te classificeren, buiten het gebied van de auto ligt. Dit geeft direct aan dat de voorspelling dat deze foto een auto voorstelt niet veel waard kan zijn.

7.4 Testen van de functionaliteit van het model

Bij het testen van een ML model wordt niet verwacht dat men voor elk afzonderlijk data-record uit de testdata een perfecte voorspelling doet. Het gaat om de algehele prestatie van het model. Het is goed mogelijk dat een test geslaagd is met een nauwkeurigheid van 80%, een enkel verkeerd voorspeld resultaat is geen bug. De vraag is ook niet of een model correct is, maar hoe goed het model een probleem oplost.

Voor het bepalen van de kwaliteit van een model is meer nodig dan de algemene nauwkeurigheid. Om een oordeel te geven over de verkregen functionaliteit van het model kan het goed zijn om ook de verschillende deelgebieden zoals grensgevallen en doelgroepen onderzocht worden.

Hieronder worden een aantal testtechnieken beschreven:

7.4.1 A/B testen

Bij een A/B test worden verschillende modelversies met elkaar vergeleken. Een beperkte groep (de B-groep) krijgt de gewijzigde versie en de controlegroep (de A-groep) de ongewijzigde versie. De verschillen in resultaat of gebruik worden geanalyseerd. Met deze test wordt het test oracle-probleem gepasseerd. Het gaat er bij deze test niet om hoe goed een model is, maar of een model ‘statistisch significant’ beter is dan een andere versie van het model. Daarom wordt deze test veelvuldig gebruikt bij het testen van ML-systemen.

De A/B test wordt niet alleen in de ontwikkelfase toegepast, maar ook in productie. Soms is het moeilijk te bepalen in welke mate een versie beter is. Denk hierbij aan het toevoegen van nieuwe functionaliteit terwijl er maar een beperkte hoeveelheid data is om deze nieuwe functionaliteit te testen, of de verwachting dat het gedrag van de gebruiker zal veranderen.

Deze testtechniek heeft meerdere varianten. Door bijvoorbeeld de A- en B-groep dezelfde data te laten gebruiken kunnen de verschillen tussen de versies geanalyseerd worden. Een andere variant is de testdata op een gelijke wijze te verdelen en hiermee het model te testen, een zogenaamde A/A test. Aangezien het dezelfde modelversie is, verwacht je geen verschillen.

Dit soort testen geven wederom een oordeel over de prestatie van het model als geheel. De prestaties tussen meerdere modellen vergelijken geeft vaak geen direct inzicht in het modelgedrag zelf. Als voorbeeld kan dienen dat een model gemiddeld beter kan zijn, maar het kan dan voorkomen dat een bepaalde groep uit de data- of gebruikers- populatie in deze vernieuwde versie bevoordeeld of benadeeld is. Om dit te testen zijn andere testmethoden nodig, methodes die het gedrag van het model testen en een begrijpelijk en voorspelbaar resultaat opleveren.

7.4.2 Equivalence Partitioning (Equivalentieklassen)

De hierboven beschreven testen richten zich voornamelijk op het optimaliseren van de algehele kwaliteit. Dit kan te algemeen zijn. Modellen die hoge algehele prestaties behalen, kunnen onaanvaardbare uitvalpercentages produceren op kritieke delen van de gegevens, denk hierbij aan de detectie van kwetsbare fietsers tijdens een autonome autorit.

Bij Equivalence Partitioning wordt de data in partities (klassen) verdeeld op een zodanige wijze dat alle leden van een bepaalde partitie naar verwachting op dezelfde manier worden verwerkt en een vergelijkbaar resultaat opleveren. Het gaat dus om consistent gedrag; in soortgelijke gevallen wordt een soortgelijk resultaat verwacht. Aangezien de data bij ML-systemen omvangrijk is, zijn er tools om dit te ondersteunen. Eventueel zou hierbij pairwise testing kunnen worden toegepast. Hierbij wordt een combinatie van variabelen (of partities) gelijk getest waardoor er met een beperkte subset van combinaties toch een redelijke dekking mogelijk is. Een andere variant is het navragen van de meest relevante partities bij inhoudelijk experts. Met een techniek zoals de Datacombinatietest kan selectief aan aantal invoervelden worden gecombineerd om daar vervolgens de juiste dekking van vast te stellen.

Met deze testen is het mogelijk om een meer inhoudelijke vergelijking tussen modelversies te maken. Bij een nieuwe versie van een model wordt niet verwacht dat het resultaat van een partitie is gewijzigd, tenzij dit een bewuste keuze is geweest voor het maken van deze nieuwe versie. Het is dus handig om een goed versiebeheer van zowel de testdata en de resultaten per test in te richten.

7.4.3 Boundary Value Analysis (Grenswaardenanalyse)

Dit is een test die op het vervolg van de Equivalentie Partitionering test kan worden uitgevoerd. De minimum- en maximumwaarden (of eerste en laatste waarden van een dataset) van een partitie zijn de grenswaarden. Het gedrag rondom de grenzen van de partities zal eerder anders zijn dan het gedrag binnen de partities. Bij Machine Learning moet dit nog wat breder worden opgevat. Denk bijvoorbeeld aan een applicatie voor beeldherkenning. Gedurende de uren met daglicht en gedurende de nacht zullen er naar verwachting weinig verschillen in het herkennen van een object zijn, maar in de grens tussen deze twee partities, de ochtend- en avondschemering met wisselende lichtsterkte en veranderende kleuren zijn er veel veranderingen. Het gedrag van het model moet hierop getest worden.

Als variant op deze test kan de Corner Test ingezet worden. Bij Corner Testing worden van alle input variabelen (features) de uiterste waarde genomen. Hierbij kan nog een onderscheid gemaakt worden tussen het minimum en maximum van deze variabele in de (test) dataset, of het mogelijke minimum of maximum dat als input van het model kan dienen.

7.4.4 Metamorphic testing

Dit is een testmethode om het test-oracle probleem te verminderen. Het idee is simpel: zelfs als we niet weten wat de juiste output (het resultaat) moet zijn van een enkele input, kunnen we nog steeds de relaties tussen de outputs van meerdere inputs kennen, vooral als de inputs zelf gerelateerd zijn. We kunnen het model controleren op deze relaties, de zogenaamde metamorphische relatie. Als deze niet in stand blijft, dan is dit zeker een signaal om nader te onderzoeken.

Ter verduidelijking kan het voorbeeld uit hoofdstuk 1 dienen, het voorspellen van de waarde van een huis. Wat de 'juiste' waarde van een huis is, is niet vast te stellen. Wel kan er aangenomen worden dat de huizenprijs in waarde stijgt zodra een huis meer oppervlakte heeft. In dit voorbeeld is er dus een metamorphische relatie tussen de inputwaarden en outputwaarden te maken, namelijk als de oppervlakte groter wordt, dan zal de huizenprijs stijgen.

Deze testsoort kan gebruikt worden in combinatie met de partities zoals beschreven bij de Equivalence Partitioning testmethode. Van de input waarden binnen een partitie verwachten we geen significante wijziging in de output. Dit wordt ook wel een invariance test genoemd. Deze testen stellen ons in staat een reeks verstoringen te beschrijven die we in de invoer zouden moeten kunnen aanbrengen zonder de uitvoer van het model te beïnvloeden. We kunnen deze verstoringen gebruiken om paren invoer-voorbeelden te produceren (origineel en verstoord) en om te controleren op consistentie in de model-voorspellingen. Het gewenst resultaat moet hetzelfde blijven binnen de bandbreedte.

Deze techniek kan ook tegenovergesteld gebruikt worden, men spreekt dan over een Directional Expectation Test. Door middel van verstoring op de input verwachten we een voorspelbaar effect op de uitvoer van het model, een bewust verschil buiten een bandbreedte.

7.4.5 User Story testen – Use Case Testen

Het beschrijven van de gewenste functionaliteit, gewenste gedrag en gestelde eisen kan opgesteld worden in de vorm van User Stories en/of Use Cases. Bij het gebruik van deze techniek worden de randvoorwaarden, condities en de acceptatiecriteria beschreven. Het samenstellen van deze testen gaat over het algemeen in samenspraak met de stakeholders en/of domeindeskundige. Deze testsoort wordt vaak gebruikt bij traditionele softwareontwikkeling, maar kan ook goed bij het testen van ML toepassingen gebruikt worden.

