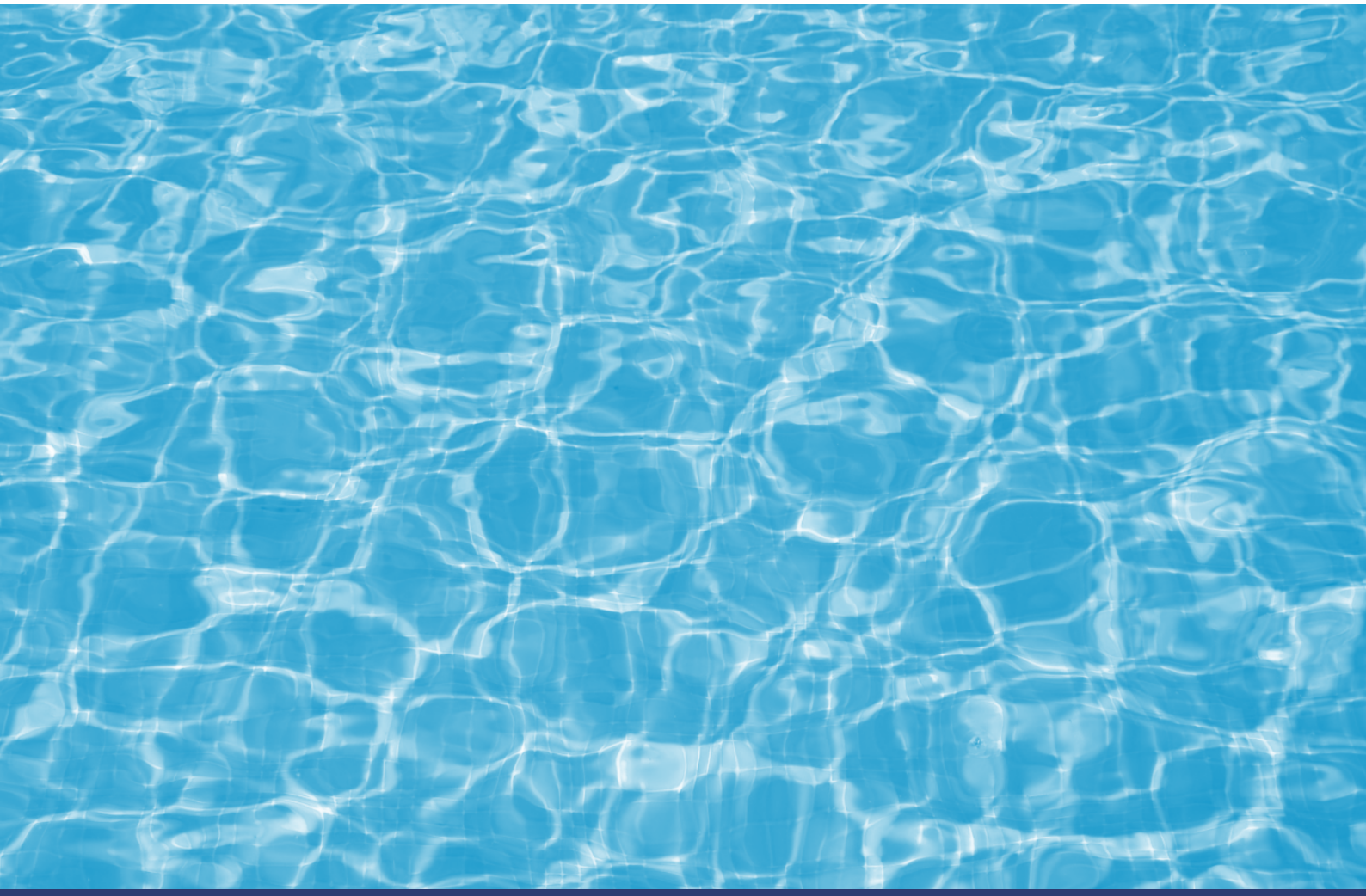
CONSIDERATI



Onvoorziene effecten van zelflerende algoritmen

Considerati

14 September 2020

Bart Schermer PhD Joas van Ham MSc Ka Wing Falkena LL.M

Contact

info@considerati.com

020 73 70 069

De Boelelaan 7, Amsterdam

www.considerati.com

Inhoudsopgave

Managementsamenvatting **6**

1 Inleiding **8**

1.1 Doelstelling 8

1.2 Vraagstelling 8

1.3 Aanpak 9

1.4 Scope 10

1.5 Leeswijzer en hoofdstukindeling 10

2 Kunstmatige intelligentie, data-analyse en algoritmen **12**

2.1 Wat is kunstmatige intelligentie? 12

2.2 Hoe 'begrijpt' kunstmatige intelligentie 13

2.3 Symbolic AI 14

2.4 Machine learning 14

2.5 Leren in *real-time*? 20

2.6 Probleemruimte en de situering van kunstmatige intelligentie 20

2.7 Emergent gedrag 22

2.8 Begrijpelijkheid en transparantie 24

2.9 Beperkingen van machine learning 25

2.10 Tussenconclusie 26

3 Kunstmatige intelligentie in de praktijk **28**

3.1 Het proces van data-analyse 28

3.2 Data science binnen de context van een organisatie. 32

3.3 Socio-technische dimensie van kunstmatige intelligentie en relevante actoren 32

3.4 De rol van de mens bij de toepassing van kunstmatige intelligentie 34

4 Onvoorziene effecten **36**

4.1 Voorziene en onvoorziene effecten 36

4.2 Oorzaken onvoorziene effecten 38

4.3 Relevantie voor de toepassing van algoritmen 39

5 Overzicht casestudies **41**

6 Case study 1: Algorithmic pricing **43**

6.1 Doel algorithmic pricing 44

6.2 Implementatie en inzet ... 45
6.3 Impact van algorithmic pricing op prijsstelling ... 47
6.4 Potentiële onvoorziene effecten van algorithmic pricing ... 48
6.5 Algorithmic pricing in de toekomst ... 52

7 Case study 2: Kredietwaardigheid en risicoprofielen ... 54
7.1 Doel kredietwaardigheidstoetsen ... 54
7.2 Implementatie en inzet ... 55
7.3 Impact AI voor bepalen kredietwaardigheid ... 55
7.4 Potentiële onvoorziene effecten van algoritmen op kredietwaardigheidstoetsen ... 57

8 Case study 3: HR Analytics ... 63
8.1 Doel HR analytics ... 64
8.2 Implementatie en inzet HR analytics ... 64
8.3 Impact van HR analytics op recruitment en employee performance ... 66
8.4 Potentiële onvoorziene effecten van HR analytics ... 67

9 Case study 4: Fysieke & geestelijke gezondheid ... 72
9.1 Doel AI voor fysieke & geestelijke gezondheid ... 73
9.2 Implementatie en inzet ... 74
9.3 Impact van algoritmen op gezondheidsadvies ... 74
9.4 Potentiële onvoorziene effecten op fysieke & geestelijke gezondheid ... 76

10 Analyse ... 78
10.1 Grondoorzaken onvoorziene effecten ... 78
10.2 Ondoorgrondelijkheid van zelflerende algoritmen ... 81
10.3 Epistemische en normatieve zorgen bij algoritmische besluitvorming ... 83
10.4 Overzicht onvoorziene effecten ... 85
10.5 Effecten op de korte-, middellange en lange termijn ... 89
10.6 Afsluitende beschouwing ... 89

11 Inzicht en mitigatie ongewenste effecten ... 91
11.1 Organisatorische maatregelen / governance ... 91
11.2 Technische maatregelen ... 98
11.3 (Extern) toezicht ... 106
11.4 Bescherming van het subject (de consument) ... 108

12 Samenvatting en conclusies 110

13 Bibliografie 114

14 Bijlagen 119

Bijlage 1: Overzicht geïnterviewden 119

Bijlage 2: Leidraad interviews 120

Managementsamenvatting

(Zelflerende) algoritmen worden ingezet om meerwaarde te leveren aan bedrijven, consumenten en de samenleving als geheel. Bij de toepassing van (zelflerende) algoritmen kunnen onvoorziene effecten optreden. De impact van deze onvoorziene effecten is, afhankelijk van de context en de aard van de toepassingen, potentieel groot. Dit onderzoek richt zich daarom op de vraag:

> *Wat zijn mogelijke onvoorziene effecten van de inzet van (zelflerende) algoritmen door bedrijven en consumenten waarvan niet duidelijk is hoe zij tot een besluit komen en hoe kunnen deze effecten geïdentificeerd, gewogen en indien ongewenst gemitigeerd worden?*

Deze vraag beantwoorden we aan de hand van literatuuronderzoek en interviews, gericht op vier cases: het bepalen van prijzen met behulp van algoritmen (*algorithmic pricing*), het beoordelen van de betrouwbaarheid van een persoon (fraude, kredietwaardigheid); het beoordelen van de geschiktheid van een persoon (*HR analytics*) en gepersonaliseerde advisering (fysieke & geestelijke gezondheid).

Uit de literatuur en de casestudies blijkt dat onvoorziene effecten sterk afhankelijk zijn van de context waarbinnen algoritmen worden toegepast. Verder blijkt dat onvoorziene effecten doorgaans negatief zijn. Dit heeft minder te maken met het gebruik van algoritmen, dan met het feit dat onvoorziene effecten door hun onvoorspelbaarheid doorgaans eerder negatief dan positief van aard zijn. Het is daarmee ook van belang om te onderstrepen dat dit niet een onderzoek betreft naar de positieve of negatieve effecten van algoritmen. In zoverre negatieve effecten van algoritmen de boventoon voeren in dit rapport, is dit omdat de focus ligt op het beschrijven van de onvoorziene effecten. De bedoelde effecten van algoritmen zullen doorgaans positief van aard zijn, dit is immers waarom ze worden ingezet.

Op basis van het onderzoek kunnen we stellen dat er drie ‘grondoorzaken’ zijn voor het optreden van onvoorziene effecten in de context van de toepassing van (zelf)lerende algoritmen.

1) Er is een onvolledig of verkeerd begrip van de probleemruimte.
2) Het zelflerende algoritme is niet goed toegerust om om te gaan met de complexe omgeving waarbinnen het wordt ingezet.
3) Het (zelflerende) algoritme wordt niet goed ingepast in een bredere (socio-technische) context.

Om de kans op onvoorziene effecten van (zelflerende) algoritmen te verkleinen is het zaak om de grondoorzaken weg te nemen die leiden tot onvoorziene effecten. We identificeren maatregelen die in de verschillende stadia van het ontwikkelproces van modellen relevant zijn. Samengevat gaat het om organisatorische maatregelen, technische maatregelen, intern en extern toezicht, en het beschermen van de rechtspositie van subjecten (consumenten).

Gebruikte afkortingen

ACM	Autoriteit Consument en Market
AI	Artificial intelligence (kunstmatige intelligentie)
AIIA	Artificial Intelligence Impact Assessment
API	Application Programming Interface
AVG	Algemene Verordening Gegevensbescherming
BKR	Stichting Bureau Kredietregistratie
CRISP-DM	Cross Industry Standard Process for Data Mining
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
ECP	Electronic Commerce Platform
FG	Functionaris Gegevensbescherming
ICE	Individual Conditional Expectations
ICO	Information Commissioner's Office
IEEE	Institute of Electrical and Electronics Engineers
ISACA	Information Systems Audit and Control Association
LIME	Local Interpretable Model-Agnostic Explanations
ML	Machine Learning
PDPC	Personal Data Protection Commission
SHAP	Shapley Additive Explanations

1 Inleiding

De toepassing van kunstmatige intelligentie staat volop in de maatschappelijke belangstelling. Enerzijds is er het hoopvolle toekomstbeeld dat kunstmatige intelligentie onze welvaart en welbevinden vergroot, anderzijds is er de angst dat kunstmatige intelligentie verkeerde beslissingen neemt die nadelige effecten hebben op individuen, groepen en de maatschappij als geheel.

In dit onderzoek staat de vraag centraal wat de mogelijke onvoorziene effecten zijn van de toepassing van kunstmatige intelligentie in het algemeen en de toepassing van 'zelflerende algoritmen' in het bijzonder. Dit onderzoek is (mede) gedaan om uitvoer te geven aan de motie van de leden van den Berg en Wörsdörfer.[1]

De motie roept de regering op om na te gaan welke ongewenste effecten zich kunnen voordoen bij de inzet van algoritmen door bedrijven en hoe deze effecten voorkomen kunnen worden. Hierbij dient in het bijzonder aandacht te worden besteed aan die algoritmen waarvan niet eenduidig te volgen is hoe de uitkomst tot stand komt.

1.1 Doelstelling

Het doel van dit onderzoek is het verschaffen van inzicht in de onvoorziene effecten van de inzet van (zelflerende) algoritmen door private partijen. In het bijzonder de effecten van de inzet van algoritmen waarvan niet duidelijk is hoe zij beslissingen nemen. Onvoorziene effecten kunnen ongewenst zijn. Bij ongewenste effecten kan worden gedacht aan een verlies aan economische kansen of welvaart, ongelijke machtsverhoudingen in de relatie bedrijf-consument en de relatie bedrijf-bedrijf, stigmatisering van personen en verlies van autonomie.

1.2 Vraagstelling

De hoofdvraag van dit onderzoek is:

> *Wat zijn mogelijke onvoorziene effecten van de inzet van (zelflerende) algoritmen door bedrijven en consumenten waarvan niet duidelijk is hoe zij tot een besluit komen en hoe kunnen deze effecten geïdentificeerd, gewogen en indien ongewenst gemitigeerd worden?*

[1] Tweede Kamer der Staten-Generaal, vergaderjaar 2018-2019, 21 501-33, nr. 748

Deze vraag valt uiteen in de volgende deelvragen:

1. Tot welke onvoorziene effecten kan het gebruik van (zelflerende) algoritmen door private partijen leiden op de korte, middellange en lange termijn?
2. Hoe kunnen bedrijven en consumenten als gebruikers van algoritmen mogelijk onvoorziene /ongewenste effecten identificeren, voorkomen en/of mitigeren?
3. Hoe kunnen aanbieders en gebruikers van algoritmen omgaan met restrisico's?

1.3 Aanpak

Voor dit onderzoek hanteert Considerati de volgende aanpak:

Literatuurstudie

Op basis van literatuuronderzoek wordt eerst de context van de probleemstelling geschetst en afgebakend. De context van het onderzoek (op hoofdlijnen) is het gebruik van (zelflerende) algoritmen door bedrijven/consumenten, waarvan de besluitvorming niet (direct) voor mensen doorgrondelijk is.

Het literatuuronderzoek biedt een theoretisch kader dat vervolgens op de verschillende *case studies* wordt toegepast. Aan de hand van interviews worden de case studies verrijkt met voorbeelden en analyses uit de praktijk.

Case study onderzoek

Om de onderzoeksvragen te beantwoorden is gekozen voor *case study* analyse. Considerati heeft gekozen voor een exploratieve, *embedded case study* aanpak, waarbij vier cases worden geanalyseerd.[2]

Om tot een goed beeld te komen van mogelijke onvoorziene en ongewenste effecten hebben wij vier cases gekozen ter analyse. Uitgangspunten bij de selectie waren dat:

- de case een potentieel grote maatschappelijke en economische relevantie heeft;
- verschillende belangen en actoren een rol spelen, waardoor een zo compleet mogelijk beeld van mogelijke effecten wordt verkregen;
- complexe (zelflerende) algoritmen toegepast worden of kunnen worden;

[2] Yin, R. K. (2009). Case study research, design and methods, fourth edition, Sage publishing

- de kans dat onvoorziene effecten optreden is toe te schrijven aan, of groter wordt door, een gebrek aan transparantie en/of begrijpelijkheid van de beslissingen.

De vier cases zijn:

- Het bepalen van prijzen met behulp van algoritmen (algorithmic pricing)
- Beoordelen van de betrouwbaarheid van een persoon (fraude, kredietwaardigheid)
- Het beoordelen van de geschiktheid van een persoon (HR analytics)
- Gepersonaliseerde advisering (fysieke & geestelijke gezondheid)

1.4 Scope

In deze rapportage gaan wij uit van systemen die in staat zijn om zelfstandig (zonder menselijke tussenkomst) beslissingen te nemen, oftewel kunstmatige intelligentie. Binnen deze overkoepelende categorie kijken wij dan in het bijzonder naar (zelf)lerende algoritmen (*machine learning*) en hun onvoorziene effecten.[3] In deze rapportage wordt beperkt aandacht besteed aan de macro-economische effecten van de inzet van zelflerende algoritmen zoals bijvoorbeeld de invloed ervan op de arbeidsmarkt. De reden hiervoor is dat deze effecten doorgaans voorzien of voorzienbaar zijn, in de zin dat het doel van de inzet van algoritmen is om menselijke besluitvorming te vervangen.