7.4.6 Expertpanel testen

Door de kennis en kunde van domeindeskundigen en experts in te zetten kan het resultaat van een ML applicatie vergeleken worden met het verwachte resultaat van deze deskundigen. Op basis van analyse van het domein worden input-testdatasets samengesteld. Zowel de ML applicatie als de deskundigen bepalen aan de hand van deze testdatasets de waarschijnlijke resultaten. Hiermee kan de kwaliteit van de ML applicatie beoordeeld worden.

Bij de inzet van deskundigen is het van belang om rekening te houden met de volgende punten:

- Menselijke experts variëren in competentie, dus de betrokken experts moeten representatief zijn zowel in aantal als kennisgebied en kennisniveau.
- Deskundigen zijn het misschien niet eens met elkaar, zelfs niet als ze dezelfde informatie krijgen. In veel gevallen wordt de juistheid van de test verschillend waargenomen door verschillende individuele gebruikers. Wat voor sommigen bijvoorbeeld als een klacht lijkt, kan voor anderen als een neutrale verklaring overkomen.

- Menselijke experts kunnen bevooroordeeld zijn voor of tegen automatisering, zie het risico angst voor AI.

Ondanks deze punten kunnen dit soort testen snel inzicht geven in de kwaliteit van een model en bijdrage aan de acceptatie in het gebruik.

7.4.7 Experience-based testen

Bij deze op ervaring gebaseerde testtechnieken zijn de testgevallen afgeleid van de vaardigheden en intuïtie van testers en de ervaring met vergelijkbare toepassingen. Deze technieken kunnen nuttig zijn bij het identificeren van testen die niet gemakkelijk te beschrijven zijn. Voor een aantal ML applicaties zal deze situatie zeker van toepassing zijn.

Experience-based testen zijn:

Exploratory Testing

Bij Exploratory Testen worden de testen niet vooraf beschreven, maar vindt de evaluatie plaats tijdens het uitvoeren van de test. Een exploratory test kijkt naar een bepaald thema, welke is vastgelegd in een test charter. Dit zorgt ook dat je achteraf kunt zien wat je allemaal getest hebt, wat vervolgens helpt bij vertrouwen en acceptatie. Test charters kunnen zich richten op bijvoorbeeld het afvangen van misbruik van het model, het vinden van ondervertegenwoordigde voorbeelden en het herkennen van bias. Dit kan dus zowel voor de analyse van de data als voor het komen tot een model.

Bij het ontwikkelen van ML modellen is het van tevoren moeilijk in te schatten wat de haalbare specificaties zijn en zijn de requirements vaak beknopt beschreven. Daarom is deze testaanpak nuttig en goed bruikbaar. Mede, omdat vooraf niet precies bedacht kan worden wat en hoe er getest moet worden. Deze testen worden vaak gebruikt als aanvulling op andere, meer formele testtechnieken, of wanneer er een aanzienlijke tijdsdruk op het testproces bestaat.

Error Guessing

Op basis van kennis van de testdataset wordt deze aangevuld met data waarbij problemen te verwachten zijn. Dit kunnen weinig voorkomende combinaties zijn over uitersten. Maar ook kan gedacht worden aan:

- Hoe het ML model in het verleden heeft gewerkt
- De fouten, zwakheden, die inherent aan het gekozen model zijn
- De fouten die data engineer kan maken
- Fouten die zich hebben voorgedaan in andere toepassingen

7.4.8 Testen vanuit Persona's

Persona's zijn fictieve personages die stakeholders- of gebruikersgroepen vertegenwoordigen. Deze methode start met het identificeren en opstellen van de gebruikersprofielen. In de profielen worden zaken die van belang kunnen zijn voor de te testen (AI) applicatie opgesteld, zoals opleidingsniveau, reden om de applicatie te gebruiken en dergelijke. De winst van het gebruik van persona's is dat er geprobeerd wordt in de huid van een ander te kruipen en de (AI) applicatie vanuit deze ogen te bekijken.

Het is een testmethode die geschikt is om specifieke vooroordelen op te kunnen sporen. Zoals aangegeven bij het hoofdstuk over ethiek mag een model niet bevooroordeeld zijn. Het is helaas zeer waarschijnlijk dat dit verschijnsel optreedt. In veel systemen wordt bijvoorbeeld een dubbele

nationaliteit bijgehouden en het kan zijn dat daar ook een andere verwerking heeft plaatsgevonden. Aangezien een model getraind wordt op data uit het verleden, kan het gedrag van het model hier ook door beïnvloed zijn geraakt.

7.5 Testen op drift

Vaak is het maken van het eerste ML-model niet eens zo ingewikkeld in vergelijking met de eerste versie van geprogrammeerde software. Het onderhouden van een model vereist echter meer aandacht, door het inspelen op verandering in de markt en beschikbaarheid van data vereist. De resultaten van een model zullen in de loop van de tijd gaan veranderen. Het kan zijn dat het patroon van invoerdata is gewijzigd, maar het kan ook zijn dat het resultaat van het model op een andere wijze wordt gewaardeerd. Het resultaat is dat het ML model minder effectief wordt. Dit verschijnsel wordt drift genoemd en kan in sommige gevallen al binnen enkele dagen plaatsvinden. Oorzaken zijn bijvoorbeeld verandering in de voorkeuren van gebruikers, marketingcampagnes, seizoensinvloed of aanpassingen bij een concurrent. In praktijk houdt dit in dat er vrijwel nooit een definitieve versie van een ML-model is.

Om deze redenen moet een model in productie altijd worden gemonitord om dit soort situaties te detecteren. Het model zal moeten worden bijgeschoold, opnieuw getraind, om weer te kunnen functioneren in deze gewijzigde wereld. Aangezien het proces van detecteren en de actie tot aanpassing van het model cruciaal is voor de effectiviteit, moet dit proces ook getest worden. Dit is namelijk een waarborg voor de kwaliteit van het model. Drift heeft dus gevolgen voor de beheerorganisatie en ook voor de regressie testen.

7.6 Regressietesten

Net als bij software zal er regelmatig een update van een model plaatsvinden. Dit kan meerdere redenen hebben:

1. Het model moet opnieuw worden getraind omdat de oorspronkelijke doelen niet meer gehaald worden, doordat de werkelijkheid is veranderd en het model daar onvoldoende op aansluit. Dit concept hebben we in de vorige paragraaf toegelicht als concept drift.
2. Er worden nieuwe of aangepaste doelen nagestreefd met het model, terwijl sommige oorspronkelijke doelen nog steeds van toepassing zijn. Denk aan het herkennen van nieuwe vragen door een chatbot of een nieuwe productcategorie bij productvoorstellen.

Het is dus van belang dat de requirements van het model duidelijk zijn en dat er gemonitord wordt dat het model nog aan deze eisen voldoet.

Een update van het model vereist enige voorzichtigheid. Bij traditionele softwareontwikkeling is het risico op regressiefouten meestal te overzien omdat er doorgaans maar een beperkte hoeveelheid code wordt aangepast bij een nieuwe release. Bij Machine Learning is dat anders, aangezien het model in zijn geheel opnieuw getraind wordt. Dus ook bij een beperkte wijziging, is het mogelijk dat het model op cruciale punten verandert. Het kan ook zo zijn dat de algemene prestatie van het model wordt verbeterd, maar toch een ongewenst negatief effect in een datagroep (zoals grenswaarden, datapartities en persona's profielen) optreedt. Dit gegeven rechtvaardigt het belang voor het uitvoeren van goede regressietesten bij Machine Learning.

Bij iedere update zal getest moeten worden in welke mate er een verschuiving heeft plaatsgevonden van het resultaat. Dit kan door middel van de besproken A/B testen. Het is van belang de originele

test dataset inclusief de bijbehorende resultaten te bewaren om de groei en de verandering van het model te onderkennen. Daarnaast is een test dataset met actuele data van belang om de huidige prestatie te testen. Het onderhouden van oude en nieuwe testcases en testmodel- versies geeft een goed beeld van de kwaliteit van de huidige en eerdere modellen.