1.5 Leeswijzer en hoofdstukindeling

In hoofdstuk 2 beschrijven wij het onderwerp waar deze rapportage betrekking op heeft: kunstmatige intelligentie. We gaan in op de verschillende vormen van kunstmatige intelligentie waarbij specifiek de nadruk wordt gelegd op machine learning, omdat dat het domein is waarbinnen (zelf)lerende algoritmen hun primaire toepassing vinden.

In hoofdstuk 3 lichten wij toe hoe algoritmen en beslismodellen in de praktijk worden ontwikkeld en toegepast.

In hoofdstuk 4 beschrijven wij wat onvoorziene effecten zijn en hoe deze kunnen ontstaan. Dit biedt ons meer inzicht in de vraag wat mogelijke oorzaken zijn waarom (zelf)lerende algoritmen onvoorziene effecten veroorzaken en welke effecten dit kunnen zijn.

[3] Omdat de wenselijkheid van een uitkomst of een effect subjectief kan zijn, gebruiken wij primair de term 'onvoorziene effecten'.

In hoofdstuk 5 bieden wij een overzicht van de case studies. Op basis van het theoretisch kader dat in de hoofdstukken 2 tot en met 4 uiteen is gezet kunnen wij deze cases gaan beschrijven en analyseren.

In hoofdstuk 6 tot en met 9 onderzoeken wij de vier geselecteerde case studies in detail. Voor elk van de cases onderzoeken wij hoe algoritmen worden ingezet. Op basis daarvan bekijken wij wat de mogelijke onvoorziene effecten zijn van deze toepassing en wat de impact daarvan is voor deze concrete domeinen.

In hoofdstuk 10 destilleren wij de belangrijkste conclusies en rode lijnen uit het case study onderzoek en stellen wij vast hoe onvoorziene effecten ontstaan. Naast de concrete bevindingen uit de cases zullen wij de onvoorziene effecten aanvullen met effecten die zijn beschreven in de literatuur.

In hoofdstuk 11 bekijken we voor de effecten die schadelijk of ongewenst zijn, welke risico-beperkende maatregelen kunnen worden geïmplementeerd om deze effecten uit te sluiten, dan wel het risico te verkleinen dat deze effecten zich voordoen. Hierbij kijken we niet alleen naar de ontwikkelaars en gebruikers van zelflerende algoritmen, maar ook naar degenen die aan de geautomatiseerde besluiten van deze toepassingen direct of indirect worden onderworpen (de consument).

In hoofdstuk 13 eindigen wij met een samenvatting en conclusies.

2 Kunstmatige intelligentie, data-analyse en algoritmen

Dit onderzoek heeft betrekking op de effecten van algoritmen. Wanneer in het populaire spraakgebruik wordt gesproken over 'algoritmen', 'slimme algoritmen' of 'zelflerende algoritmen', dan wordt doorgaans gedoeld op (elementen van) computersystemen die zelf beslissingen kunnen nemen, oftewel kunstmatige intelligentie. Voor een beter begrip van de context waarbinnen algoritmen worden toegepast zetten wij in dit hoofdstuk uiteen wat kunstmatige intelligentie is en welke plaats (zelflerende) algoritmen innemen binnen het bredere concept kunstmatige intelligentie.

2.1 Wat is kunstmatige intelligentie?

Om kunstmatige intelligentie helder te kunnen definiëren is het van belang dat we eerst definiëren wat we onder menselijke intelligentie verstaan. Psycholoog David Wechsler definieert menselijke intelligentie als volgt:

> *"The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment."*[4]

Oftewel, handelen met een doel, rationeel denken en effectief kunnen omgaan met de omgeving. Deze definitie verschilt niet veel van de definitie die de Europese Commissie geeft aan kunstmatige intelligentie:

> *"systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals."*[5]

Hier wordt ook benadrukt dat het gaat om handelen met een specifiek doel en het analyseren van de omgeving. Kunstmatige intelligentie wordt logischerwijs vergeleken met menselijke intelligentie, omdat kunstmatige intelligentie gemodelleerd wordt naar menselijke intelligentie. Het doel van kunstmatige intelligentie is doorgaans ook het evenaren of voorbijstreven van menselijke intelligentie.

Ten tijde van het schrijven van dit rapport is kunstmatige intelligentie op bepaalde gebieden de menselijke intelligentie reeds voorbijgestreefd (denk bijvoorbeeld aan schaken of het herkennen van gezichten), terwijl op andere vlakken de menselijke intelligentie nog steeds

[4] Wechsler, D. (1958). *The* Measurement *and Appraisal of Adult Intelligence*. Baltimore, MD: The Williams & Wilkins Company.

[5] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussel 25 april 2018, COM(2108) 237 final.

superieur is (bijvoorbeeld autorijden). De meeste AI-toepassingen worden echter slechts ontworpen en getraind om één taak uit te voeren. Algemene kunstmatige intelligentie (*artificial general intelligence*), die in staat is om allerlei taken op een (boven)menselijk niveau uit te voeren, bestaat (nog) niet.

Algoritmen vormen de bouwstenen voor kunstmatige intelligentie. Een algoritme is een eindige reeks instructies die leidt tot de berekening van een uitkomst voor een probleem. Een algoritme is als het ware een recept: door de verschillende stappen te volgen eindig je met het eindresultaat (de maaltijd). Een algoritme kan heel eenvoudig zijn, maar ook bijzonder complex. In het populaire spraakgebruik wordt de term algoritme, of 'slim algoritme' vaak gelijkgesteld met kunstmatige intelligentie.

2.2 Hoe 'begrijpt' kunstmatige intelligentie

Om effectief om te kunnen gaan met de omgeving heb je een zeker begrip nodig van de omgeving. De mate van intelligentie die een individu nodig heeft om keuzes te maken en handelingen uit te voeren is afhankelijk van de complexiteit van de omgeving en/of het op te lossen probleem.

Mensen creëren een begrip van hun omgeving door de input van hun zintuigen en hun bestaande kennis en ervaring om te zetten naar een beeld van de werkelijkheid. Dit beeld van de werkelijkheid gebruiken wij niet alleen om te interacteren met de wereld, maar ook om te bedenken hoe onze acties de wereld kunnen beïnvloeden, welke acties het meest gunstig zijn enzovoorts. We creëren als het ware een 'mentaal model' van de werkelijkheid en plannen en handelen op basis van dit begrip van de werkelijkheid. Een mentaal model is een versimpelde versie van de werkelijkheid die ons in staat stelt om te beredeneren wat er gebeurt in de wereld en hoe ons handelen deze werkelijkheid nu en in de toekomst beïnvloedt. Mentale modellen kunnen op verschillende niveaus worden gevormd. Zo kunnen wij ons een beeld van de wereld vormen waarin we ons bevinden, maar we kunnen ook een model maken van een heel specifiek vraagstuk of probleem.

Wanneer wij een computer een taak willen laten uitvoeren die betrekking heeft op onze werkelijkheid, dan moet de computer ook een begrip hebben van deze werkelijkheid. Om een kunstmatige intelligentie een begrip van de werkelijkheid te geven moet dus een mentaal model worden gecreëerd. Dit kan expliciet geprogrammeerd worden door een mens, of het kan 'geleerd' worden door de computer op basis van trainingsdata en een leeralgoritme. De eerste manier wordt ook wel *symbolic AI* genoemd, de tweede variant wordt aangeduid met de term *machine learning*.

2.3 Symbolic AI

Hoewel de *symbolic AI*-stroming indrukwekkende resultaten heeft geboekt, kent het een in principe onoverkomelijke beperking: de mens moet in staat zijn om de werkelijkheid en alle daarbinnen mogelijke acties voldoende precies te omschrijven, zodat de computer in alle situaties weet wat het moet doen. In dit kader wordt ook wel over *action space* gesproken, dat wil zeggen de hoeveelheid 'opties' die de kunstmatige intelligentie heeft in een concrete situatie om een beslissing te nemen. Zo heeft het spel schaak een kleinere *action space* dan het spel Go waar er op basis van de grootte van het bord en de regels van het spel veel meer mogelijke zetten zijn. Met de 'werkelijkheid' bedoelen wij in deze context de omgeving waarbinnen de kunstmatige intelligentie moet handelen, zoals eerder naar voren kwam in de definitie van kunstmatige intelligentie. Dit kan de gehele fysieke wereld zijn, maar ook een afgebakende omgeving met een afgebakende taak, zoals het spelen van een bordspel of het herkennen van een gezicht.

Het expliciet omschrijven van de werkelijkheid en de regels voor het handelen daarbinnen is voor een afgebakende omgeving en een afgebakende taak goed mogelijk, maar het is nagenoeg onmogelijk wanneer een omgeving complex en/of dynamisch is. In complexe en/of dynamische omgevingen zijn er dusdanig veel variabelen dat het op voorhand niet mogelijk is om alle mogelijke situaties die zich voor gaan doen uit te schrijven. In dergelijke omgevingen is de 'probleemruimte', dat wil zeggen alle elementen waarmee een kunstmatige intelligentie rekening moet houden simpelweg te groot om uit te schrijven.[6]

2.4 Machine learning

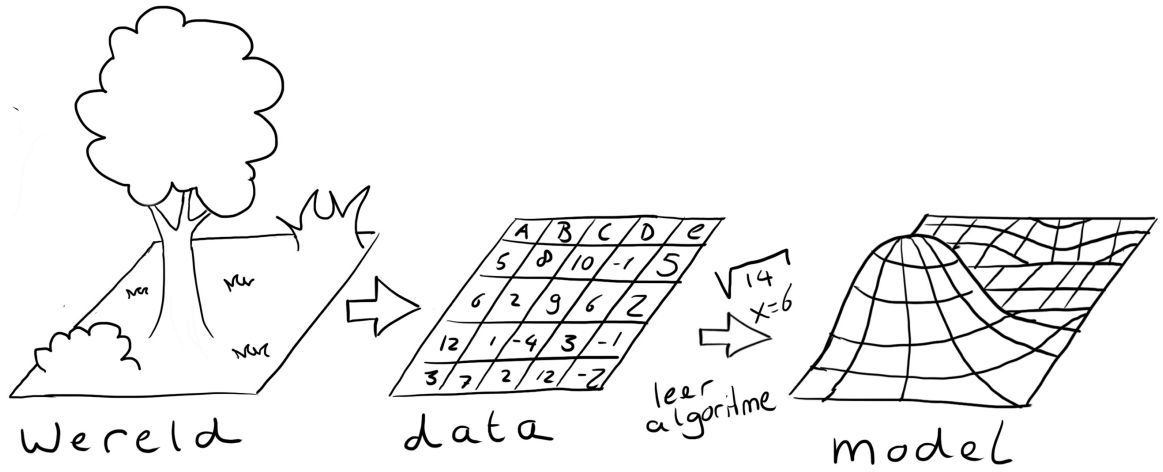
Een deel van het 'antwoord' op de in de vorige paragraaf besproken beperking van de symbolic AI benadering is *machine learning*. Machine learning omzeilt het probleem van het uitputtend moeten beschrijven van de werkelijkheid door mensen, door de computer zelf te laten leren wat het probleem is dat opgelost moet worden en de ruimte waarbinnen dat probleem speelt. Het Joint Research Centre van de Europese Commissie definieert machine learning als:

> *"A subfield of AI to predict a behaviour from many examples given as input. The computer receives input data as well as the answers expected from the data, and the ML agent needs to produce the rules."*[7]

[6] Zie bijvoorbeeld: Pang, Z. et al. (2019) *On Reinforcement Learning for Full-length Game of StarCraft*. Voor een nadere bespreking van wat wij verstaan onder het concept 'probleemruimte' zie paragraaf 2.6

[7] Annoni, A. et al. (2018) . Artificial Intelligence: A European Perspective. JRC Working Papers JRC113826. Joint Research Centre

Door middel van data krijgt de computer informatie over de werkelijkheid. Op basis van deze input en de gebruikte algoritmen vormt de computer een model van deze werkelijkheid dat het kan gebruiken om beslissingen te nemen en/of voorspellingen te doen.



Figuur 1 Vertaling van de werkelijkheid via data naar een model (ontleend aan Molnar 2019)

Het doel van *machine learning* is om te komen tot een functie (f) die optimaal de inputwaarde (X) kan relateren aan de outputwaarde (Y). Dit kan worden voorgesteld als:

$$Y = f(X)$$

Op basis van de beschikbare gegevens (X) en de functie (f) kun je dus de onbekende output (Y) berekenen. Dit stelt je bijvoorbeeld in staat om te voorspellen wat de optimale prijs is voor een tweedehandsauto, of te bepalen of het dier op een plaatje een hond of een kat is.

Om tot en optimale functie (f) te kunnen komen, moet de computer leren begrijpen hoe de inputwaarde(n) en de outputwaarde zich tot elkaar verhouden. Het leerproces van de computer resulteert uiteindelijk in de functie (f). Dit wordt ook wel het model genoemd.[8] Een model kan worden verkregen door trainingsdata door een leeralgoritme te voeren.[9]

2.4.1 Supervised learning, unsupervised learning en self-supervised learning

Binnen machine learning wordt een onderscheid gemaakt tussen *supervised learning*, *unsupervised learning* en *self-supervised learning*.

[8] Voor de leesbaarheid worden in deze rapportage de termen algoritme, model en kunstmatige intelligentie door elkaar gebruikt.

[9] Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Geraadpleegd via: https://christophm.github.io/interpretable-ml-book/

Bij *supervised learning* begeleidt een mens het leerproces van het algoritme. Bij *supervised learning* is in de trainingsfase zowel de input (X) als de output (Y) bekend en is het de taak van de computer om tot een zo goed mogelijke functie (f) te komen, die de input relateert aan de output. Dit proces kunnen we verduidelijken aan de hand van een voorbeeld. Stel, je wilt een algoritme leren om onderscheid te maken tussen honden en katten. De trainingsdata bestaat uit een grote hoeveelheid plaatjes van honden en katten die correct gelabeld zijn (het is duidelijk wat een hond en wat een kat is). Het leeralgoritme wordt op basis van deze kennis getraind en kan vervolgens met een bepaalde accuraatheid toekomstige, niet gelabelde data herkennen.

Bij *unsupervised learning* is de trainingsdata niet gelabeld en is er geen vooraf gedefinieerde output. Het idee is dat het algoritme zelf in de data relevante verbanden leert herkennen. Het doet dit door bijvoorbeeld clusters van gerelateerde datapunten te identificeren.

Een relatief nieuwe ontwikkeling is *self-supervised learning*. Self-supervised learning lijkt het meest op de manier waarop baby's en peuters leren: door te observeren en te interacteren met hun omgeving, zonder dat iemand hen uitlegt hoe iets werkt.[10] Vertaald naar een machine learning context is *self-supervised learning* de mogelijkheid om op basis van verbanden/structuren in de data zelf labels voor deze data te creëren die vervolgens gebruikt kunnen worden voor toekomstige voorspellingen. In die zin komt een *self-supervised learning* algoritme het dichtst in de buurt van een 'zelflerend algoritme'.

2.4.2 Machine learning varianten

Machine learning is een breed veld en computers kunnen op verschillende manieren leren. Momenteel worden de grootste stappen gezet in het oplossen van problemen door gebruik te maken van *deep learning*, *reinforcement learning* en *imitation learning*. Het gaat hierbij overigens niet persé om discrete categorieën en deze verschijningsvormen zijn ook niet wederzijds exclusief. Zo kunnen deep learning algoritmen worden toegepast voor reinforcement learning (*deep reinforcement learning*) en kunnen verschillende methoden naast elkaar gebruikt worden (bijvoorbeeld eerst imitation learning toepassen en daarna reinforcement learning) om tot optimale resultaten te komen.