7.7 Tot besluit

Wij vinden dat met de beschreven testmethodes en kwaliteitskenmerken inzicht kan worden gekregen in de kwaliteit van een model. Uiteraard is zo'n overzicht nooit volledig en zal per situatie en per type model de ene test meer geschikt zijn dan de andere test. Daarnaast is het goed mogelijk om een aantal testmethoden in een test te combineren, zoals testers dit al vele jaren doen. Ook zal per test bekeken moeten worden welke risico's, kwaliteitskenmerken en ethische aspecten men wil afdekken. Veel van deze dingen zijn anders bij ML-systemen, maar sommige aspecten blijven ongeveer hetzelfde. Zo zijn integratie-, security- en gebruikersacceptatietesten niet beschreven in dit document omdat deze testen voor een ML model niet fundamenteel afwijken.

Het is veel gemakkelijker voor mensen om technologieën te vertrouwen waarover ze volledige, constante controle kunnen uitoefenen. Dit is één van de grootste uitdagingen binnen AI en ML. Door het testen op consistent en voorspelbaar gedrag en het documenteren van de testresultaten is het mogelijk om vertrouwen te krijgen in de werking van ML-systemen.

8 Testen van AI in de praktijk

De rol van testers of Quality Assurance-specialisten in AI-trajecten is relatief nieuw. Dit biedt veel kansen om zelf vorm aan te geven aan deze rol en bij te dragen aan een succesvolle implementatie van een AI-toepassing. In dit hoofdstuk gaan we dieper in op projecten, rollen en vaardigheden.

8.1 Hoe verloopt een AI project?

Ieder AI-initiatief (of dus specifiek: ML-initiatief) zal data willen inzetten om een bepaald doel te bereiken. Soms zal dit ontstaan vanuit de nieuwsgierigheid of behoefte bij één van de medewerkers. Als de data beschikbaar is, kan gestart worden met een experiment. Een eerste conceptmodel is snel opgezet en getraind.

Een ander startpunt is een business case, vaak om een proces binnen de organisatie te versnellen of de klant beter te bedienen. Dit soort initiatieven zullen eerder als volwaardig project worden opgezet, maar ook hier begint het meestal met een idee vanuit een medewerker. Het startpunt waarbij directie of management aangeeft dat ‘we toch eens iets met AI moeten’ verlaagt de kans op een succesvolle implementatie aanzienlijk.

Een AI-project is een constante verkenningstocht. Men leert welke data beschikbaar is, welke kwaliteit het heeft en in hoeverre het helpt om een doel te behalen. In de tussentijd zullen er diverse modellen worden getraind. Iedere iteratie geeft nieuwe inzichten en nieuwe uitdagingen om op te lossen.

Bij een iteratie zullen de volgende activiteiten een rol spelen, al zal de focus bij de eerste iteraties vooral op de eerste activiteiten liggen en bij de laatste iteraties op de laatste activiteiten.

- Ideevorming. Dit is een fase waarin de balans tussen ambitie en haalbaarheid wordt onderzocht. Het is ook een goed moment om kritische vragen te stellen, bijvoorbeeld of ML inderdaad de beste keuze voor het probleem is, of de juiste kennis en middelen beschikbaar zijn, of de data beschikbaar en bruikbaar is, welke algemene eisen er zijn rond onderwerpen zoals ethiek, uitlegbaarheid en privacy, enzovoort.
- Dataverzameling en -bewerking. Dit is het ophalen of opvragen van interne of externe databronnen, het beoordelen op consistentie, compleetheid en bruikbaarheid. Ook het selecteren, bewerken en aanvullen gebeurt hier. Uiteraard volgens duidelijke, logische en vastgelegde keuzes.
- Modelleren, trainen en valideren. Hier wordt de modelvariant gekozen, mogelijk op basis van proeven met meerdere varianten. Daarbij worden de modellen geconfigureerd, wordt data aangeboden en komen er modelscores uit. Deze scores worden beoordeeld op basis van een separate dataset.
- Testen. Dit is de inhoudelijke, handmatige beoordeling van het model. Hierbij wordt gekeken naar individuele voorbeelden, zoals grenswaarden en outliers. Het doel is om vast te stellen dat de scores er logisch uitzien door er kritisch naar te kijken, liefst met hulp van inhoudelijke experts. Daarnaast zullen hier ook de technische testen worden gedaan om het model in een eventuele keten aan het werk te zien.

- Accepteren. Dit is de acceptatietest, waarbij alle belanghebbenden overtuigd moeten raken dat het AI-model bruikbaar en prettig werkbaar is. Belanghebbenden zijn vertegenwoordigers van klanten of andere eindgebruikers, van eigen medewerkers die ermee te maken krijgen, maar ook een vertegenwoordiging van de opdrachtgever die de business case zal willen valideren. Daarnaast zullen de juristen en de security afdeling van de organisatie hun mening willen geven.
- Implementeren en gebruiken. Hier wordt de organisatie klaargemaakt om het model te gebruiken in de praktijk. Zo kunnen er bij de klantenservice vragen komen die moeten worden beantwoord. Ook moeten de scores van het model in de praktijk worden gemonitord en moet er een nieuwe versie van het model - of een extra controleregel als vangnet - kunnen worden geïmplementeerd als dat nodig blijkt te zijn.

8.2 Welke rollen zijn er in een AI project

Een veelgebruikte rolverdeling in de huidige softwareontwikkeling is de volgende.

- Een ontwerper of architect die bedenkt wat er moet worden ontwikkeld.
- Een programmeur die het plan in detail uitwerkt en vervolgens uitschrijft in code.
- Een tester die, op basis van risico's, controleert of er fouten of onduidelijkheden zijn.
- Een acceptant, die controleert of het geheel werkbaar is en waarde toevoegt.
- Een projectleider of scrum master, die voor voortgang en samenwerking zorgt.

Het werken volgens de iteratieve agile of devops methode past erg goed bij de manier waarop AI-modellen ontwikkeld en beheerd worden. AI-ontwikkeling is namelijk bij uitstek een gezamenlijk verkenningstraject om uit data voorspellende waarde te halen, waarbij iedereen vanuit zijn eigen specialisme ideeën inbrengt en vragen stelt. De ontwikkeling van een AI-applicatie is ook een iteratief proces. Bovendien zal een AI-model, nadat het in gebruik is genomen, altijd nauwlettend gemonitord moeten worden en zal er regelmatig doelgericht hertraind moeten worden.

Ten opzichte van agile en devops zien we een aantal nieuwe rollen ontstaan:

- Een ML datascientist, die zorgt dat data verzameld en bruikbaar gemaakt wordt.
- Een ML-engineer, die een model bouwt en test op basis van bruikbare data.
- Een AI Ethiek-functionaris, vergelijkbaar met de Data Privacy Officer die na de invoering van de AVG verantwoordelijk is voor het bereiken en het bewaken van de naleving.

Een ander verschil met agile is dat de acceptant veel eerder in het project betrokken wordt. Hij of zij is actief bezig met het bruikbaar maken van de data, op basis van kennis van de praktijk. Er is dus al meteen een actieve samenwerking tussen de dataspecialist en de acceptant. Daarna, zodra er een eerste versie van het model getraind is, zal de acceptant bezig gaan met het interpreteren van de gevonden trends en met de meest opvallende uitkomsten. Er zal samen worden nagedacht over het maken van de volgende iteratie van het model.

Het gestructureerd samenwerken in een AI-project is van belang. De ML-engineer levert harde cijfers op en de acceptant interpreteert de resultaten. Er zijn echter veel meer activiteiten die bijdragen aan het verbeteren van de kwaliteit. Dit is waar testers of QA-specialisten een rol kunnen vervullen, zodat uiteindelijk voldoende zekerheid over dekking en openstaande risico's wordt verkregen. Het

deelnemen aan een AI-traject biedt kansen voor testers om hun toegevoegde waarde te laten zien en om zich verder te ontwikkelen in een nieuw vakgebied.

8.3 Kennis en vaardigheden

Bij vaardigheden die nodig zijn in een AI-traject kan gedacht worden aan traditionele vaardigheden van testen die samenhangen met kwaliteitsborging, zoals het vinden van fouten en uitzonderingen, het leggen van verbanden en het uitproberen van onverwachte combinaties, en het inschatten van risico's. Het best passende profiel van de tester of QA-specialist zal per project anders zijn. Soms zullen er meer coördinerende vaardigheden gevraagd worden, soms ligt het accent meer op de techniek.

Bij het testen in AI-projecten kan er behoefte zijn aan nieuwe of meer diepgaande kennis en vaardigheden zoals:

- Uitgebreide vaardigheid met het gebruik van black box testtechnieken, zeker bij AI-ontwikkeling waar deep learning wordt gebruikt en het onmogelijk is om de inwendige werking van het model te beoordelen.
- Kennis van ethische overwegingen die gemaakt moeten worden, op basis van richtlijnen vanuit de overheid of vanuit de eigen organisatie. Waarbij het ook van belang is om dit op de juiste momenten en op de juiste manier onder de aandacht te kunnen brengen.
- Kennis van data en vaardigheid in het beoordelen ervan. Hoe is het opgezet, hoe moet het worden geïnterpreteerd, welke problemen zijn er doorgaans met data, welke manieren zijn er om dit op te lossen.
- Basiskennis van AI / ML / DL, kennis van verschillende manieren om te modelleren. Welke modellen van belang zijn, hangt uiteraard afhankelijk van het traject. Beeldherkenning is anders dan patroonherkenning in datasets, modellen die iets genereren zijn ook weer anders.
- Kennis van programmeertalen en frameworks die bij AI veel gebruikt worden. Dit kunnen tools zijn om met data om te gaan, maar ook tools om te modelleren. Soms gaat dit op basis van programmeertalen Python of R, soms met behulp van libraries zoals TensorFlow, Pytorch, Keras, Pandas en Scikit-learn. Eventueel aangevuld met kennis van platformen die modellering of zelfs AutoML aanbieden, zoals Microsoft Azure, Google Cloud, Amazon AWS en IBM Watson.

De term 'tester' is dus een erg rekbaar begrip. Als men bij 'tester' nog denkt aan de traditionele testrol, kan het gebeuren dat er pas laat in het project wordt gedacht aan kwaliteitsborging en aan het stellen van de echt kritische vragen. Deze whitepaper heet bewust 'Kwaliteit en testen van AI', niet alleen 'Testen van AI'. Bovendien gebruiken we bewust regelmatig de term QA-specialist naast de term 'tester'.

Een nieuwe vaardigheid is om je als tester te positioneren bij projecten waarbij AI wordt ontwikkeld. Op dit moment is het nog geen vanzelfsprekendheid dat er testers (op tijd) bij betrokken raken. Maar dat wil niet zeggen dat het niet kan!

8.4 Samen de kwaliteitsrol invulling geven

Het inzetten van testers in AI-projecten staat nog in de kinderschoenen. Doordat het relatief nieuw is, biedt het kansen om er zelf vorm aan te geven.

Persoonlijk, om jezelf verder te ontwikkelen in een meer technische of andere richting. Het helpt als je al ervaring hebt opgedaan bij eerdere AI-projecten. Maar sta je aan het begin van zo'n traject of project dan kan de kennis van anderen je helpen. Waarbij de rol en plaats van tester of kwaliteitswaarborger in een AI-project zeker ook van belang is.

En als groep kunnen testers en QA-functionarissen gezamenlijk de kwaliteitsrol binnen AI vervullen, waaronder het gebruik van nieuwe tools en technieken, het testen van ethiek, het ontwikkelen van best practices en cursussen voor verdere training. Deze aanpak maakt het mogelijk om gezamenlijk de kwaliteit en het testen in een AI-project te benadrukken, wat de adoptie van AI-oplossingen zal versnellen.

Deze whitepaper is dan ook een oproep aan alle kwaliteitsmedewerkers om elkaar te vinden, successen te delen en te leren van fouten. We zoeken de samenwerking graag op via communities als TestNet en EuroSTAR. Heb je ideeën over AI-testen of heb je er al ervaring mee, dan nodigen we je van harte uit om je inzichten en feedback met ons te delen. Contactgegevens zijn op de laatste pagina te vinden.

Bijlage A: Bronnen

Algemeen

- Human Compatible - Stuart Russell, 2019
  https://en.wikipedia.org/wiki/Human_Compatible
- The Next Decade in AI - Gary Marcus, 2020
  https://arxiv.org/ftp/arxiv/papers/2002/2002.06177.pdf
- Rebooting AI - Gary Marcus and Ernest Davis, 2019p
  http://rebooting.ai/
- Wat maakt AI testen anders - Peter Collewijn (werkgroep TestNet)
  https://www.testnet.org/testnet/download/common/2020-11-wat-maakt-ai-testen-anders.pdf
- Hidden Technical Debt in Machine Learning Systems, D. Sculley, 2015
- The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction, Eric Breck, 2017
- Testing machine learning based systems: a systematic mapping, Vincenzo Riccio, 2020
- Why Should I Trust You?", Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro
- Combinatorial Testing for Deep Learning Systems, Lei Ma, 2018
  https://arxiv.org/abs/1806.07723
- Effective testing for machine learning systems, Jeremy Jordan, 2020
  https://www.jeremyjordan.me/testing-ml/
- A Software Testing View on Machine Learning Model Quality, Christian Kästner, 2020
  https://ckaestne.medium.com/a-software-testing-view-on-machine-learning-model-quality-d508cb9e20a6
- Continuous Delivery for Machine Learning, Martin Fowler.com
  https://martinfowler.com/articles/cd4ml.html
- Snorkel Intro Tutorial: Data Slicing
  https://www.snorkel.org/use-cases/03-spam-data-slicing-tutorial
- Towards Robust and Verified AI: Specification Testing, Robust Training, and Formal Verification
  https://deepmind.com/blog/article/robust-and-verified-ai
- Interpretable Machine Learning, Christoph Molnar, 2020
  https://christophm.github.io/book/
- Discrimination, AI and Algorithmic Decision-Making,. Frederik Zuiderveen Borgesius, 2018,
  https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73
- Metamorphic Testing of Machine-Learning Based Systems, Teemu Kanstrén, 2020
  https://towardsdatascience.com/metamorphic-testing-of-machine-learning-based-systems-e1fe13baf048
- Katten selectie van Cassy Kozyrkov
  https://towardsdatascience.com/in-ai-the-objective-is-subjective-4614795d179b
- Decision trees versus neural networks
  http://www.312analytics.com/decision-trees-vs-neural-networks/

Kwaliteitsnormen

- Testing in the digital age, Sogeti Nederland B.V. ISBN: 978-90-75414-87-5
- Quality and AI-based Systems with Adam Leon Smith (ISO 25059)
  https://youtu.be/OadJbNeTmiY
- Artificial Intelligence – Life Cycle Processes and Quality Requirements (DIN SPEC 92001-1)

Ethiek

- AI heeft geen stekker meer, Rudy van Belkom
  https://detoekomstvanai.nl/artikelen/ai-heeft-geen-stekker-meer/
- Ethics guidelines for trustworthy AI
  https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- Regulatory framework proposal on Artificial Intelligence
  https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Checklist

- Checklist for Data Science Research Review
  https://medium.com/@ptannor/checklist-for-data-science-research-review-8a817b50697b
- Checklist for Artificial Intelligence in Medical Imaging (CLAIM)
  https://pubs.rsna.org/doi/10.1148/ryai.2020200029
- AI Checklist Cards
  https://www.tmforum.org/resources/reference/ai-checklist-cards/
- Chatbot test
  https://chatbottest.com/

Assessments

- EU: https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence
- ECP: https://ecp.nl/publicatie/artificial-intelligence-impact-assessment-volledige-versie/
  https://ecp.nl/publicatie/artificial-intelligence-impact-assessment-english-version/
- Pair: https://pair.withgoogle.com/