[10] Wiggers, K., Y. LeCun and Y. Bengio (2020). *Self-supervised learning is the key to human-level intelligence*. Geraadpleegd via: https://venturebeat.com/2020/05/02/yann-lecun-and-yoshua-bengio-self-supervised-learning-is-the-key-to-human-level-intelligence/.

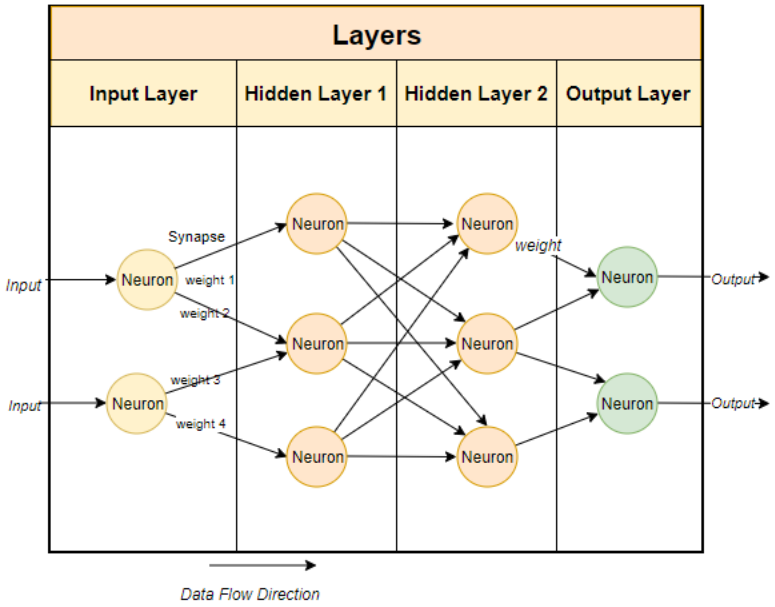
Deep learning

Deep learning maakt gebruik van neurale netwerken om problemen op te lossen. Neurale netwerken zijn netwerken die losjes gebaseerd zijn op de werking van het menselijke brein. Het menselijk brein is opgebouwd uit neuronen: cellen die informatie van de ene aan de andere cel doorgeven. Afhankelijk van de situatie 'vuurt' een neuron (het geeft wel of geen signaal door). Clusters van neuronen die gezamenlijk vuren vormen gedachten. Door het complexe samenspel van neuronen kunnen wij denken. Neurale netwerken imiteren dit proces.[11]

Neurale netwerken zijn opgebouwd uit lagen die bestaan uit 'neuronen'. Het neurale netwerk heeft minimaal twee lagen: een *inputlaag* en een *outputlaag*, maar doorgaans één of meer tussenlagen. Als het netwerk meerdere tussenliggende lagen heeft dan spreken we van een diep neuraal netwerk, vandaar de term deep learning.

De lagen in het netwerk vormen een hiërarchisch geheel. Informatie wordt doorgegeven van laag naar laag, beginnend met de inputlaag en eindigend met de outputlaag (de oplossing). Alle neuronen in een laag zijn verbonden met alle neuronen in de volgende laag en zo verder. Elke neuron berekent een waarde op basis van de input die komt uit de voorgaande laag. Deze input is de waarde van elke voorgaande neuron vermenigvuldigd met een bepaald gewicht (*weight*). De neuron berekent de gewogen som van al deze waarden. Er wordt aan de som van alle inputs een *bias* toegevoegd en vervolgens wordt de waarde door een activatie functie gevoerd om een getal te krijgen tussen 0 en 1. Dit is de nieuwe 'waarde' van deze neuron. Deze waarde wordt vervolgens doorgegeven als input aan de neuronen in de volgende laag, waar het proces zich herhaalt.

[11] Aggarwal, C. C. (2018), *Neural Networks and deep learning*. Springer International Publishing AG.



Figuur 2 Weergave van een neuraal netwerk[12].

Het idee is dat de latere lagen concepten met een hogere abstractie kunnen afleiden uit de informatie die komt uit de voorgaande lagen. De werking van een diep neuraal netwerk dat plaatjes van honden moet herkennen zou bijvoorbeeld als volgt kunnen werken: De neuronen in de inputlaag representeren de pixels in het plaatje en hebben allemaal een waarde. Alle neuronen in de inputlaag 'vuren' naar de neuronen in de volgende laag die daar patronen zoals rondingen of rechte lijnen proberen te herkennen. De volgende laag herkent in de rondingen en lijnen vormen zoals poten of ogen. De laatste laag neemt deze vormen bij elkaar en beoordeelt of de patronen samen een hond vormen of niet. In de outputlaag komt tenslotte de finale output, zijnde de oplossing voor het probleem: is dit een afbeelding van een hond of niet.

Het grote voordeel van deep learning is dat het zelf relevante kenmerken (*features*) kan ontdekken in de input data. Het is niet nodig voor een mens om aan te geven welke features relevant zijn of welk patroon van belang is voor het herkennen van een hond, dit zoekt het neurale netwerk zelf uit door middel van een proces van *trial and error*. In eerste instantie zullen de resultaten van het neurale netwerk veel fouten bevatten, maar door de gewichten aan te passen in de connecties tussen de neuronen komt het netwerk uiteindelijk vanzelf tot een

[12] Malik, Farhad.,' Neural Network Layers. Understanding How Neural Network Layers Work', 18-05-2019, https://medium.com/fintechexplained/neural-network-layers-75e48d71f392.

passend(er) model voor het probleem (de input kan goed gerelateerd worden aan de output) en leert het de features herkennen die relevant zijn.

Reinforcement learning

Reinforcement learning is een manier van leren waarbij een algoritme zelf moet proberen een bepaald doel te bereiken zonder expliciete instructies (wederom door middel van *trial and error*). Een algoritme moet bijvoorbeeld een bepaald spel winnen zonder dat het de regels van het spel kent. Om te zorgen dat het algoritme leert hoe het doel bereikt moet worden, wordt het 'beloond' voor acties die bijdragen aan het oplossen van het probleem en 'gestraft' voor handelingen die dat niet doen (de zogenaamde *reward function*).[13] Het algoritme 'onthoudt' de acties die bijdragen aan het oplossen van het probleem en wordt zo steeds beter in het uitvoeren van de gewenste taak. Omdat het algoritme met voldoende rekenkracht eindeloos opties kan uitproberen, 'leert' het uiteindelijk hoe het spel gespeeld moet worden en kan het dit vaak ook nog op een bovenmenselijk niveau.

Imitation learning

Daar waar de probleemruimte (zie paragraaf 2.6) dusdanig groot is dat het moeilijk is om te bepalen wanneer een bepaalde actie beloond of bestraft moet worden (de reward function kan niet berekend worden), is het niet goed mogelijk om reinforcement learning toe te passen. In dergelijke gevallen kan *imitation learning* worden toegepast. Imitation learning is een specifieke vorm van supervised learning waarbij een algoritme leert van de wijze waarop een expert (een mens) de taak uitvoert. Een goed voorbeeld is AlphaStar. AlphaStar leerde het computerspel Starcraft te spelen door partijen van menselijke spelers te analyseren.[14] Op basis daarvan kon het een goede Starcraft speler worden, door vervolgens ook reinforcement learning toe te passen kon AlphaStar uiteindelijk zo goed worden dat het internationale topspelers kon verslaan.

[13] Zie bijvoorbeeld: Samuditha, G. (2019), What is reinforcement learning, via: https://towardsdatascience.com/what-is-reinforcement-learning-b047d9bb05cc

[14] The Alphastar Team (2019). AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. Deepmind. Geraadpleegd via: https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii.

2.5 Leren in *real-time*?

Alle vormen van machine learning kennen een 'leerfase' en een 'handelingsfase'. Het model wordt getraind op basis van de beschikbare trainingsdata en wordt daarna in productie genomen. Hoewel het model nieuwe, onbekende input data krijgt, verandert het zelf niet meer: het geeft een output op basis van de input, maar het gebruikt de input data niet om zichzelf te veranderen. Met andere woorden, machine learning modellen veranderen niet met de tijd en worden ook niet 'slimmer' of 'beter'. Sterker nog, ze worden doorgaans 'slechter', omdat de wereld om hen heen verandert.

Omdat modellen niet in *real time* leren, neemt de voorspellende waarde ervan af met de tijd (*model degradation*). De oorzaak hiervoor ligt in het feit dat modellen doorgaans statisch zijn, dat wil zeggen: ze worden getraind op basis van historische data en veranderen niet op basis van de input data die zij vervolgens gaan verwerken. Met andere woorden: wanneer de werkelijkheid verandert, dan verandert het model niet mee. Dit fenomeen wordt ook wel aangeduid met de term *concept drift*.[15]

Wil het model gebruik kunnen maken van nieuwe data of veranderende inzichten, dan moet het model opnieuw worden getraind. Dit betekent dat het model uit productie wordt gehaald, opnieuw wordt getraind en daarna weer in productie wordt genomen. Het besluit om een model te hertrainen is aan de degene die het model inzet.

Het continu opnieuw trainen van modellen wordt *continual learning* genoemd. Bij *continual learning* wordt de input data die het model gebruikt voor de voorspellingen ook bewaard in een aparte database, om later als trainingsdata te kunnen functioneren. Hoewel de interval waarmee een model wordt hertraind steeds korter wordt, is het nog niet vergelijkbaar met de manier waarop een mens kan leren van een ervaring en deze direct in een nieuwe situatie (en zelfs een andere context) kan toepassen.

2.6 Probleemruimte en de situering van kunstmatige intelligentie

Zoals beschreven in de definities van intelligentie wordt intelligentie afgemeten aan de capaciteit om binnen een bepaalde omgeving optimaal een doel te bereiken. Om het doel te kunnen bereiken is een goed begrip van de omgeving noodzakelijk en hoe jouw handelen als intelligentie deze omgeving beïnvloedt.

[15] Wat ook mogelijk is, is dat ons denken over de wereld verandert in de tijd, bijvoorbeeld door nieuwe wetenschappelijke inzichten of veranderende maatschappelijke opvattingen over een bepaald fenomeen. Dat betekent dat we dezelfde data op een andere manier interpreteren of wegen. Daar waar we vroeger label 'A' aan iets zouden hangen, hangen we daar nu label 'B' aan. Een model dat getraind is op data met de oorspronkelijke labels zal verkeerde voorspellingen gaan doen (A is nog steeds A voor het model, niet B).

Het vormen van een beeld van de omgeving waarbinnen of waarmee het model moet werken, kan afhankelijk van de situatie bijzonder ingewikkeld zijn. Voor gezichtsherkenning bijvoorbeeld, zijn alleen beelden van gezichten relevant. De data die gebruikt worden voor het trainen van een model om gezichten te herkennen bestaan uit foto's of video's van mensen (genomen uit verschillende hoeken, met uiteenlopende belichting *et cetera*). Voor het 'probleem' gezichtsherkenning is de probleemruimte dus redelijk goed afgebakend.

Wanneer de probleemruimte groter is, omdat er heel veel variabelen zijn die tegelijkertijd een beslissing kunnen beïnvloeden, dan zijn meer gegevens nodig om de probleemruimte te definiëren. Wanneer er te veel mogelijkheden en/of variabelen zijn, dan is het niet meer mogelijk om de probleemruimte te simuleren of daarvoor een representatieve dataset te creëren. In dergelijke gevallen is bijvoorbeeld supervised learning niet meer mogelijk (omdat de mens niet meer alle datapunten kan labelen). Voor dit soort complexe omgevingen worden andere methoden zoals imitation learning en reinforcement learning gebruikt.[16]

Een kunstmatige intelligentie die daadwerkelijk in onze fysieke wereld actief is (een *gesitueerde intelligentie*), moet zich een beeld van de fysieke werkelijkheid vormen (in ieder geval van die elementen die relevant zijn voor de goede uitvoering van de taken). Los van het goed kunnen uitvoeren van de taak zijn er allerlei randvoorwaarden waarmee de kunstmatige intelligentie rekening moet houden, bijvoorbeeld de veiligheid van mensen en objecten in de fysieke wereld. Zo is het, bijvoorbeeld, voor een robotstofzuiger niet zo erg om tegen een stoelpoot te botsen en op basis daarvan zijn koers aan te passen, maar als een autonome auto deze strategie hanteert dan ontstaan uiteraard gevaarlijke situaties. De autonome auto moet daarom met meer factoren rekening houden dan de robotstofzuiger en een andere oplossing voor het probleem zoeken.

Naarmate de dynamiek of de complexiteit van de omgeving waarin de kunstmatige intelligentie gesitueerd is (of dit nu de fysieke wereld is of een digitale omgeving) groter is, wordt de 'probleemruimte' voor de kunstmatige intelligentie dus ook groter. Om effectief met de omgeving om te gaan is dan doorgaans een hogere intelligentie vereist.

Definitie probleemruimte

Wij definiëren 'probleemruimte' voor de context van deze rapportage daarom als volgt:

> *"De omgeving waarbinnen een kunstmatige intelligentie ingezet wordt en waarbinnen het specifieke doelstellingen moet bereiken."*

[16] Lőrincz, Z. (2019). A brief overview of Imitation Learning. Geraadpleegd via: https://medium.com/@SmartLabAI/a-brief-overview-of-imitation-learning-8a8a75c44a9c.

Een voorbeeld van een probleemruimte is het spel Schaak. De probleemruimte van het spel Schaak is alle mogelijke posities op het bord voor de kunstmatige intelligentie en diens tegenstander en alle legale zetten om deze posities te bereiken. Binnen deze probleemruimte is het doel van de kunstmatige intelligentie (de *eindstaat*) het verslaan van de tegenstander. Meer specifiek: de kunstmatige intelligentie moet de stukken dusdanig positioneren dat de tegenstander geen legale zetten meer kan doen met zijn koning.

Voor een zelfrijdende auto is de openbare weg en alle situaties die zich daar voor kunnen doen de probleemruimte. Binnen deze probleemruimte is het snel en veilig van punt A naar punt B komen de doelstelling.

Maar de probleemruimte kan ook zijn: Het financiële gedrag betreffende een persoon. Binnen deze probleemruimte is het doel bijvoorbeeld het zoeken naar afwijkende patronen die kunnen duiden op fraude. Voor deze probleemruimte ligt de moeilijkheid in de vraag of de probleemruimte effectief 'gevangen' is door de data: zijn bijvoorbeeld alle variabelen die het gedrag vormen van een persoon meegenomen en zijn alle variabelen die frauduleus gedrag bekend en vastgelegd?

Daar waar voor spellen zoals Schaak en Go de probleemruimte duidelijk afgebakend is, is dit voor de fysieke wereld en menselijk gedrag veel minder helder. Het is dus van groot belang dat er een goed begrip is van de probleemruimte en hoe keuzes tijdens de hele levenscyclus van een algoritme deze beïnvloedden.

2.7 Emergent gedrag

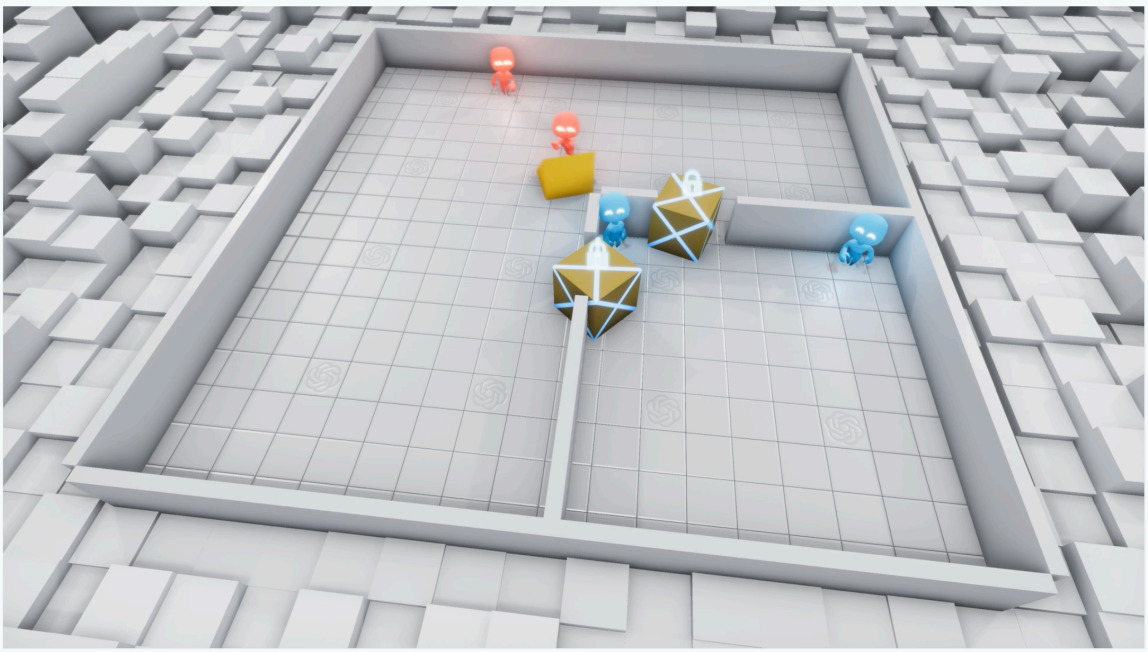
Zoals hierboven beschreven is intelligentie nauw verbonden met de omgeving waarin een intelligentie zich bevindt. Maar naast de interactie met de omgeving is ook de interactie met anderen relevant voor het ontwikkelen van intelligentie. Op onze planeet is complexe intelligentie ontstaan door de co-evolutie en strijd tussen verschillende organismen.[17] Wanneer een organisme een nieuwe strategie verzint die beter werkt dan dat van een organisme waarmee het concurreert, dan moet de concurrent zijn gedrag aanpassen. Deze interactie leidt tot nieuwe strategieën en daarmee tot intelligent(er) gedrag.

Dit concept vertaalt zich ook naar kunstmatige intelligentie. Wanneer kunstmatige intelligentie (een *agent)* in een omgeving komt waarin het gedwongen wordt om zich aan te aan te passen door het handelen van andere agenten (een *multi-agent systeem*), dan kan 'spontaan'

[17] Dawkins, R., & Krebs, J. R. (1979). *Arms races between and within species. Proceedings of the Royal Society of London*. Series B. Biological Sciences, 205(1161), 489-511; Baker, B. et al. (2019). *Emergent tool use from multi-agent autocurricula*. arXiv preprint arXiv:1909.07528.

intelligent gedrag ontstaan. De agent ontwikkelt nieuwe strategieën op basis van de veranderde omstandigheden die niét van tevoren gemodelleerd zijn door de mens.[18] Dergelijk *emergent gedrag* ontstaat uit de interacties tussen verschillende agenten en is niet geprogrammeerd. Dit betekent ook dat emergent gedrag veelal niet voorspelbaar of verklaarbaar is.

Een zeer treffend voorbeeld van emergent intelligent gedrag in de context van kunstmatige intelligentie is het programma *Hide and Seek* (verstoppertje) van OpenAI.[19] In een virtuele omgeving spelen twee teams van intelligente agenten verstoppertje. De agenten hebben een (relatief) beperkte set aan mogelijkheden: ze kunnen elkaar zien, ze kunnen bewegen en ze kunnen objecten verplaatsen en vergrendelen. Zowel de verstoppers als de zoekers komen tot complexe strategieën om het spel te winnen. Zo blokkeren de verstoppers deuren met blokken, zodat de verstoppers niet in een ruimte kunnen komen. Om toch in de ruimte te komen leren de verstoppers om een helling te gebruiken waarmee ze over de muur heen klimmen.



Episodes 8.62–14.5 million
Ramp Use Seekers learn to use the ramp to jump over obstacles.

Figuur 3 Hide and Seek van OpenAI

[18] in tegenstelling tot bijvoorbeeld de aanpak die wordt gehanteerd bij *symbolic AI*.

[19] Baker, B. et al. (2019). *Emergent tool use from multi-agent autocurricula*. arXiv preprint arXiv:1909.07528.

Geen van deze strategieën zijn geprogrammeerd of gepland: het ontstaat allemaal uit de spontane interactie tussen de agenten onderling en de omgeving waarin zij zich bevinden.

Een ander voorbeeld is AlphaStar, de kunstmatige intelligentie die de allerbeste Starcraft spelers ter wereld heeft weten te verslaan. Om het niveau van de menselijke wereldtop te halen werd een virtuele competitie opgesteld waarin verschillende versies van AlphaStar tegen elkaar speelden. Uiteindelijk werd de best presterende agent ingezet tegen de menselijke spelers. De geselecteerde agent had op basis van de trainingsperiode in de competitie het equivalent van 200 jaar (!) aan menselijke Starcraft ervaring.[20]

2.8 Begrijpelijkheid en transparantie

Een vraagstuk dat binnen machine learning speelt is de begrijpelijkheid van de gebruikte modellen. Dit wordt ook wel aangeduid als het '*black box* probleem'. Een *black box* is een apparaat waarvan de werking voor de gebruiker verborgen is. Je kan als het ware niet 'in de doos kijken'. Het tegenovergestelde van een *black box* is een *white box* (ook wel een *glass box* genaamd). Een *white box* is een apparaat waarvan de interne werking volledig transparant is voor de gebruiker.[21]

Een machine learning model kan een black box zijn, omdat de maker of gebruiker ervan het de werking van het model niet openbaar wenst te maken. De reden hiervoor kan bijvoorbeeld de bescherming van intellectueel eigendom zijn of angst dat door kennis van het model mensen in staat zijn de resultaten te manipuleren (*gaming the system*).

Een machine learning model kan ook een black box zijn omdat het model simpelweg te complex is voor mensen om te bevatten. Zelfs al zou het model volledig transparant en openbaar zijn, dan nog kunnen mensen de volledige werking van het model niet begrijpen, omdat het door de complexiteit niet voor mensen meer is te bevatten.[22] Feitelijk gezien is een dergelijk model dus een white box model, maar wel een onbegrijpelijk white box model. Voor het gemak zullen wij in deze rapportage voor beide oorzaken de term *black box* hanteren.

[20] The Alphastar Team (2019). AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. Deepmind. Geraadpleegd via: https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii.

[21] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160.

[22] Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Geraadpleegd via: https://christophm.github.io/interpretable-ml-book/

2.9 Beperkingen van machine learning

Hoewel machine learning indrukwekkende resultaten heeft geboekt in diverse domeinen zijn er een aantal beperkingen die in relevant zijn voor de context van deze rapportage, omdat zij naar het idee van de onderzoekers kunnen bijdragen aan onvoorziene effecten.

Getraind voor specifieke taken[23]

Een machine learning model wordt getraind voor een zeer specifieke taak en kan niet zomaar voor een andere taak worden ingezet. Zo kan een model dat getraind is in het herkennen van honden niet worden gebruikt om te schaken: het moet worden hertraind voor deze nieuwe taak. Dit betekent dat het toepassingsgebied van *machine learning* modellen relatief afgebakend is en niet persé dynamisch kan worden omgegaan met onvoorziene omstandigheden.

Vereist grote hoeveelheden gegevens

Voor het effectief trainen van *machine learning* modellen zijn grote hoeveelheden gegevens noodzakelijk. Omdat deze data op basis van menselijke keuzes gegenereerd en geselecteerd worden, is het in veel gevallen de vraag of de gegevens die gebruikt worden adequaat zijn voor het beschrijven en oplossen van het hele probleem. Met andere woorden: is de data waarop de computer zich een beeld van de wereld vormt daadwerkelijk representatief voor deze werkelijkheid?

Correlatie in plaats van causaliteit

Machine learning modellen (en statistische modellen in het algemeen) kunnen correlaties tussen verschillende attributen (*features*) in de data herkennen, maar dat betekent niet direct dat er sprake is van een causaal verband.

Ondoorgrondelijkheid van de besluitvorming

Complexe machine learning modellen zijn doorgaans ondoorgrondelijk (het black box probleem). Dit maakt het moeilijk om te beoordelen of wat zij doen daadwerkelijk correct is. In domeinen waar de juiste uitkomst van een probleem niet eenduidig vaststaat is daardoor niet altijd met voldoende zekerheid vast te stellen of een uitkomst correct is, of hoe het model gaat reageren op een onvoorziene situatie.

[23] Voss, P. (2016). Why Machine Learning won't cut it'. Geraadpleegd via: https://medium.com/@petervoss/why-machine-learning-wont-cut-it-f523dd2b20e3#.wifeugkuq.

Het spel schaak bijvoorbeeld heeft een duidelijk einddoel: de koning van de tegenpartij kan geen geldige zetten meer doen. Als de kunstmatige intelligentie deze eindstaat weet te bewerkstelligen, dan heeft het goed zijn taak uitgevoerd. Hoe het dat heeft gedaan doet in feite niet ter zake: het heeft de taak goed uitgevoerd.[24] Dit is een indicatie dat de besluitvorming 'correct' of 'accuraat' is geweest. We hoeven de werking van het model niet persé te begrijpen om wat te zeggen over de bruikbaarheid van de uitkomsten.

Voor veel probabilistische vraagstukken ligt dit anders, omdat de eindstaat (nog) niet bekend is op het moment van het nemen van het besluit. Je kan op het moment van de beslissing dus nog niet zeggen of het model goed werkt, omdat de eindstaat nog niet is bereikt en je dus niet objectief kunt vaststellen of het 'goed' is gegaan. Dit is inherent aan probabilistische vraagstukken en is op zich ook niet uniek voor machine learning: ook mensen moeten vaak op basis van incomplete informatie een besluit over de toekomst nemen. Maar in tegenstelling tot machine learning modellen kunnen mensen motiveren hoe zij tot een bepaald besluit zijn gekomen.

Wil je weten hoe een machine learning model tot een besluit is gekomen, dan moet je 'onder de motorkap kijken' en verifiëren of het model correct werkt (kiest het de juiste *features*, wordt alles meegewogen *et cetera*). Maar juist dat is door de complexiteit van het model niet (goed) mogelijk.[25]

Uiteraard kun je in de trainingsfase testen en valideren of het model goed werkt, maar dit is altijd binnen de context van de geselecteerde dataset. Mocht deze dataset niet een volledig of objectief beeld van de werkelijkheid geven, bijvoorbeeld omdat de data vooringenomenheid bevat (*bias*) dan is dat moeilijk te herkennen op basis van de gebruikte gegevens of de verkregen uitkomsten.

2.10 Tussenconclusie

In dit hoofdstuk hebben wij verschillende verschijningsvormen van kunstmatige intelligentie beschreven. Momenteel worden de grootste vooruitgangen op het gebied van kunstmatige intelligentie geboekt in het veld van machine learning. Binnen deze discipline worden modellen getraind op basis van grote hoeveelheden gegevens.

[24] Er uiteraard vanuit gaande dat het zelf geen illegale zetten heeft gedaan.

[25] Nu zijn er wel mogelijkheden om *post hoc* voor een specifieke beslissing te achterhalen wat een model heeft gedaan (zie hoofdstuk 11), maar dit geeft nog geen inzicht in de werking van het model als geheel.

Hoe intelligent een kunstmatige intelligentie kan en moet zijn hangt af van de probleemruimte: naarmate er meer variabelen zijn waarmee rekening moet worden gehouden, wordt meer gevergd van de kunstmatige intelligentie.

3 Kunstmatige intelligentie in de praktijk

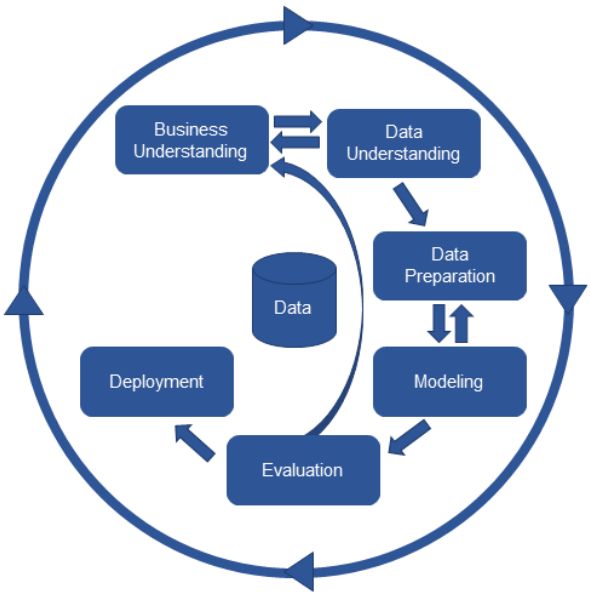
In het voorgaande hoofdstuk hebben we de verschillende verschijningsvormen van kunstmatige intelligentie besproken en gekeken hoe algoritmen (zelf) leren. In dit hoofdstuk richten wij ons op de manier waarop algoritmen en modellen in de praktijk worden ontwikkeld en ingezet.

De ontwikkeling en toepassing van algoritmen en modellen die in staat zijn om inzicht te krijgen in data duiden wij aan met de term *data science*.

3.1 Het proces van data-analyse

Om beter te begrijpen hoe lerende modellen tot stand komen behandelen we in deze paragraaf het *Cross Industry Standard Process for Data Mining* model, oftewel het CRISP-DM model. Dit is een open standaard procesmodel dat wordt gebruikt voor de ontwikkeling van algoritmische modellen.

In de onderstaande afbeelding is het CRISP-DM model weergegeven. De CRISP-DM methodologie is verdeeld in verschillende niveaus, fases en taken. Ieder niveau heeft een bepaald niveau van abstractie. Dit betekent dat wanneer een niveau 'dieper' wordt gegaan, de daarbij horende fases en taken specifieker en gedetailleerder zijn dan in het niveau daarboven.[26]



Figuur 4 Het CRISP-DM model

[26] Wirth, 'CRISP-DM: Towards a Standard Process Model for Data Mining', p.3.

Een voorbeeld dat door Wirth en Hipp wordt genoemd is dat in een hoger niveau een taak voorkomt die kan worden samengevat als 'build model'. Op een niveau dieper is dezelfde taak aanwezig, maar dan beschreven als 'build response model', waarbij er meer voorwaarden aan hetzelfde model worden gesteld.[27]

De CRISP-DM methode heeft de ontwikkeling van een goed werkend en stabiel model tot doel. Hiertoe dienen alle stappen in het proces te worden doorlopen, want het overslaan van een stap leidt er mogelijk toe dat het model problemen kent bij de implementatie. De beschrijving van de fases, niveaus en taken is hoe de ontwikkeling van een algoritmisch model idealiter verloopt via de CRISP-DM methode.

Hoewel alle stappen in de methode doorlopen dienen te worden, betekent het niet dat al deze stappen altijd in dezelfde volgorde worden doorlopen tijdens de ontwikkeling. Het CRISP-DM model is geen rigide model, waarbij telkens een stap vooruit wordt gedaan. Iedere stap in het proces staat in relatie tot de vorige en de volgende stap in het proces. Wanneer een probleem wordt ontdekt, dan moet de ontwikkeling van het model misschien een paar stappen terug doen. Bijvoorbeeld, wanneer bij de 'data preparation' de data-analisten erachter komen dat de gebruikte trainingsdata niet geschikt is voor het doel van het algoritmisch model, dan moet een stap terug worden gedaan naar 'data understanding' of misschien wel 'business understanding' om zodoende tot wel geschikte uitgangspunten te komen voor de ontwikkeling van het model en de bijbehorende trainingsdata.[28]

De cirkel rondom het model geeft aan dat de CRISP-DM methode cyclisch van aard is. Het proces stopt namelijk niet na implementatie (*deployment*) van het model. Na de implementatie van het model worden nieuwe ideeën opgedaan en zijn lessen geleerd die gebruikt kunnen worden voor de ontwikkeling van een beter model. Zo wordt het proces van algoritmische modelontwikkeling telkens nieuw leven ingeblazen.

3.1.1 Business understanding

De eerste fase, *business* understanding, bestaat uit verschillende sub-fases. Ten eerste dienen de *business objectives* van de organisatie duidelijk te zijn. Het moet duidelijk zijn voor het management van een organisatie wat het doel is van het project en wat de vereisten zijn voor het project. Vervolgens moet worden onderzocht wat de context is waarbinnen het algoritmisch model ontwikkeld en ingezet wordt. Relevante factoren om in ogenschouw te nemen zijn onder

[27] Wirth, 'CRISP-DM: Towards a Standard Process Model for Data Mining', p.3.