Trainingen

- Ai United Certified Tester in AI (CTAI)
  https://www.ai-united.org/
- A4Q, AI and Software Testing
  https://www.alliance4qualification.info/a4q-ai-and-software-testing
- KSTQB & CSTQB, Certified Tester AI Testing (Testing AI-Based Systems)
  https://imbus.cn/upFile/Uploadfiles/AI%20Testing_Testing%20AI-Based%20System%20Syllabus%20v1.3.pdf
- Datascience Academy, Testing & Monitoring ML Deployments
  https://data-science-academy3.teachable.com/p/testing-monitoring-machine-learning-model-deployments

Nieuwsartikelen

- Uber moet account chauffeur herstellen na beschuldiging fraude
  https://nos.nl/artikel/2376697-uber-moet-account-chauffeur-herstellen-na-beschuldiging-fraude.html
- Tesla's geweerd bij Chinese legerbases wegens bezorgdheid over spionage
  https://www.nu.nl/tech/6122885/teslas-geweerd-bij-chinese-legerbases-wegens-bezorgdheid-over-spionage.html
- Rapport: gezichtsherkenning Britse politie faalt in vier van vijf gevallen
  https://tweakers.net/nieuws/154870/rapport-gezichtsherkenning-britse-politie-faalt-in-vier-van-vijf-gevallen.html
- Rookgordijn rondom fraudesysteem Nissewaard
  https://webcache.googleusercontent.com/search?q=cache:VHqc2uuOkTMJ:https://www.binnenlandsbestuur.nl/sociaal/nieuws/rookgordijn-rondom-fraudesysteem-nissewaard.13026523.lynkx+&cd=4&hl=nl&ct=clnk&gl=nl
- Microsoft spreekt van gecoördineerde Tweet-aanval tegen chatbot
  https://webcache.googleusercontent.com/search?q=cache:8f1teYguutsJ:https://tweakers.net/nieuws/109711/microsoft-spreekt-van-gecoordineerde-tweet-aanval-tegen-chatbot.html+&cd=7&hl=nl&ct=clnk&gl=nl
- Google just gave a stunning demo of Assistant making an actual phone call
  https://www.theverge.com/2018/5/8/17332070/google-assistant-makes-phone-call-demo-duplex-io-2018
- De impact van gezichtsherkenning: een gezicht als bewijs voor criminaliteit
  https://www.nu.nl/tech-achtergrond/6121506/de-impact-van-gezichtsherkenning-een-gezicht-als-bewijs-voor-criminaliteit.html
- Anti-fraudesysteem SyRI moet van tafel, overheid maakt inbreuk op privéleven
  https://nos.nl/artikel/2321704-anti-fraudesysteem-syri-moet-van-tafel-overheid-maakt-inbreuk-op-priveleven.html
- Algoritmes zoeken naar bijstandsfraudeurs, welke rol speelt etnisch profileren
  ttps://nos.nl/artikel/2366962-algoritmes-zoeken-naar-bijstand-fraudeurs-welke-rol-speelt-etnisch-profileren.html
- GPT-3 Facts
  https://en.wikipedia.org/wiki/GPT-3
- Amazon scraps secret AI recruiting tool that showed bias against women
  https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
- Top 10 Reasons Why 87% of Machine Learning Projects Fail
  https://dzone.com/articles/top-10-reasons-why-87-of-the-machine-learning-proj
- Why So Many Data Science Projects Fail to Deliver
  https://sloanreview.mit.edu/article/why-so-many-data-science-projects-fail-to-deliver/?utm_source=ONTRAPORT-email-broadcast&utm_medium=ONTRAPORT-email-broadcast&utm_term=&utm_content=Data+Science+Insider%3A+March+5th%2C+2021&utm_campaign=06032021
- Google assistant phones hairdresser
  https://www.youtube.com/watch?v=JvbHu_bVa_g

Bijlage B: Woordenlijst

Begrip	Toelichting met linkjes naar tekst in de hoofdstukken
A/B testen	Zie Hoofdstuk Testen van AI
Acceptant	Zie Hoofdstuk Testen van AI in de praktijk
Adversarial testen	Zie hoofdstuk Ethiek
Amazon AWS	Cloud omgeving van Amazon
AutoML	Zie Algemene risico's bij uitdagingen voor testers
Beginselethiek	Zie Hoofdstuk Ethiek
Bias	Zie bijlage C
Bias trade off	Zie bijlage C
Boundary Value Analysis	Zie Hoofdstuk Testen van functionaliteit
Configuratietest	Zie Hoofdstuk Hoe test van AI
Convolutional Neural Network (CNN)	Zie Hoofdstuk 3 Beeldherkenning
Corner test	Zie Hoofdstuk Testen van functionaliteit
Classificatie	Classificatie is het indelen van items in voor gedefinieerde categorieën. Zie als voorbeeld Patroonherkenning, Beeldherkenning en Sequentieherkenning uit hoofdstuk Verschijningsvorm van AI.
Datacombinatietest	Zie Equivalence Partitioning
Datascientist	Zie Testen van AI in de praktijk
Deep learning	Zie hoofdstuk Artificial Intelligence
Deep fake	Zie Deep fake
Deugdethiek	Zie Hoofdstuk Ethiek
DIN SPEC 92001-1	Zie hoofdstuk Kwaliteitsattributen

Begrip	Toelichting met linkjes naar tekst in de hoofdstukken
Directional Expectation Test	Zie Hoofdstuk Testen van AI, Metamorphic testen
Drift	Zie hoofdstuk Testen van AI,Testen op drift
Equivalence Partitioning	Zie Hoofdstuk Testen van AI, Equivalence Partitioning
Expertpanel test	Zie Hoofdstuk Testen van functionaliteit
Explainable AI	Het uitleggen van een model. Zie Het testen van een model en Bijlage D
Exploratory testing	Zie Hoofdstuk Testen van Functionaliteit
Extrapolatie	Zie verschijningsvorm regressie
FLOPS	Zie Tekstgeneratie FLOPS is een eenheid die wordt gebruikt om de rekenkracht van CPU's aan te duiden. https://en.wikipedia.org/wiki/FLOPS
GAN	https://en.wikipedia.org/wiki/Generative_adversarial_network
Gevolgenethiek	Zie Hoofdstuk Ethiek
Google Cloud	Cloud omgeving van Google
Heatmap	zie Hoofdstuk Testen van AI, testen van het model zie Hoofdstuk Verschijningsvormen, beeldherkenning
High-Level Expert Group on Artificial Intelligence	Zie hoofdstuk Ethiek
IBM Watson	Cloud omgeving van IBM met AI toepassingen en tools
Input variabelen	Zie Hoofdstuk 1 Artificial Intelligence ook wel invoervariabelen genoemd
Invariance Test	Zie Hoofdstuk Testen van ai, Metamorphic testen
ISO 25010	Zie Hoofdstuk Kwaliteitsattributen. Meer informatie: https://iso25000.com/index.php/en/
Keras	Software bibliotheek met veel gebruikte ML functies
Kwaliteitsattributen / Kwaliteitskenmerken	Zie Hoofdstuk Verschillende kwaliteitsattributen

Begrip	Toelichting met linkjes naar tekst in de hoofdstukken
Kwaliteitsnormen	Zie Bijlage A: Bronnen
Labels	Zie Algemene risico's
Lineaire regressie	Zie bijlage C
Metamorphic testen	Zie Hoofdstuk Testen van ai, Metamorphic testen
Microsoft Azure	Cloud omgeving van Microsoft
ML-engineer	Zie Testen van AI in de praktijk
Neuraal Netwerk	Zie hoofdstuk Artificial Intelligence
OpenAI	Zie Tekstgeneratie
Outliers	Zie Afhankelijkheid van data
Overfitting	Zie bijlage C
Pairwise testing	Zie Hoofdstuk Testen van ai, Equivalence Partitioning
Persona	Een persona is een archetype van een gebruiker, ofwel een karakterisering van een bepaald type gebruiker. Zie Testen vanuit Persona's.
Preprocessing	Zie Testen van data
Polynomiale regressie	Zie bijlage C
Python	Programmeertaal, vaak gebruikt voor ML development
Pythorch	Deep Learning Framework voor het configureren van DL modellen
R of Rstudio	Programmeertaal vaak gebruikt voor ML development met name in een Academische omgeving
Regressie	Zie Hoofdstuk Verschijningsvormen
Regressietesten	Zie Hoofdstuk Testen van AI
Robuustheid	Zie Robuuste AI
Scikit-learn	Software bibliotheek met veel gebruikte ML functies
Sequentieherkenning	Zie hoofdstuk Verschijningsvormen.

Begrip	Toelichting met linkjes naar tekst in de hoofdstukken
TensorFlow	Deep Learning Framework voor het configureren van DL modellen
Testdata	Zie bijlage C
Test oracle	Zie Hoofdstuk 1
Testen op Drift	Zie hoofdstuk Testen van AI,Testen op drift
Testen van een model	Zie Hoofdstuk Hoe test je AI
Testen vanuit persona's	Zie Hoofdstuk Testen van functionaliteit
TMAP	Zie Hoofdstuk Gradaties in Autonomie
Token	Zie Tekstgeneratie
Trainingen	Zie Bijlage A
Trainingsdata	Zie bijlage C
Transfer learning	Transfer learning is the reuse of a pre-trained model on a new problem. It's currently very popular in deep learning because it can train deep neural networks with comparatively little data.https://builtin.com/data-science/transfer-learning
Underfitting	Zie bijlage C
User Story test - Use Case test	User Story testen – Use Case Testen
Variance	Zie bijlage C

Bijlage C: Trainen en Evalueren

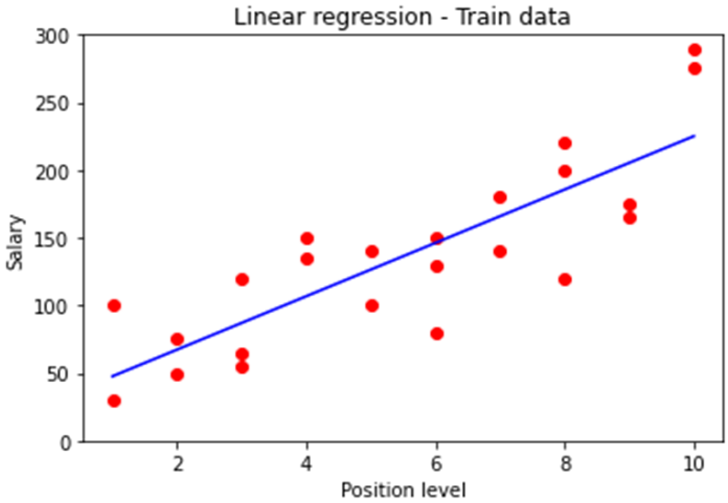
Een Machine Learning algoritme is gebaseerd op voorbeelden, data die in het verleden een zeker resultaat tot gevolg hebben gehad. Die voorbeelden zijn per definitie slechts een deel van de werkelijkheid. Om een Machine Learning model te ontwikkelen heb je (veel) data nodig. Hoe meer data, hoe beter de kans is dat er een algoritme gemaakt kan worden dat een goede voorspelling doet van de aangereikte voorbeelden die we het model hebben aangereikt.

In bijgaand voorbeeld gebruiken we als data het salaris afgezet tegen de salarisschaal.

Het is gebruikelijk om deze data te verdelen in twee delen:

- Een deel dat gebruikt wordt om het model te trainen (rood)
- Een deel dat gebruikt wordt om het model te testen (groen)

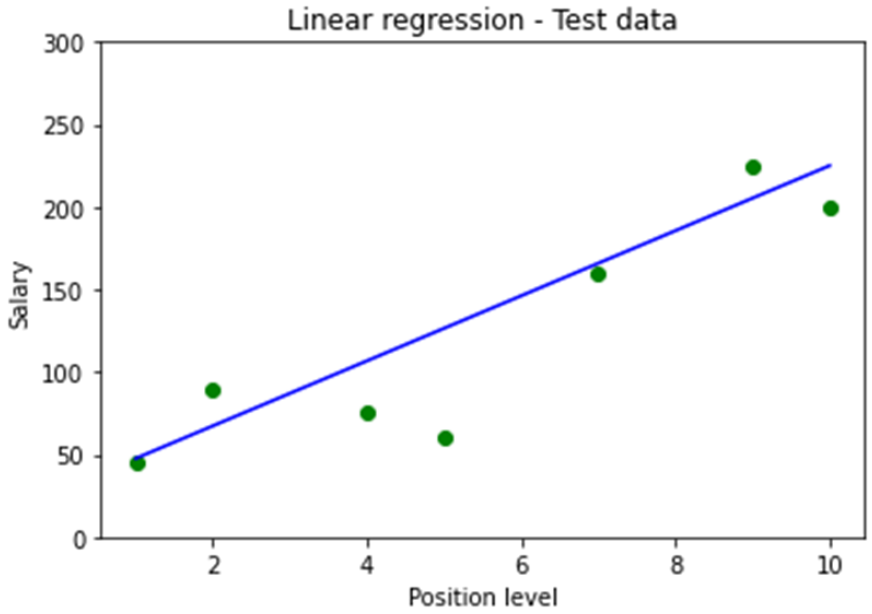
Na het trainen van het model wordt getest hoe goed, hoe nauwkeurig, het model is. Dit is een soort unit test op het model uitgevoerd door een ML engineer. In ons voorbeeld levert het model een rechte lijn op en op basis van de trainingsdata, de rode punten, krijgen we de volgende grafiek:



Deze lijn voorspelt voor ongeveer 70% nauwkeurig het patroon van de trainingsdata.

Echter, deze data is gebaseerd op trainingsdata, het model weet hier de uitkomst al van. Het is daarom van belang te weten hoe goed dit model presteert op basis van de testdata. Dat is immers een indicatie hoe goed het model gaat werken in een productie omgeving.

In de volgende grafiek wordt het model, de lijn, afgezet tegen de testdata.



Het model blijkt op de testdata zelfs beter te presteren. Na berekening blijkt het model voor 78% nauwkeurig het patroon van de test data te voorspellen.

Uiteraard is het doel om zo dicht mogelijk bij de 100% te komen. Het verschil tussen deze 100% en de werkelijke nauwkeurigheid (zie dit als de fout), wordt **Bias** genoemd.

Het streven is deze bias zoveel mogelijk te verlagen en dus het verschil tussen een willekeurige voorspelling en de juiste voorspelling van onze voorbeelden zo klein mogelijk te maken. Hiermee beperken we **Underfitting** (ofwel we zorgen dat het model beter past bij de voorbeelden).

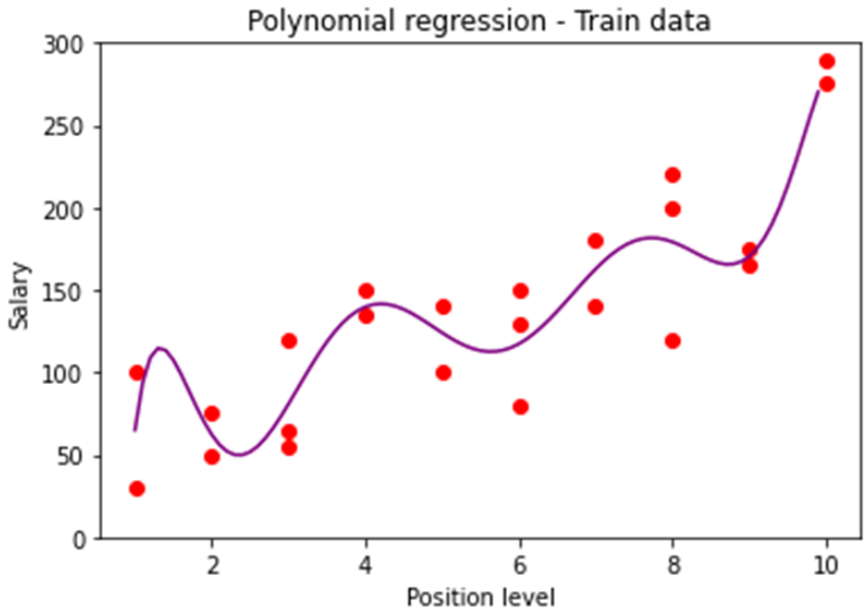
In ons voorbeeld is dus een 22% bias (fout) op de test data, en 30% bias op de trainingsdata.