[28] Chapman, P. et al. (2000). CRISP-DM 1.0. CRISP-DM Consortium, 76(3), p. 10.

andere: de beschikbare middelen (*resources*), de beperkingen van het project en de aannames die men heeft betreffende het project. Vervolgens dient al deze informatie omgezet te worden in een projectplan waaruit blijkt dat het doel en de vereisten bereikt kunnen worden door middel van het ontwikkelen van een model.[29]

De fase van *business understanding* is in het kader van dit rapport in het bijzonder van belang omdat in deze de fase de probleemruimte wordt gekozen/beschreven.

3.1.2 Data understanding

In het projectplan zijn de middelen opgenomen die beschikbaar zijn voor het project, inclusief de beschikbare datasets. Dit kan bijvoorbeeld data zijn die de organisatie reeds heeft verzameld, of wenst te verzamelen.

De tweede fase *data understanding* begint met het bijeenbrengen van deze datasets. Vervolgens moet de data, zoals de naam van de fase aangeeft, worden begrepen: de data wordt beschreven, gelabeld, de datakwaliteit wordt gecontroleerd en eerste simpele analyses worden uitgevoerd. Bij iedere stap wordt gerapporteerd of er problemen zijn, of deze zijn opgelost, hoe deze zijn opgelost en of problemen zich blijven voordoen.

Wanneer problemen zich blijven voordoen, ondanks pogingen van het projectteam om deze problemen weg te nemen, dan zal weer een stap terug moeten worden gedaan naar '*business understanding*'. In deze stap kunnen bijvoorbeeld veranderingen worden aangebracht aan het doel van het project en de middelen die ter beschikking staan. Er is een nauwe relatie tussen '*business understanding*' en '*data understanding*', want het formuleren van het probleem dat het algoritmisch model moet oplossen in de fase van *'business understanding'* wordt direct beïnvloed door het begrijpen van de beschikbare data in de fase *'data understanding'*.[30]

3.1.3 Data preparation

In de derde fase worden de data geselecteerd die gebruikt gaan worden ter ontwikkeling van het algoritmisch model. Het selecteren van de juiste data betekent onder meer dat data worden opgeschoond om de datakwaliteit te garanderen, data met elkaar gecombineerd worden om nieuwe informatie te genereren en de data worden omgezet naar bestandsformaten die bruikbaar zijn bij de ontwikkeling van het model.[31]

[29] Chapman, P. et al. (2000). CRISP-DM 1.0. CRISP-DM Consortium, 76(3), p. 17-19.

[30] Wirth, Hipp, CRISP-DM: Towards a Standard Process Model for Data Mining, p. 5

[31] Wirth, Hipp, CRISP-DM: Towards a Standard Process Model for Data Mining, p. 5

3.1.4 Modeling

In de vierde fase wordt het algoritmisch model ontwikkeld. Echter, voordat het model daadwerkelijk wordt gebouwd, moet eerst worden besloten wat voor soort model gecreëerd wordt. Na deze keuze wordt eerst een testontwerp gemaakt van de gekozen variant om de kwaliteit van het model te testen. Wanneer het testontwerp niet door de test komt, betekent dit dat er mogelijk iets mis is met de data die zijn voorbereid in de vorige fase. Als dat het geval is, dan gaat de ontwikkeling van het algoritmisch model terug naar de '*data preparation*' of zelfs terug naar '*data understanding*' fase.[32]

Wanneer het testontwerp naar tevredenheid functioneert, wordt het model gebouwd. Ook dit gebouwde model wordt gecontroleerd en op kwaliteit beoordeeld. Deze beoordeling wordt alleen gedaan op het technische aspect van het model, nog niet op de elementen zoals deze zijn opgesteld in het projectplan ten tijde van de '*business understanding*' fase. Dat gebeurt in de volgende fase.[33]

3.1.5 Evaluation

De vijfde fase in het CRISP-DM proces wordt gebruikt om het gebouwde model te evalueren, om te beoordelen of het model voldoet aan de scope, de doelen en de vereisten zoals vastgesteld in de fase van '*business understanding*'. Tijdens het evaluatieproces is ook sprake van een *review proces*, waar geanalyseerd wordt of belangrijke aspecten tijdens de ontwikkeling van het algoritmisch model in de vorige fases over het hoofd zijn gezien.[34]

3.1.6 Deployment

De laatste stap is de ingebruikname van het gebouwde model en het daarbij behorende beheer. Zo moet bijvoorbeeld de monitoring en het onderhoud van het model worden gedocumenteerd. Verder moeten instructies worden opgesteld voor de personen die het model in gebruik nemen, zodat zij het model goed gebruiken.[35] Deze gebruikers moeten onder meer weten wat de reikwijdte van het model is, voor welk doel het model is ontwikkeld en wat de beperkingen van het model zijn. Wanneer deze stap is afgerond wordt een eindrapport

[32] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39).

[33] Chapman, P. et al. (2000). CRISP-DM 1.0. CRISP-DM Consortium, 76(3)., p. 23-25.

[34] Ibid., p. 26-27.

[35] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39).

opgesteld en het proces geëvalueerd. Een dergelijk rapport zorgt ervoor dat alle lessen die geleerd zijn tijdens het project meegenomen worden naar de volgende cyclus.[36]

3.2 Data science binnen de context van een organisatie.

Het is van belang om ons te beseffen dat *data science* geen op zichzelf staande activiteit of discipline is. Doorgaans staan voorspelmodellen en andere kunstmatige intelligentie toepassingen in dienst van specifieke bedrijfsprocessen en strategieën. Hierbij gaat het bovenal om het beter begrijpen van processen (*business intelligence*) om deze processen vervolgens effectiever en efficiënter te maken. Hoewel ook wel modellen worden gebruikt om tot totaal nieuwe inzichten en toepassingen te komen, lijkt de nadruk toch te liggen op de eerste categorie toepassingen.

De ontwikkeling en toepassing van algoritmen vinden ook niet plaats in isolatie. De hierboven beschreven CRISP DM methodologie wordt – als het goed is – binnen de organisatie ingebed in de algemene processen voor het ontwikkelen van producten en diensten. Het gaat dan niet alleen om het opstellen van de *business case* en het bijbehorende projectmanagement, maar ook om de *governance* (risicomanagement, verandermanagement, compliance *et cetera*) binnen de organisatie.

Met het oog op mogelijke onvoorziene effecten van (zelflerende) algoritmen is *model risk management* in het bijzonder relevant. *Model risk management* is het proces van het beheersen van risico's bij de ontwikkeling en toepassing van modellen. Hierbij moet primair gedacht worden aan de ontwikkeling van modellen, het opstellen van beleid en procedures voor het beheersen van risico's, het goed monitoren en beheren van modellen gedurende hun levenscyclus en het (extern) valideren van de gebruikte modellen.[37] In de financiële wereld is *model risk management* een verplicht onderdeel van het risico management van financiële instellingen.

3.3 Socio-technische dimensie van kunstmatige intelligentie en relevante actoren

Algoritmen bestaan niet op zichzelf maar vormen een onderdeel van bredere systemen en ecosystemen waarin zijn opereren. Een kunstmatige intelligentie is in die zin een 'socio-technisch systeem'. Socio-technische systemen hebben mechanische (hardware),

[36] Chapman, P. et al. (2000). CRISP-DM 1.0. CRISP-DM Consortium, 76(3)., p. 28-29.

[37] Zie bijvoorbeeld: Bank of England Prudential Regulation Authority (2017), *Stress Test Model Management.*

informationele (software), psychologische (mensen) en sociale componenten die interacteren en kunnen veranderen met de tijd.[38]

Bij de toepassing van kunstmatige intelligentie zijn diverse partijen betrokken. In deze rapportage onderscheiden wij de volgende actoren/rollen:1) ontwikkelaar, 2) gebruiker 3) eindgebruiker, en 4) subject.

Ontwikkelaar

De ontwikkelaar is de natuurlijke- of rechtspersoon die het algoritme ontwikkelt. Verder is het gewoonlijk dat de ontwikkelaar zelf de trainingsdata selecteert en voorbereid. Het is mogelijk dat de ontwikkelaar ook degene is die het algoritme zelf inzet voor haar eigen doelen, echter veelal is het zo dat de ontwikkelaar een algoritme ontwikkelt in opdracht van een gebruiker.

Gebruiker

De gebruiker is de natuurlijke- of rechtspersoon die het algoritme inzet om een vooraf bepaald doel te bereiken. Over het algemeen zal het een bedrijf of organisatie zijn die de gebruiker is van een algoritme. Hierbij kan je denken aan een verzekeringsmaatschappij die algoritmen inzet om fraude te detecteren, sociale media platformen die inbreukmakende content detecteren en banken die algoritmen inzetten om de kredietwaardigheid van een persoon te beoordelen. Mocht het algoritme intern ontwikkeld zijn, dan komen de rollen 'ontwikkelaar' en 'gebruiker' samen in dezelfde persoon/organisatie.

Hoewel wij hier de ontwikkelaar en de gebruiker als discrete rollen hebben beschreven worden veel modellen in nauwe samenspraak gecreëerd. Niet alleen is de gebruiker de behoeftesteller, ook is het degene die data ter beschikking heeft. De ontwikkelaar op zijn beurt is degene die adviseert over de toepassing van het model en eventuele aanpassingen aan de bedrijfsvoering die daarbij horen.

Eindgebruiker

De eindgebruiker is de persoon die het algoritmische model daadwerkelijk 'bedient' of gebruikt. Een voorbeeld van een eindgebruiker is een dokter die advies krijgt van een algoritmisch model dat wordt ingezet door het ziekenhuis waar de dokter werkt (de gebruiker). Downing et al. beschrijven dit als volgt:

[38] Whitworth, B., & Ahmed, A. (2020). Socio-technical system design. The Encyclopedia of Human-Computer Interaction, 2nd Ed.

> *"The end-user is the person ultimately intended to use a product, as opposed to people involved in developing or marketing it."*[39]

Subject

Het subject is de natuurlijke- of rechtspersoon die 'onderworpen' is aan het algoritmische model. Dat is bijvoorbeeld de persoon die naar een webshop gaat en een speciale aanbieding voorgeschoteld krijgt en de persoon die een zoekterm invult in de zoekbalk van Google.

Nota bene: wij hebben deze bovenstaande rollen als discrete rollen omschreven, maar in de praktijk kunnen deze rollen ook samenvallen. Bijvoorbeeld, de bestuurder van een autonome auto is tegelijkertijd gebruiker, eindgebruiker en subject. Tegelijkertijd is de autofabrikant doorgaans ook de gebruiker. Van geval tot geval moet beoordeeld worden wat de rollen zijn van verschillende actoren en wat dit betekent voor hun verhouding tot de kunstmatige intelligentie toepassing.

3.4 De rol van de mens bij de toepassing van kunstmatige intelligentie

Bij kunstmatige intelligentie toepassingen is de rol van de mens uiteraard van belang. De Personal Data Protection Commission (PDPC), de privacy toezichthouder van Singapore, maakt een nuttig onderscheid tussen *human in the loop*, *human out of the loop* en *human over the loop*.[40]

Human in the loop

In dit model is de mens actief betrokken en heeft volledige controle. De kunstmatige intelligentie toepassing maakt zelf geen autonome beslissingen maar adviseert de eindgebruiker, die zelf een gewogen keuze kan maken.

Human out of the loop

In dit model heeft de mens geen directe controle over de toepassing. De kunstmatige intelligentie kan volledig autonoom beslissen, zonder menselijke tussenkomst of controle.

[39] Downing, D. A et al. (2000). *Dictionary of computer and Internet terms*. Barron's Educational Series Inc..

[40] PDPC (2018). Discussion Paper on Artificial Intelligence (AI) and Personal Data - Fostering Responsible Development and Adoption of AI.

Human over the loop

In dit model kan de mens de parameters voor de uitvoering van de taak van de kunstmatige intelligentie tijdens de uitvoer aanpassen, maar blijft de kunstmatige intelligentie beslissen. Bijvoorbeeld, een mens kan tijdens het autonoom navigeren door een auto de route handmatig wijzigen.

4 Onvoorziene effecten

In dit hoofdstuk zetten we uiteen hoe we de term 'onvoorziene effecten' definiëren, welke distincties hierin gemaakt moeten worden en welke aspecten de (on)voorzienbaarheid van effecten beïnvloeden.

4.1 Voorziene en onvoorziene effecten

In deze paragraaf gaan wij in op verschillende effecten van handelingen. Door een bepaalde menselijke handeling kan een gevolg optreden. Deze handeling kan gewenste of ongewenste gevolgen hebben en deze kunnen voorzien of onvoorzien zijn.

Wanneer je een eenvoudige handeling uitvoert, zoals het oversteken van een lege straat, dan zal je doorgaans de effecten daarvan goed kunnen voorzien en beoordelen. Het effect van het oversteken is dat je aan de andere kant van de straat eindigt. Wanneer deze handeling complexer wordt, of de omgeving waarbinnen de handeling plaatsvindt dat wordt, dan kunnen er effecten optreden die je van tevoren niet had voorzien of bedoeld. Wanneer er bijvoorbeeld veel verkeer is, het mistig is en je tegelijkertijd aan het bellen bent, zijn er veel meer mogelijke uitkomsten, waaronder uitkomsten die je niet voor ogen had vóór het oversteken. Je kan bijvoorbeeld worden aangereden door een fiets of een auto, een auto moet uitwijken en botst tegen een paal, een auto toetert waardoor je je telefoon laat vallen *et cetera*.

Healy maakt een onderscheid tussen *anticipated consequences* (verwachte/voorziene gevolgen) en *unanticipated consequences* (onverwachte/onvoorziene gevolgen).[41]

Voorziene gevolgen kunnen:

- bedoeld en gewenst zijn,
- niet gewenst zijn maar wel redelijkerwijs te verwachten (*probable*),
- niet gewenst zijn en niet redelijkerwijs te verwachten (*improbable*),

Onvoorziene gevolgen zijn:

- gewenst
- ongewenst.[42]

Healy geeft het voorbeeld van de bouw van een kerncentrale aan de kust. Het voorziene en bedoelde effect is dat de kerncentrale energie opwekt. Een ongewenst doch voorzienbaar

[41] Healy, T. (2012). The unanticipated consequences of technology. Nanotechnology: ethical and social Implications, 155-173., p. 1.

[42] Ibid.

effect is de opwarming van het oceaanwater bij de kerncentrale als gevolg van het lozen van koelwater uit de centrale. Een ongewenst maar niet redelijkerwijs te verwachten gevolg is een ontploffing in de kerncentrale. Een onvoorzien doch gewenst gevolg is het spontaan ontdekken van nieuwe werkprocedures die kernenergie veiliger of efficiënter maken. Een onvoorzien en ongewenst gevolg van de bouw van de kerncentrale is dat opwarming van het oceaanwater leidt tot de versnelde evolutie van een nieuwe roofvis die al het andere zeeleven uitroeit.[43]

Pringle *et al.* richten zich naast de voorzienbaarheid van de gevolgen ook op de bedoeling van de acties die leiden tot deze gevolgen.[44] Aldus komen zij tot het volgende schema:

Gevolgen	**Verwacht (voorzien)**	**Onverwacht (onvoorzien)**
Bedoeld	De daadwerkelijke doelstellingen die we nastreven. Omdat we voorzien dat een algoritme een bepaald effect heeft zetten we het in. *Bijvoorbeeld, we zetten een algoritme in om een betere inschatting te kunnen maken van een kredietrisico en dit is ook het geval.*	Geen geldige categorie. Een effect dat onverwacht is kan nooit bedoeld zijn.
Onbedoeld	Effecten die we niet beogen maar wel redelijkerwijs kunnen voorzien. Het kan gaan om zowel positieve als negatieve effecten. Wanneer het negatieve effecten betreft zijn het gevolgen (of risico's) die we moeten uitsluiten of beperken. *Bijvoorbeeld de inzet van een algoritme heeft als bijwerking dat personen onterecht worden uitgesloten van krediet op basis van kenmerken die zij toevallig gemeen hebben met een risicovolle groep.*	Effecten waarvan met redelijkheid vooraf niet kon worden voorzien dat ze zouden optreden. Zogenaamde 'unknown unknowns'. Doorgaans zijn deze effecten door hun onvoorspelbaarheid negatief. *Bijvoorbeeld een autonoom rijdende auto stopt ineens midden op de snelweg omdat het de schaduw van een langs vliegende helikopter aanziet voor een vrachtauto.*

Tabel 1 Categorisering effecten door Pringle et al. (2016), voorbeelden toevoegingen Considerati

[43] Healy, T. (2012). The unanticipated consequences of technology. Nanotechnology: ethical and social Implications, 155-173., p. 2.