Het verschil in bias tussen deze twee datasets wordt **Variance** genoemd. In ons voorbeeld is de variance 30% - 22% = 8%.

Zoals aangegeven is het doel een model te ontwikkelen dat zo dicht mogelijk tegen de 100% nauwkeurigheid aankomt. Het feit dat het model met de testdata al beter presteert is ook een indicatie dat er ruimte voor verbetering is. Het model is nu **Underfit**.

We kunnen een andere trainingsmethode proberen.

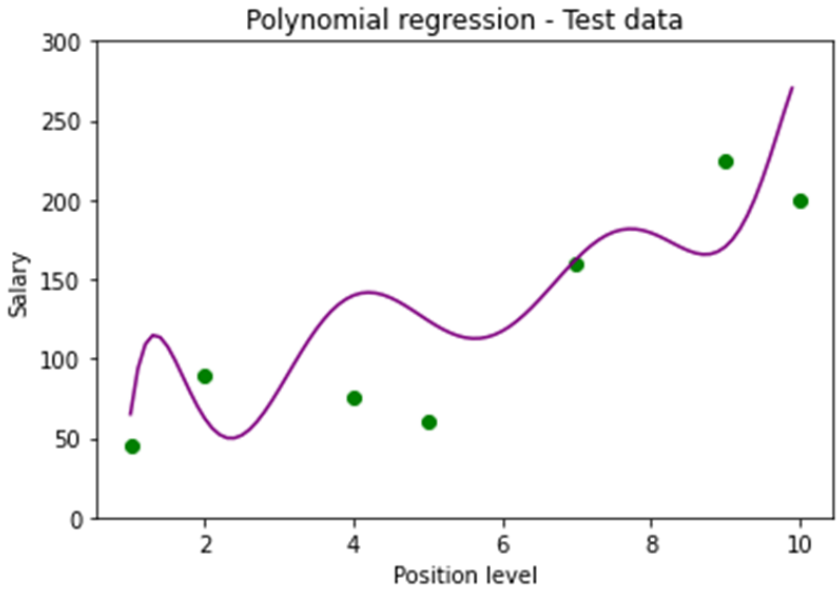
Door deze andere aanpak voorspelt het model nu een lijn die het patroon van de trainingsdata beter volgt. Zie onderstaande grafiek van de trainingsdata.



Dit model heeft een nauwkeurigheid van 84%. Dit is een grote verbetering t.o.v. de 70% van het eerste model.

Echter, de vraag is wederom: hoe goed presteert het model op de testdata?

Dat valt helaas tegen, zie het bewijs in onderstaande grafiek:



De nauwkeurigheid is nu slechts 38%.

In een situatie zoals in dit voorbeeld:

- het model presteert goed met de trainingsdata, echter
- het model presteert slecht met de testdata,

spreekt men van een model dat **Overfit** is.

Vergelijk het met het uit het hoofd leren van alle antwoorden van een proef examen. Dat is meestal ook geen goede leermethode om te slagen bij het werkelijke examen.

Eveneens is de Variance van dit model erg groot. Het verschil is nu 84% - 38% = 46%.

We moeten dus oppassen dat we het algoritme niet al te veel vormen naar de voorbeelden die we hebben. We willen tenslotte dat het algoritme goed generaliseert, zodat het ook goed scoort op de gevallen die we niet als voorbeeld hebben gebruikt. Hiermee verlagen we de variance (het verschil tussen de voorspelling van de gekozen voorbeelden en de voorspelling van de complete werkelijkheid) en beperken we overfitting (het model voorspelt vooral de voorbeelden goed, maar is minder goed in het voorspellen van de complete werkelijkheid).

Conclusie:

We zijn dus op zoek naar een model met een zo hoog mogelijke nauwkeurigheid. Of met andere woorden een zo klein mogelijke Bias.

Eveneens willen we dat deze nauwkeurigheid generiek is, dus dat er weinig verschil in deze nauwkeurigheid is bij verschillende datasets. Dus een zo klein mogelijke Variance.

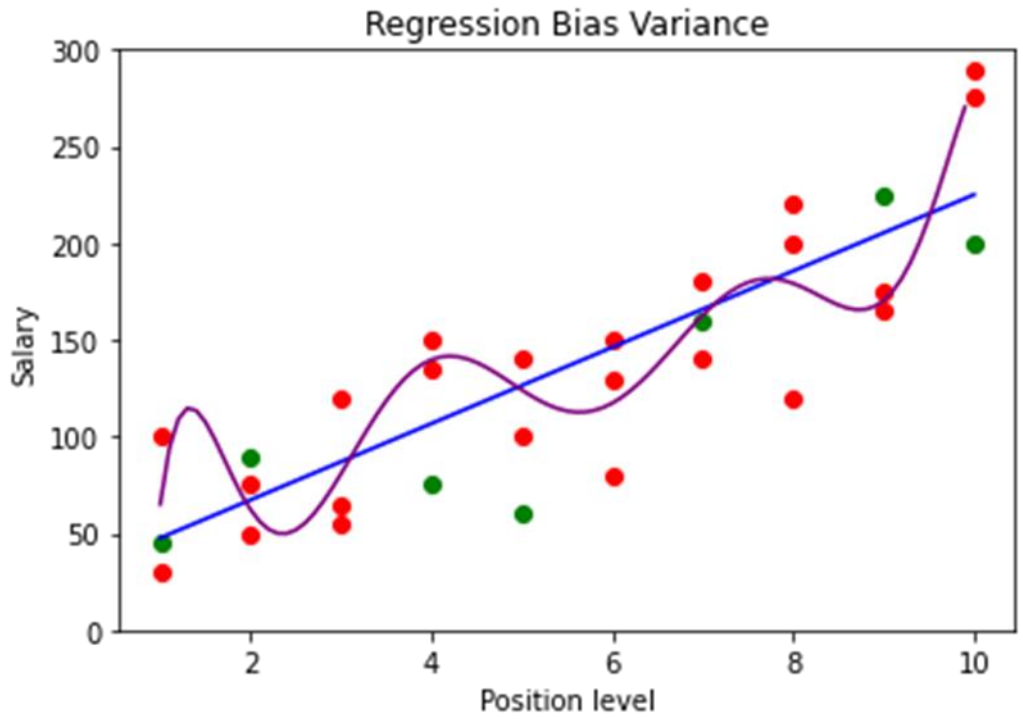
Als je een model beter of langer gaat trainen is het mogelijk dat de nauwkeurigheid op de trainingsdata beter wordt. De Bias van het model op deze data wordt minder.

Echter, vaak wordt dan het verschil met de nauwkeurigheid van de testdata groter. De Variance wordt dus groter.

Dit verschijnsel wordt de **Bias- Variance Trade Off** genoemd.

Het doel is dus een model te maken met een zo hoog mogelijke nauwkeurigheid én dat een vergelijkbare nauwkeurigheid heeft met zowel de trainings- als test-data.

Dat is immers een goede indicatie hoe het model in productie zal presteren.



Bijlage D: Risico's en Testen

In verschillende hoofdstukken zijn de risico's en aandachtspunten van AI-ontwikkeling besproken. In het hoofdstuk 'Testen van AI', wordt stapsgewijs aangegeven welke kwaliteits- en testactiviteiten bijdragen aan het wegnemen van de risico's. In deze bijlage is een koppeling tussen risico's en testactiviteiten beschreven.

Onzekere uitkomsten

Grip krijgen op onzekere uitkomsten kan deels met bestaande technieken zoals Equivalence Partitioning en Boundary Value Analysis, waarmee helderheid verkregen kan worden over grijze gebieden in voorspellingen. Inhoudelijke experts kunnen daarmee bepalen welke resultaten het model in elk van de gedefinieerde situaties zou moeten geven. Ook kan een persona's test inzicht geven in welke mate de uitkomst afhankelijk is van de persoonsgroep.

Afhankelijkheid van data

Het testen van data blijft één van de kernpunten bij het testen van AI. Data kan worden beoordeeld op de manier waarop het is verzameld, geselecteerd, bewerkt en ingelezen. Ethische overwegingen zijn daarbij vaak van groot belang. De whitepaper heeft daar een hoofdstuk aan besteed, in de hoop dat dit helpt om zelf de ethische overweging naar de eigen praktijk te vertalen.

Verder is het goed om te beseffen dat Equivalence Partitioning, Boundary Value Analysis en andere technieken voor data-analyse ook op de verzamelde (invoer-)data kan worden toegepast. Ook hier kunnen experts helpen om deze data te beoordelen op inhoudelijke juistheid en onderlinge verhouding.

Beperkte uitlegbaarheid

ML-modellen, en dan met name de DL-modellen zullen altijd moeilijk uitlegbaar blijken, maar er bestaat altijd de mogelijkheid om deze voor een gedeelte te interpreteren. Het uitleggen van AI, of te wel eXplainable AI is een gebied dat nog volop in ontwikkeling is. Vanuit traditoneel testprerpectief ligt het voor de hand om minimaal een aantal blackbox-technieken zoals use case testen, exploratory testen of persona's toe te passen. Metamorphic testing is ook mogelijk, omdat het bewaken van logische relaties tussen invoer en uitvoer heel belangrijk is voor de begrijpelijkheid en daarmee uitlegbaarheid. Daar waar whitebox-technieken mogelijk zijn, is dat natuurlijk altijd een meerwaarde. Denk aan het voorbeeld in de whitepaper over de Dakar-auto in het zand.

Dit soort testen verhoogt het vertrouwen dat het AI-model begrijpelijk functioneert en zorgt voor een acceptatie van het gebruik.

Veranderende werkelijkheid of behoefte

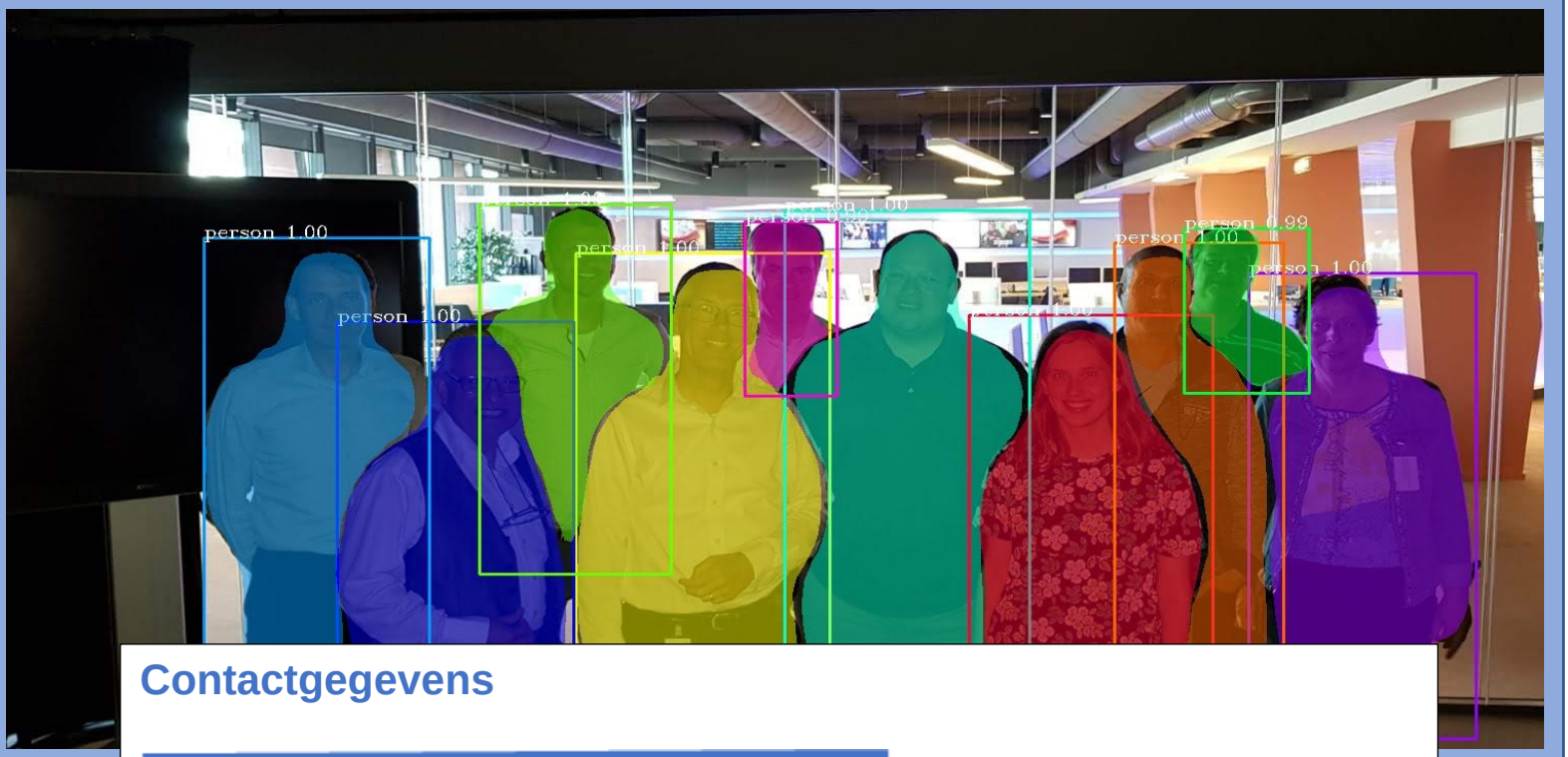
De veranderende werkelijkheid of behoefte staat bekend als drift en is in het hoofdstuk 'Testen van AI' kort beschreven. Met behulp van A/B-testen is het mogelijk om dit risico van veranderende gedrag in beeld te krijgen.

Deze testen gaan gepaard met monitoring, zoals een wijziging in de verhouding van de invoergegevens of de nauwkeurigheid van de voorspelling. In de whitepaper is hier beperkt op ingegaan maar het spreekt voor zich dat monitoring ervoor zorgt dat het AI-model in praktijk acceptabel blijft. Aangezien verschillende groepen belanghebbenden andere scores belangrijk vinden, kan het testen met persona's helpen om alle groepen blijvend tevreden te houden.

A/B-testen kunnen ook worden gebruik om een AI-model in te praktijk te controleren op veranderende behoefte. Dit is uiteraard ook een vorm van gedrag, maar in dit geval toont het aan dat de interesse in het algemeen is verschoven, bijvoorbeeld als men op een videoplatform een bepaalde categorie films opeens veel meer aanklikt. Dit zijn subtiele veranderingen in behoefte. Als vanuit de eigen organisatie een grote verandering van behoefte ontstaat, bijvoorbeeld het genoemde voorbeeld over het herkennen van andere soorten ruitschade, dan is regressietesten een belangrijke activiteit.

Algemene angst voor AI

Nieuwe technieken leiden in het begin vrijwel altijd voor angstgevoelens. De belangrijkste activiteit om angst weg te nemen is het geven van de juiste inzichten. Een van de doelstellingen van de EU is het gebruik van AI te stimuleren en tegelijkertijd de angst weg te nemen door ethisch richtlijnen voor te stellen en op termijn ook verplicht te stellen. Om inzicht te krijgen in de werking van het model zijn bij deze richtlijnen checklist en assessments opgesteld. Eveneens kunnen exploratory testen, het testen met persona's ook een rol vervullen bij het krijgen van inzicht in de werking van een AI-model. Maar feitelijk draagt elk uitgevoerde test bij aan het verhogen van dit inzicht en levert dus een bijdrage aan het verminderen van het risico angst voor AI.



Contactgegevens

Werkgroep Testen en AI

Sander Mol
Peter Collewijn
Hannie van Kooten

email: ai.workgroup.testnet@gmail.com
site: https://www.testnet.org/testnet/p000610/werkgroepen/werkgroep-testen-en-ai

TestNet

site: https://www.testnet.org/testnet/home