[44] Pringle, R., Michael, K., & Michael, M. G. (2016). Unintended Consequences of Living with AI: The Paradox of Technological Potential? Part II [Guest Editorial]. IEEE Technology and Society Magazine, 35(4), 17-21.

In dit rapport ligt de focus op de term 'onvoorziene effecten'. Dit zijn effecten/gevolgen die onverwacht optreden bij de toepassing van een algoritmisch model.

Het is van belang dat 'onvoorzien' een subjectief en relatief begrip is. In hoeverre een effect onvoorzien is hangt in sterke mate af van hoe goed je de toepassing en de mogelijke effecten daarvan doordacht hebt. Wat extreem gesteld: voor een roekeloos of naïef persoon zijn de gevolgen van het eigen handelen misschien onvoorzien, maar voor een voorzichtig iemand zijn diezelfde gevolgen wel voorzien.

4.2 Oorzaken onvoorziene effecten

Om adequaat te kunnen reageren op onvoorziene effecten is het van belang dat we weten waarom effecten onvoorzien zijn. Onvoorziene effecten kunnen volgens Dörner worden toegeschreven aan ons onvolledige begrip van een systeem.[45] Dit onvolledige begrip leidt tot een onvermogen om alle gevolgen van ons handelen te overzien.

Volgens Dörner hebben complexe systemen drie eigenschapen die primair bijdragen aan ons onvermogen om de gevolgen van ons handelen te overzien:

1. Interdependentie;
2. Dynamiek; en
3. Ondoorzichtigheid

Interdependentie

In een systeem beïnvloeden verschillende variabelen elkaar. Naar mate de complexiteit van een systeem groeit, neemt het aantal componenten dat elkaar beïnvloedt toe en wordt het moeilijker om de interactie tussen deze variabelen en de effecten van ons handelen op deze variabelen te overzien.

Dynamiek

Met dynamiek wordt bedoeld dat een complex systeem zelden een statische omgeving is. De omgeving waarin een handeling plaatsvindt is meestal aan veranderingen onderhevig. Hierdoor komen er variabelen bij, verdwijnen variabelen en beïnvloeden variabelen elkaar op een andere wijze. Ook verandert ons handelen en dat van anderen de omgeving.

[45] Dörner, D. (1997). The logic of failure: Recognizing and avoiding error in complex situations. Basic Books.

Ondoorzichtigheid

Wanneer bepaalde aspecten van een systeem voor de gebruiker verborgen zijn, is er sprake van ondoorzichtigheid. De aspecten zijn op zichzelf wel kenbaar / te begrijpen, maar degene die handelt heeft geen toegang tot de benodigde informatie.

Deze drie eigenschappen maken het moeilijk om een complex systeem te doorgronden waardoor ons handelen eerder zal leiden tot voor ons onvoorziene effecten. Hierbij moet wederom de menselijke psyche in ogenschouw worden genomen: sommige mensen zullen de consequenties van hun handelen minder goed overzien dan anderen, schatten risico's anders in enzovoorts.

4.3 Relevantie voor de toepassing van algoritmen

Wanneer wij het bovenstaande beschouwen in het licht van zelflerende algoritmen dan kunnen we vaststellen dat onvoorziene effecten bij de toepassing van algoritmen op twee manieren kunnen ontstaan.

Ten eerste kan het doel waartoe het algoritme wordt ingezet leiden tot onvoorziene effecten. Met andere woorden, het is niet persé het algoritme dat zorgt voor onvoorziene effecten, maar breder gezien de acties die worden genomen om het doel te bereiken.

Ten tweede kan het gebruik van (zelflerende) algoritmen om een bepaald doel te bereiken leiden tot onvoorziene effecten. Hierbij zijn de gevolgen primair toe te schrijven aan de onvoorziene effecten die het gebruik van het algoritme heeft.

Het is de tweede categorie die voor dit onderzoek primair relevant is. Onvoorziene effecten van algoritmen vloeien naar ons inzicht primair voort uit de complexiteit van systemen zoals hierboven beschreven. Zoals beschreven in hoofdstuk 2 is een goed begrip van de 'probleemruimte' cruciaal. Wanneer de probleemruimte een complex systeem is of beschrijft, dan kan een gebrekkig begrip van deze probleemruimte leiden tot algoritmische besluiten die 'verkeerd' zijn en daarmee waarschijnlijk tot onvoorziene effecten leiden.

Interdependentie

Zonder goed begrip van de werking van een systeem (welke variabelen en rol spelen, wat is hun interdependentie en welke effecten heeft ons handelen op deze variabelen en daarmee op het systeem als geheel) is de kans groot dat het algoritme fouten maakt, bijvoorbeeld omdat het een onvolledig begrip heeft van de situatie en de context. Deze fouten leiden doorgaans tot onvoorziene effecten.

Dynamiek

Wanneer een systeem dynamisch is en het beslismodel dat in deze omgeving wordt ingezet is dat niet, dan kunnen de veranderende omstandigheden tot situaties leiden waarop het beslismodel naar verwachting geen adequaat antwoord heeft. De verkeerde besluiten die hieruit voortvloeien kunnen leiden tot onvoorziene effecten.

Ondoorzichtigheid

Ook ondoorzichtigheid kan leiden tot een onvoldoende begrip van de situatie. Als bepaalde informatie voor mensen onbekend of verborgen is, dan betekent dit doorgaans dat het beslismodel dat wordt getraind eenzelfde gebrekkige begrip van de situatie heeft. Het machine learning model wordt namelijk getraind op basis van de beschikbare gegevens. Wanneer deze data een incompleet beeld van het probleem geven, of er door onwetendheid of onbegrip bijvoorbeeld een *bias* in de data zit, dan zal het machine learning model deze onwetendheid overnemen (en mogelijk zelfs versterken).

5 Overzicht casestudies

In dit hoofdstuk geven we een kort overzicht van de verschillende casestudies te weten: 1) algorithmic pricing, 2) kredietwaardigheid en risicoprofielen, 3) HR analytics en 4) fysieke & geestelijke gezondheid.

Algorithmic pricing

De eerste case heeft betrekking op algorithmic pricing. Het dynamisch prijzen van producten of diensten op basis van de inzichten van een algoritme kan grote economische meerwaarde hebben, bijvoorbeeld omdat minder wordt verspild en vraag en aanbod elkaar optimaal weten te vinden. Tegelijkertijd kan dynamisch prijzen mogelijk ook leiden tot oneerlijke prijsdiscriminatie of tot collusie tussen marktpartijen. Zonder inzicht in de totstandkoming van de prijzen en de factoren die daarbij een rol spelen zijn deze ongewenste effecten niet te detecteren.

Kredietwaardigheid en risicoprofielen

Een domein waar algoritmische modellen reeds lange tijd gebruikt worden zijn risico inschattingen. Algoritmen bepalen op basis van risicoprofielen of afwijkende patronen welke transacties en personen verdacht zijn. Financiële instellingen detecteren bijvoorbeeld frauduleuze betalingen of witwassen met behulp van algoritmische modellen. Een concrete casestudy voor de toepassing van algoritmen voor risico-inschatting is het toetsen van de kredietwaardigheid van personen. De bank analyseert bijvoorbeeld op basis van allerlei gegevens over de kredietaanvrager of hij/zij nu en in de toekomst de lening kan en gaat terugbetalen.

HR Analytics

De derde case heeft betrekking op het gebruik van algoritmen voor het monitoren of beoordelen van sollicitanten en medewerkers. Het gebruik van algoritmische besluitvorming kan tot betere HR-beslissingen leiden, maar kan ook de verhouding tussen werkgever en werknemer ingrijpend veranderen en discriminatie in de hand werken. Het gaat hier zowel om het automatisch beoordelen van sollicitanten op basis van CV's en motivatiebrieven, als om het algoritmisch analyseren van de prestaties van medewerkers en beoordelen waar binnen de organisatie verbeteringen kunnen plaatsvinden.

Fysieke & geestelijke gezondheid

De vierde case heeft betrekking op het gebruik van algoritmen om consumenten te helpen (betere) beslissingen te nemen over hun fysieke en/of mentale gezondheid. Hierbij kan gedacht worden aan *nutrition trackers* die advies geven over wat wel/niet te eten, maar ook aan applicaties die helpen bij het omgaan bij mentale problemen, zoals depressie. Deze applicaties registreren de gegevens over het subject, geven advies of verwijzen de persoon door naar een professional. Deze toepassingen ondersteunen mensen om betere keuzes te maken in hun leven en kunnen de zorg ontlasten. Deze toepassingen roepen echter ook vragen op met betrekking tot de autonomie van de consument, en de aansprakelijkheid voor foutieve adviezen.

6 Case study 1: Algorithmic pricing

Vroeger kwamen prijzen tot stand door onderhandeling tussen koper en verkoper. Het prijskaartje en daarmee de 'vaste prijs' deed pas zijn intrede aan het einde van de 19e eeuw.[46] Dit leverde een aantal voordelen op: klanten hoefden niet meer voor elke aankoop te onderhandelen en winkelbedienden konden met minder ervaring in winkels werken. Prijzen veranderden nog steeds, maar een op een individuele situatie afgestemde prijs werd steeds ongebruikelijker. Met de beschikbaarheid van meer data, meer mogelijkheden om deze data te gebruiken en nieuwe 'interfaces' (bijvoorbeeld digitale prijskaartjes, webwinkels, platforms) zijn de mogelijkheden voor *algorithmic pricing* sterk toegenomen.

Algorithmic pricing is een manier van prijsstelling die plaatsvindt op basis van data-analyse en een daaruit voortvloeiend algoritmisch model. Deze analyse zorgt ervoor dat bedrijven op geautomatiseerde wijze de prijzen voor hun diensten en producten in *real time* kunnen aanpassen per potentiële klant.[47] Dit mechanisme wordt over het algemeen ingezet door bedrijven die producten en diensten verkopen die op de korte termijn verkocht moeten worden. Voorbeelden hiervan zijn: hotelboekingen, vliegtickets en seizoensgebonden mode.[48]

Algorithmic pricing is onder te verdelen in twee categorieën, namelijk *dynamic pricing* en *personalised pricing*. Dynamic pricing, ook wel *surge*, *yield* of *real-time pricing* genoemd, kan als volgt worden gedefinieerd:

> *"the practice of dynamically adjusting prices in order to achieve revenue gains, while responding to a given market situation with uncertain demand"*[49]

Hierbij worden de prijzen dus aan de omstandigheden, in het bijzonder de vraag, aangepast.

Personalised pricing kan als volgt worden gedefinieerd:

> *"first-degree price discrimination, customized, or targeted pricing, and represents a pricing strategy "whereby firms charge different prices to different consumers based on their willingness to pay."*[50]

---

[46] Fountain, N. (producer). (2015). Planet money Episode 633: The Birth and Death of The Price Tag [Podcast]. Geraadpleegd via: https://www.npr.org/sections/money/2015/06/17/415287577/episode-633-the-birth-and-death-of-the-price-tag?t=1599817295236

[47] Seele, P. et al. (2019). Mapping the Ethicality of Algorithmic Pricing: A Review of Dynamic and Personalized Pricing. Journal of Business Ethics, 1-23.

[48] Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. Management science, 49(10), 1287-1309.

[49] Seele, P. et al. (2019). Mapping the Ethicality of Algorithmic Pricing: A Review of Dynamic and Personalized Pricing. Journal of Business Ethics, 1-23.

[50] Ibid.

Het idee achter personalised pricing is dus dat voor ieder product of dienst de meest optimale prijs wordt gevraagd aan de individuele klant. De theorie is dat iedere (potentiële) klant een ander maximaal bedrag uitgeeft voor hetzelfde product of dienst.

De twee vormen van prijsstelling kunnen als volgt van elkaar worden onderscheiden. Waar een systeem gebaseerd op *dynamic pricing* de prijzen doet fluctueren op basis van vraag en aanbod en andere externe factoren, verandert een systeem gebaseerd op *personalised pricing* de prijzen ook op basis van individuele kenmerken van de (potentiële) klant.[51]

Personalised pricing lijkt nog weinig gebruikt te worden. Kortingen gebaseerd op persoonlijke kenmerken worden wel al toegepast. Bijvoorbeeld bij de "Persoonlijke bonusaanbiedingen" van Albert Heijn. Hier is niet echt sprake van individuele korting, maar van een vaste korting die aan een subgroep van de klanten wordt aangeboden. In de praktijk kan dit op de volgende manier werken: leverancier X geeft aan dat product Y met een korting aangeboden mag worden. Deze korting is 15%. Klant A koopt regelmatig product Y. Het model weet dit op basis van aankoophistorie. Klant A krijgt dus product Y aangeboden met de vaste korting van 15%. Iedere klant die een soortgelijke aankoophistorie heeft als klant A krijgt voor hetzelfde product dezelfde vaste korting. In andere woorden:

> *Als je persoonlijke korting krijgt, dan krijgt iedereen die korting krijgt dezelfde korting.*
>
> *Sven Schagen, Director Data Science Albert Heijn*

6.1 Doel algorithmic pricing

Een *algorithmic pricing* model wordt toegepast om op basis van vooraf bepaalde data de markt waarin een product of dienst zich bevindt te herkennen en een optimale prijsstrategie uit te voeren. Handmatige analyse en prijsbepaling is door de hoeveelheid producten en de verschillende data die hiervoor gebruikt kunnen worden niet meer praktisch uitvoerbaar. Algorithmic pricing automatiseert (een deel van) dit proces. Algorithmic pricing modellen worden intern ontwikkeld of ingekocht (vaak als dienst) van een andere partij.

Bij de analyse van de markt (voor een product) kunnen zowel interne data (verkoopcijfers, voorraad), prijzen van concurrenten, productkenmerken, als externe factoren (seizoen, trends, et cetera) gemodelleerd worden. Deze inzichten vormen het uitgangspunt voor het bepalen en uitvoeren van een strategie en het stellen van de uiteindelijke prijs. Deze prijsstrategieën zijn

[51] OESO (2018), Personalised Pricing in the Digital Era. DAF/COMP(2018)13, p. 9.

bestaande strategieën zoals het volgen van een concurrent, of juist marktaandeel winnen door lagere prijzen te hanteren.

Algorithmic pricing wordt ook ingezet om klanten te bewegen producten te kopen die anders snel hun waarde verliezen. Een voorbeeld hiervan is het dynamisch afprijzen van versproducten om te voorkomen dat een supermarkt deze weg moet gooien.[52]

Hoe de prijsstelling verloopt hangt af van de context van de toepassing. In (retail) winkels, in het bijzonder webshops, kunnen prijzen bijvoorbeeld automatisch worden aangepast. Pricing algoritmen worden ook toegepast met een adviserende functie voor verkopers. Als verkopers veel autonomie hebben om prijzen te bepalen en kortingen te geven kan een pricing algoritme een advies geven over de prijs en maximale korting.

6.2 Implementatie en inzet

De implementatie van algorithmic pricing modellen zorgt ervoor dat de gebruiker inzicht krijgt in wat voor prijzen gevraagd kunnen worden om op een zo efficiënt mogelijke manier een bepaalde bedrijfsstrategie te bereiken. De implementatie leidt er ook toe dat daar waar het veranderen van prijzen voorheen door mensen werd bedacht en uitgevoerd, dit nu grotendeels wordt overgenomen door het algoritme en de software die daar omheen is gebouwd.

Verder zorgt het voor een verandering in de manier waarop met klanten wordt omgegaan, voornamelijk bij business-to-business verkopen waar onderhandelingen over prijzen plaatsvinden. De onderhandelingen vonden voorheen tussen mensen plaats met veel domeinkennis maar met gelimiteerde kennis over alle data gerelateerd aan het te verkopen product. Nu worden de eindgebruikers in deze rol ondersteund door het algoritmisch model, dat wel enorme 'kennis' heeft over de factoren die relevant zijn bij het prijzen van een product of dienst. Dit zorgt voor een andere dynamiek bij de prijsonderhandelingen.

De gebruiker en de eindgebruiker hebben met de implementatie van het algoritme dus meer kennis en controle over het prijzen van hun producten of diensten.

Wanneer de implementatie echter niet succesvol plaatsvindt kan dat uiteenlopende effecten hebben. Het is bijvoorbeeld mogelijk dat het algoritmisch model onbedoeld significant lagere

[52] Business Insider, 'Albert Heijn doet test met digitaal prijskaartje: kip en vis afprijzen op basis van de houdbaarheid', 20-05-2019, https://www.businessinsider.nl/albert-heijn-doet-test-met-digitaal-prijskaartje-kip-en-vis-afprijzen-op-basis-van-de-houdbaarheid/.

prijzen vraagt voor producten of diensten.[53] Verder is het ook mogelijk dat vanwege onverwachte situaties onethisch hoge prijzen worden gevraagd voor producten/diensten.[54]

Om dit te voorkomen kunnen ontwikkelaars en gebruikers in het algoritmische model bepaalde parameters en/of grenzen voor de *output* van het model vaststellen. Omdat de prijsstelling van producten en diensten grote impact kan hebben op bedrijfsresultaten worden diverse mechanismen toegevoegd om de prijsstelling binnen vooraf bepaalde grenzen te laten verlopen.[55] Deze grenzen kunnen op verschillende manieren en op verschillende niveaus ingesteld worden. Een product kan bijvoorbeeld automatisch uit een webwinkel gehaald worden, of een eindgebruiker kan een signaal krijgen om een bepaalde prijs te controleren (*human over the loop*).

Verder hangt de implementatie van pricing algoritmen af van het specifieke doel en de systemen van een organisatie. Uit de interviews blijkt dat (eind-)gebruikers van algoritmen, verkopers (die advies krijgen van het systeem) of *category managers* (die verantwoordelijk zijn voor een productgroep) inzicht willen hebben hoe een output van een algoritme tot stand is gekomen. Aanbieders geven inzicht in deze logica door bijvoorbeeld een dashboard beschikbaar te stellen waar informatie over het systeem en specifieke uitkomsten wordt ontsloten. Naast deze manier om besluiten te monitoren wordt ook vaak een mogelijkheid gegeven om de werking van het model of een specifieke uitkomst te wijzigen.

Uit de interviews blijkt evenwel dat technische maatregelen alleen niet voldoende zijn. Om tot een succesvolle implementatie te komen is voortdurend contact tussen ontwikkelaar en gebruiker noodzakelijk. De ontwikkelaar helpt de gebruiker bij de succesvolle inzet van het algoritme en fungeert als technische support, vraagbaak, en sparringpartner.

Er zijn natuurlijk ook organisaties die intern algoritmen en de bijbehorende systemen ontwikkelen. Hier is vaak sprake van projecten die het hele traject van het formuleren van de businesscase, ontwikkeling, implementatie en evaluatie beslaan.

[53] Kollmeyer, Barbara., 'Lucky travelers score $6.99 tickets to Hawaii after Delta glitch', 27-12-2013, MarketWatch, https://www.marketwatch.com/story/lucky-travelers-score-699-tickets-to-hawaii-after-delta-glitch-1388138286.

[54] Riley, C. (2017). Uber criticized for surge pricing after London terror attack . CNN. Geraadpleegd via: https://money.cnn.com/2017/06/04/technology/uber-london-attack-surge-pricing/index.html.

[55] Bijvoorbeeld: nooit onder de inkoopprijs, nooit een negatief bedrag

> *Het gaat niet om het algoritme. Het gaat om: wat ben je aan het leren? En wat ben je aan het optimaliseren met dat leerproces? Daarom beginnen we met de eenvoudigste modellen die we kunnen maken, want die kun je begrijpen en ook gaan begrijpen wat wel en niet voor de klant werkt.*
>
> Sven Schagen, Director Data Science Albert Heijn

6.3 Impact van algorithmic pricing op prijsstelling

De inzet van algoritmen voor de prijsstelling heeft een aantal effecten op de manier waarop prijsstellingen worden verricht.

Automatisering

Het toepassen van algoritmen voor prijsstelling zorgt ervoor dat een deel van het werk dat eerst door een eindgebruiker werd gedaan geautomatiseerd wordt. Het gevolg hiervan is dat op hogere snelheid en grotere schaal prijzen kunnen worden bepaald op basis van traditionele en nieuwe kenmerken.

Voor zowel de eindgebruiker als voor de gebruiker is de impact groot omdat een deel van het werk wordt geautomatiseerd en er nieuwe werkzaamheden ontstaan. De eindgebruiker hoeft bijvoorbeeld niet zelf grote berekeningen te maken over de prijsstelling, maar de eindgebruiker moet nu de resultaten van het algoritmisch model interpreteren, begrijpen en daarin goede keuzes maken. Voor dezelfde rol binnen de organisatie van de gebruiker moet de eindgebruiker dus beschikken over nieuwe eigenschappen om de rol succesvol uit te kunnen voeren.

Voor de gebruiker is de impact van algorithmic pricing zo dat een veel gedetailleerder en actueler inzicht ontstaat over alles wat leidt tot de prijsstelling van een product of dienst. Verder heeft de gebruiker mensen nodig die inzicht hebben in de werking en het gebruik van het algorithmic pricing model. De gebruiker dient de eindgebruiker te trainen, zodat deze bekend is met de werking van het algoritme, of de gebruiker dient nieuwe mensen aan te nemen met expertise in *data science* en de werking van algoritmische modellen.

Uit onze interviews kwam de training van de gebruiker en de eindgebruiker ook naar voren. Voor de ontwikkelaar is het namelijk van belang dat de gebruiker en de eindgebruikers begrijpen hoe het model werkt om zo tot een succesvolle samenwerking te komen.

> *Daar proberen we heel duidelijk in te zijn. Als de klant niet voldoende klaar is voor onze toepassing, dan wordt het niet succesvol.*
>
> Ontwikkelaar pricing algoritmen

Wederzijdse afhankelijkheid tussen ontwikkelaar en gebruiker

Het toepassen van een extern ontwikkeld algoritme betekent dat er meer van de (informatie die relevant is voor) besluitvorming gebaseerd wordt op de kennis van een externe partij. Dat zorgt onder meer voor wederzijdse afhankelijkheid tussen ontwikkelaar en gebruiker. Het feit dat specifieke marktkennis in te kopen is kan zorgen voor een lagere drempel om toe te treden tot de markt (nieuwkomers hoeven geen diepgaande kennis over de markt te hebben).

Efficiëntere uitvoer van prijsstrategie

Het algoritmisch uitvoeren van een prijsstellingsstrategie zorgt ervoor dat deze accurater en efficiënter uitgevoerd kan worden vergeleken met menselijke prijsstellers. Dit zorgt er bijvoorbeeld voor dat de gebruiker sneller kan inspelen op veranderingen in de markt, doelgerichter producten of diensten kan prijzen aan de hand van de gekozen bedrijfsstrategie en met meer informatie een prijsonderhandeling succesvol kan afronden. Een ander gevolg is dat het voor concurrenten en consumenten zichtbaarder kan worden welke strategie wordt gehanteerd. In het kader van vliegtickets en hotelkamers hebben veel consumenten een algemeen idee hoe de prijzen zich ontwikkelen in de tijd (een vliegticket een week voor vertrek goedkoper dan op de dag van vertrek). Hierdoor is denkbaar dat in de toekomst ook voor andere productgroepen of specifieke aanbieders dergelijke informatie beschikbaar is.

6.4 Potentiële onvoorziene effecten van algorithmic pricing

Bij de toepassing van algoritmische modellen in de context van algorithmic pricing kunnen allerlei effecten ontstaan die voor een of meerdere actoren onvoorzien zijn. In deze paragraaf noemen we de effecten die wij op basis van het literatuuronderzoek en de interviews hebben geïdentificeerd als mogelijk onvoorzien.

Eerlijkheid prijzen

Omdat pricing algoritmen prijzen aanpassen is het mogelijk dat in de optiek van bepaalde actoren deze veranderingen als oneerlijk worden gezien. Voor diverse productgroepen is het gebruikelijk dat prijzen regelmatig veranderen (vliegtickets, elektronica, platformdiensten).

Voor andere producten is het minder gebruikelijk dat prijzen veel variëren. Het blijkt dat consumenten uiteenlopend denken over verschillende manieren om prijzen te variëren.[56]

Veel pricing algoritmen richten zich op de *willingness to pay* van een potentiële klant. In normale situaties is wat als een eerlijke prijs wordt gezien en wat men bereid is te betalen redelijk gelijk. Maar in situaties waarbij een de vraag naar een product of dienst snel groeit kan een corresponderende prijsverandering als oneerlijk worden gezien. Een extreem voorbeeld zijn de prijzen van een taxirit van Uber die omhoogschoten tijdens een terroristische aanslag in Londen.[57]

Een dergelijk onvoorzien effect kan zich voordoen wanneer onvoldoende of geen monitoring plaatsvindt bij de *deployment* van het model. Door de *output* van het model goed te monitoren, en limieten in te stellen voor bepaalde waarden, kunnen onbedoelde uitschieters voorkomen en opgemerkt worden.

> *Niemand heeft het door, maar je kan wel allemaal dingen doen die de business negatief beïnvloeden, […] gebrek aan governance en monitoring op algoritmen is het allergrootste hiaat op dit moment.*
>
> *Ruud Schmeink, CEO MarketRedesign*

Onjuiste prijzen

Het scenario van een algoritme dat per ongeluk veel te lage prijzen vraagt spreekt tot de verbeelding. Er zijn talloze voorbeelden van producten die voor 'verkeerde' prijzen worden aangeboden.[58] Net als bij traditionele prijsstelling kunnen bij algorithmic pricing foutieve prijzen gevraagd worden. Automatisering en de frequentie waarmee prijzen kunnen worden aangepast maakt dat de impact van een foutieve prijs groot kan zijn. Om dit te voorkomen worden diverse controlemechanismen ingebouwd in het systeem:

> *We maken de klant bewust van de risico's. Bij de meeste klanten zitten er 'safety valves' ingebouwd in het systeem, zodat prijzen niet hele gekke sprongen kunnen maken.*
>
> *Ontwikkelaar pricing algoritmen*

[56] Poort, J. & Zuiderveen Borgesius, F. J. (2019). Does everyone have a price? Understanding people's attitude towards online and offline price discrimination. Internet Policy Review, 8(1). DOI: 10.14763/2019.1.1383

[57] Riley, C. (2017). Uber criticized for surge pricing after London terror attack . CNN. Geraadpleegd via: https://money.cnn.com/2017/06/04/technology/uber-london-attack-surge-pricing/index.html.

[58] Bijvoorbeeld een stapelbed met 91% korting: https://nos.nl/artikel/2186332-net-als-bij-leen-bakker-waren-ook-deze-aanbiedingen-te-mooi-om-waar-te-zijn.html

Onderdeel van dergelijke mechanismen is dat verdachte veranderingen van prijzen worden geblokkeerd en bijvoorbeeld ter attentie van een eindgebruiker worden gebracht.

Onjuiste prijzen kunnen ook het effect zijn van externe oorzaken die de effecten van het algoritme beïnvloeden. Een voorbeeld is gebruik door organisaties die niet de capaciteit hebben het systeem goed te implementeren, of (eind-)gebruikers te weinig kennis hebben voor een correcte toepassing.

Door het model uitvoerig te testen, governance in te richten, mogelijkheden voor monitoring te ontwerpen, en gebruikers en eindgebruikers te trainen kunnen onjuiste prijzen voorkomen worden.

> *Het is gewoon kwaliteitsborging. Dat iemand erop toeziet dat er gecontroleerd wordt wat dat ding doet.*
>
> *Ruud Schmeink, CEO MarketRedesign*

Beïnvloeding

Beïnvloeding van koopgedrag is van alle tijden. Reclame, kortingen, prijzen die aantrekkelijker lijken (bijvoorbeeld een prijs van €9,99 in plaats van €10,-), zijn slechts een enkele voorbeelden van manieren waarop aanbieders de vraag naar hun product proberen te beïnvloedden. De toepassing van personalised pricing maakt het mogelijk om een aanbod te doen dat aantrekkelijker is voor specifieke (groepen) consumenten. Het effect hiervan hangt sterk af van de context. De ACM stelt dat een gepersonaliseerd aanbod in het voordeel van de consument kan zijn, maar dat er ook ingespeeld kan op specifieke kwetsbaarheden van consumenten (met nadelige gevolgen voor de consument).[59]

Dit effect ontstaat bij het bepalen van het doel waarvoor een pricing algoritme wordt ingezet. In zekere zin *is* dit het doel waarvoor algorithmic pricing wordt ingezet. Het voorkomen van ongewenste beïnvloeding is dus al bij het bepalen van de *business case* aan de orde. Een manier om hiermee om te gaan is het documenteren van besluiten in het hele proces:

> *Documenteer alle keuzes. Ontwikkelaars doen allemaal dingen met het algoritme, maar je moet het wel gevalideerd hebben met de klant. Bijvoorbeeld het vooraf afstemmen van performance limieten en modelkwaliteit.*
>
> *Ruud Schmeink, CEO MarketRedesign*

[59] Autoriteit Consument en Markt (2020). Leidraad Bescherming online consument. Geraadpleegd via: https://www.acm.nl/nl/publicaties/leidraad-bescherming-online-consument

Dat de normatieve implicaties van beïnvloeding per casus sterk kunnen verschillen laat de casus van het dynamisch afprijzen van versproducten zijn. Hier is het doel om minder versproducten weg te hoeven gooien, door gericht versproducten af te schrijven (op basis van de voorraad, weer, houdbaarheidsdatum, enzovoorts). Als klanten niet beïnvloed worden dan moeten de producten worden weggegooid.

Prijsdifferentiatie / prijsdiscriminatie

Het doel van algorithmic pricing is het aanpassen van prijzen in de loop van de tijd, bij bepaalde kenmerken van de aankoop (kwantumkorting, bundelkorting) of tussen verschillende (soorten) klanten.

Het aanpassen van prijzen op basis van persoonlijke kenmerken is het relevantst in het licht van onvoorziene effecten. In essentie worden hierbij kenmerken van een subject, of kenmerken van anderen waar het subject op lijkt gebruikt voor het bepalen een prijs. De kans is aanwezig dat deze profilering onverwacht plaatsvindt of onvoorziene effecten heeft. De profilering kan onjuist zijn of (per ongeluk) discrimineren. Onterechte profilering als onvoorzien effect ontstaat onder andere als kenmerken ten onrechte gebruikt worden om subjecten in groepen in te delen.

Voor dit effect geldt hetzelfde als voor 'oneerlijke profilering' en 'discriminatie', dat de fases waarin dit onvoorziene effect gemitigeerd kan worden in de fases zitten waarin wordt bepaald welke data wordt gebruikt voor het model.

Afhankelijkheid

De inzet van algorithmic pricing modellen ook leiden tot verlies van controle en onafhankelijkheid bij de (eind)gebruiker.

Voor de gebruiker is het bepalen van een prijsstrategie en het bepalen van een specifieke prijs voor een product of dienst van groot belang. Door de inzet van extern ingekochte algoritmen kan de gebruiker afhankelijk worden van een externe leverancier van het prijsstellingsmechanisme. Voor een deel is deze afhankelijkheid inherent aan het uitbesteden van een deel van de werkzaamheden aan een externe leverancier. Maar uit de interviews blijkt dat er zeker niveau van (technische-)kennis nodig is in de organisatie van de gebruiker om de pricing algoritmen op een verantwoorde manier in te zetten.

Ook voor de eindgebruiker kan een mate van afhankelijkheid van het pricing algoritme ontstaan. Een deel van de werkzaamheden worden immers efficiënt en effectief overgenomen van de eindgebruiker.

6.5 Algorithmic pricing in de toekomst

De toekomst van algorithmic pricing lijkt te vatten in drie trends: bredere adoptie, verdere automatisering, en meer personalisering.

Bredere adoptie

Er is een verschil tussen het niveau van adoptie van algorithmic pricing binnen en tussen sectoren. Het is de verwachting dat de beschikbaarheid van algorithmic pricing toepassingen toeneemt. Hierdoor kunnen ook nieuwe onvoorziene effecten ontstaan, met name op het gebied van interactie tussen pricing algoritmen. Bij dat laatste moet wel opgemerkt worden dat bij bredere adoptie met gelijkaardige algoritmen slechts één *component* van de prijsbepaling gelijkaardig is tussen aanbieders, andere zaken die de prijs bepalen variëren nog steeds tussen aanbieders, en dus de prijs ook.

> *Het enige potentiële risico dat er is, maar daar zijn we heel ver vanaf, als systemen echt meer voorspellen en zo efficiënt worden, dan zou er in theorie een deadlock kunnen ontstaan. Omdat zoveel variabelen een rol spelen, moeten we nog maar zien of dat daadwerkelijk gaat gebeuren. Hieruit zou een onbedoeld effect kunnen komen.*
>
> *Ontwikkelaar pricing algoritmen*

Verdere automatisering

Pricing algoritmen automatiseren een deel van het pricing proces. Het is de verwachting dat deze automatisering verder doorzet. Bijvoorbeeld op het gebied van advies over prijsstrategieën. Dit is relevant omdat dit wellicht impact heeft op de verantwoordelijkheid voor de gekozen strategie. Daarnaast blijkt dat met verdere automatisering ook meer aandacht geschonken moet worden aan monitoring en risicobeheersing.

Meer en diepere personalisering

Hoewel persoonlijke prijzen nog weinig voor lijken te komen is het de verwachting dat personalisatie van prijzen en kortingen toeneemt. Personalisering kan bijvoorbeeld, plaatsvinden door op steeds diepere niveaus trends te ontdekken bij individuele klanten of

groepen klanten.[60] Daardoor voelen klanten steeds meer dat desbetreffende verkoper naar de wensen van hen 'luistert'.

> *Uiteindelijk is de long term loyality en customer intimacy - aansluiting bij de behoeften van de consument - belangrijk.*
>
> *Sven Schagen, Director Data Science Albert Heijn*

Het is voor organisaties steeds duidelijker wat de kosten zijn om een nieuwe klant te krijgen. Personalisatie van prijzen en kortingen is een manier om klanten te binden en te kunnen blijven concurreren met branchegenoten, maar ook om te voldoen aan groeiende verwachtingen van klanten (die op steeds meer gepersonaliseerde producten en diensten gebruiken).

[60] Een klantengroep kan bijvoorbeeld gesignaleerd worden door de ontwikkelaar of gebruiker op basis van aankoophistorie vergeleken met de aankoophistorie van andere klanten. Wanneer de aankoophistorie soortgelijk is tussen een groep mensen, dan kunnen deze mensen geclusterd worden.

7 Case study 2: Kredietwaardigheid en risicoprofielen

Kredietwaardigheidstoetsen en risicoprofielen van consumenten spelen een belangrijke rol in financiële dienstverlening. Om de risico's van een lening in te kunnen schatten worden door financiers kredietwaardigheidsscores, -rapportages, en -profielen gebruikt. Kredietgevers kunnen hiermee beter inschatten wat de kans is dat een lening wordt terugbetaald en daarmee de risico's (en dus de kosten) terugdringen. Daarnaast dragen kredietbeoordelingen bij aan het voorkomen van overkreditering van consumenten en bedrijven. Betrouwbare, waarheidsgetrouwe inschattingen van kredietwaardigheid zijn dus een belangrijk onderdeel van verantwoorde toegang van bedrijven en consumenten tot krediet. [61] Kredietwaardigheidstoetsen worden gebruikt bij het inschatten van het risico voor alle soorten consumentenkrediet variërend van geldkrediet (lening, doorlopend krediet) tot goederenkrediet (kopen op afbetaling, leasing).[62]

Traditioneel wordt kredietwaardigheid beoordeeld op basis van individuele 'positieve' en 'negatieve' informatie zoals het afbetalen van een lening of het krijgen van een aanmaning[63], en statistische/demografische gegevens (vaak op postcode niveau).[64] Tegenwoordig worden deze traditionele indicatoren van risico aangevuld met nieuwe (big) data uit verschillende bronnen.[65]

7.1 Doel kredietwaardigheidstoetsen

De rol van de kredietwaardigheidstoetsen, de gebruikte gegevens, en de wijze van 'scoring' varieert tussen landen en organisaties.[6667] In Nederland zijn naast het door de banken en andere financiële dienstverleners opgerichte Stichting Bureau Kredietregistratie (BKR) een aantal commerciële kredietbeoordelaars actief. Een relatief recente toepassing van

[61] World Bank Group. (2019). Credit Reporting Knowledge Guide 2019. Washington: World Bank Group; Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. Yale Journal of Law and Technology, 18(148).

[62] AFM, 'Consumptief Krediet', https://www.afm.nl/nl-nl/professionals/onderwerpen/consumptief-krediet-fp.

[63] World Bank Group. (2019). Credit Reporting Knowledge Guide 2019. Washington: World Bank Group. Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. Yale Journal of Law and Technology, 18(148).

[64] Definitie 'credit scoring' in: Gedragscode verwerking persoonsgegevens. Nederlandse vereniging van (handels)informatiebureaus. Geraadpleegd op 29 juni 2020 via https://www.nvhinfo.nl/htm/gedragscode.pdf

[65] Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. Yale Journal of Law and Technology, 18(148).

[66] World Bank Group. (2019). Credit Reporting Knowledge Guide 2019. Washington: World Bank Group. Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. Yale Journal of Law and Technology, 18(148).

[67] Curley, Christopher., 'Many countries don't use credit scores like the US – here's how they determine your worth', 20-08-2018, Business Insider, https://www.businessinsider.nl/credit-score-around-the-world-2018-8?international=true&r=US.

kredietwaardigheidsbeoordelingen (hierna: kredietbeoordelingen) is een 'betaal later' optie bij webwinkels (als aanvulling op of vervanging van de traditionele acceptgiro).[68]

Kredietgevers die voor hun potentiële klanten kredietbeoordelingen doen baseren deze beoordeling vaak op een combinatie van informatie van externe kredietbeoordelaars en interne (klant-)gegevens.

Inaccurate kredietbeoordelingen kunnen een grote impact hebben op zowel de kredietaanvrager als de kredietgever. Enerzijds zorgt een onterechte negatieve beoordeling in veel gevallen ervoor dat de kredietaanvrager de gewenste lening of financiering niet kan krijgen. Anderzijds kan een onterecht positieve beoordeling juist zorgen voor risicovolle leningen (voor zowel de kredietgever als de kredietaanvrager). Een adequate beoordeling uitgevoerd door een algoritme is dus in het belang van zowel de kredietgever (de gebruiker van het algoritme) als de kredietaanvrager (het subject van de kredietbeoordeling). [69]

7.2 Implementatie en inzet

De implementatie van kredietbeoordelingen hangt af van het verkoopkanaal. Doorgaans vormt de uitkomst van het algoritmische model (de kredietscore) de basis voor het besluit van de kredietgever. Het is dus niet de kredietbeoordelaar die besluit of iemand een lening krijgt of niet, maar de kredietgevers zelf op basis van hun risicobereidheid.[70] De beoordeling kan volledig geautomatiseerd plaatsvinden (bijvoorbeeld bij een webwinkel) of voorgelegd worden aan een eindgebruiker die een besluit neemt.

De modellen waarmee kredietbeoordelingen worden verricht zijn traditioneel gebaseerd op historische gegevens over de kredietwaardigheid consumenten. Het is aan de aanbieder van het krediet te bepalen welke risico's aanvaardbaar zijn (of welke voorwaarden dan gelden).

7.3 Impact AI voor bepalen kredietwaardigheid

De impact van algoritmische modellen op kredietwaardigheidstoetsen moet in deze context met name begrepen worden als het inzetten van machine learning voor

[68] Daarbij is het in het verboden (Art. 7:26 BW) in het algemeen kopers te verplichten het hele aankoopbedrag vooruit te laten betalen. De 'achteraf betalen' optie is een manier om aan deze vereiste te voldoen.

[69] Dit neemt niet weg dat een kredietbeoordeling, net als andere gegevens van subjecten, ook gebruikt kan worden voor illegale of oneerlijke besluiten, bijvoorbeeld door subjecten met een verhoogd risico onterecht te weigeren voor een krediet.

[70] Indien het algoritmisch model intern is ontwikkeld door de kredietgever, dan vallen de rollen van kredietbeoordelaar en kredietgever vanzelfsprekend samen.

kredietwaardigheidsbeoordelingen op basis (big) data. Deze implementatie heeft op een aantal aspecten impact op de praktijk van kredietbeoordeling.

Gedetailleerdere profielen kredietaanvrager

Er worden meer, en persoonlijkere, gegevens gebruikt bij het bepalen van kredietwaardigheid met een algoritme. Traditioneel wordt geautomatiseerde kredietbeoordeling gebaseerd op (een combinatie van) gegevens als betaalgedrag, en geaggregeerde gegevens zoals betaalgedrag van mensen in een bepaalde regio (via postcodes + huisnummers). Nu meer data beschikbaar zijn voor de kredietbeoordelaar en er systemen in staat zijn al deze data snel te verwerken, doen bedrijven kredietbeoordelingen op basis van het principe "*all data is credit data*".[71] Data die in deze additionele modellen gebruikt worden zijn bijvoorbeeld:[72]

- Beroep, werkhistorie
- Afgesloten abonnementen (telefoon, media, clubs)
- Sociale media gebruik[73] en content
- Spaargeld, eigendommen, faillissementen

Traditionele kredietbeoordelingen werden al met algoritmen uitgevoerd maar op basis van een beperkt aantal indicatoren. Het toevoegen van nieuwe kenmerken aan de analyse zorgt ervoor dat de profielen van kredietaanvragers veel gedetailleerder zijn voordat de kredietbeoordeling plaatsvindt.

Gedetailleerdere beoordelingen

Uit de interviews blijkt dat het toevoegen van meer kenmerken tot gedetailleerdere en preciezere resultaten kan leiden. Ook kan het modelleren van traditionele gegevens met behulp van machine learning betere modellen opleveren. Dit komt doordat traditionele modellen (vaak logistische regressies) een relatief simplistische (lineaire) beslissingsgrens (*decision boundary)* kennen. Hierdoor kunnen individuen 'verkeerd' beoordeeld worden. Modellen die tot stand zijn gekomen met behulp van machine learning hebben veel complexere (multi-dimensionale) beslissingsgrenzen en kunnen dus preciezere uitkomsten genereren.

[71] Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. Yale Journal of Law and Technology, 18(148).

[72] Ibid.

[73] Er zijn zelfs partijen die stellen dat alleen op basis van de gegevens verkregen uit sociale media een accurate creditscore kan worden verkregen, zie bijvoorbeeld: https://www.bigdatascoring.com/study-of-credit-scorecard-using-only-facebook-data-3/

> *One of the benefits of machine learning is that you can narrow down and distinguish very well between people who normally would have been rejected. With machine learning models you can classify people better, you can assess the risks better, you can have better loans for more people.*
>
> Wereldwijde aanbieder van (krediet) informatiediensten en producten

Actuelere modellen

Traditionele modellen van kredietbeoordeling zijn gebaseerd op historische data van tenminste een jaar tot enkele jaren oud. De inzet van machine learning geeft de mogelijkheid sneller en meer data te analyseren en modelleren. Dat is bijvoorbeeld relevant als de historische modellen door externe factoren slechter presteren. Een economische crisis kan ervoor zorgen dat subjecten veel lager scoren op een bepaalde indicator voor kredietwaardigheid in het model. Het gevolg is dat veel minder subjecten kredietwaardig worden geacht, terwijl dat niet noodzakelijk zo hoeft te zijn. Deze vorm van *concept-drift,* de gebruikte kenmerken zijn niet meer adequaat om een bepaalde voorspelling te doen, kan voorkomen worden door aanvullende (machine learning) modellen te gebruiken, modellen sneller te hertrainen, of nieuwe kenmerken toe te voegen die een beter beeld geven van de situatie van een subject.

Snellere doorloop beoordelingen

Het gebruik van algoritmen zorgt ervoor dat een kredietbeoordeling sneller kan plaatsvinden dan voorheen, ondanks het gebruik van gedetailleerdere profielen van kredietaanvragers. Indirect leidt dit ertoe dat de klantervaring over het algemeen beter is (buiten het resultaat van de beoordeling om), want het aanvragen van een beoordeling en het verkrijgen van het resultaat kan zeer snel gebeuren. De verbetering van de klantervaring en de snelheid waarop een beoordeling kan plaatsvinden zorgt ervoor dat de drempel voor gebruikers lager is om producten of diensten te verkopen op basis van een kredietaanvraag met een kredietbeoordeling.

7.4 Potentiële onvoorziene effecten van algoritmen op kredietwaardigheidstoetsen

In deze paragraaf beschrijven we onvoorziene effecten die kunnen ontstaan bij de inzet van algoritmen voor de beoordeling van kredietwaardigheid met behulp van machine learning en nieuwe (big) data. Het betreft hier niet noodzakelijkerwijs een uitputtende opsomming.

Machtsverschillen en informatieasymmetrie tussen kredietgever- en aanvrager

Kredietwaardigheidstoetsen worden ingezet om informatieasymmetrie te verkleinen tussen kredietgever en kredietaanvrager (in het voordeel van de kredietgever). De vraag lijkt te zijn of het verkleinen van de asymmetrie niet omslaat in het voordeel van de kredietgever. Dit is bij traditionele kredietbeoordelingen aan de orde en lijkt ook bij inzet van machine-learning modellen het geval te zijn.[74] Aan de andere kant blijkt uit de interviews dat het gebruik van nieuwe data juist ook het subject in staat kan stellen een beslissing bij te stellen door aanvullend 'bewijs' aan te leveren.

De inzet van nieuwe gegevens voor de beoordeling maakt het machtsverschil en de informatieasymmetrie tussen de kredietgever en het kredietaanvrager potentieel groter. Wanneer de kredietgever een complex model inzet met een groot aantal kenmerken, is het moeilijker om uit te leggen aan de kredietaanvrager hoe een resultaat tot stand is gekomen. Sowieso kan het in het belang van de kredietbeoordelaar zijn om het specifieke model waarmee besluiten worden genomen geheim te houden, dit om te voorkomen dat kredietaanvragers de beoordeling manipuleren (*gaming the system*).[75]

Voor inzicht in de gebruikte gegevens en de logica van de besluitvorming zijn een aantal fasen in het ontwikkel- en implementatietraject van belang. Bij het bepalen van de te analyseren kenmerken kan vastgelegd worden hoe deze besluiten tot stand zijn gekomen. Bij het implementeren van het model in bredere systemen moet gezorgd worden voor voldoende inzicht in de logica van de besluitvorming zodat het subject inzicht krijgt in welke gegevens volgens welke logica tot welke kredietbeoordeling leiden.

Kredietaanvrager onder druk meer persoonlijke gegevens te delen

Het gebruik van meer gegevens leidt welhaast onvermijdelijk tot grotere risico's op het gebied van gegevensbescherming en privacy. Het gebruik van kenmerken waarvan het subject niet bewust is dat deze gebruikt worden, of hoe deze verhouden tot zijn of haar kredietwaardigheid, hebben impact op privacy. In de interviews is naar voren gekomen dat, logischerwijs, het algoritmische model met meer data over de kredietaanvrager een preciezere kredietbeoordeling kan uitvoeren. Kredietaanvragers kunnen zelfs gevraagd worden om meer te delen over zichzelf en op basis daarvan een nieuwe beoordeling te laten uitvoeren. Hierdoor komt de kredietaanvrager onder druk te staan, omdat het voor de kredietaanvrager van groot

[74] Hijink, M. (2018).'Hoe wordt je kredietscore berekend?', Geraadpleegd via: https://www.nrc.nl/nieuws/2018/12/27/hoe-wordt-je-kredietscore-berekend-a3127138.

[75] Het algoritme van de FICO-score, een traditionele kredietbeoordelingsstandaard in de V.S. die daar voor 90% van kredietbesluiten gebruikt wordt is een voorbeeld van 'black box', alleen de categorieën gegevens zijn bekend.

belang is dat de kredietaanvraag wordt goedgekeurd. Dit leidt er mogelijk toe dat de kredietaanvrager het gevoel krijgt geen andere keuze te hebben dan meer gegevens vrij te geven bij een negatieve kredietbeoordeling.

Een inbreuk op privacy is niet te vermijden bij het doen van een kredietbeoordeling. Er wordt immers getracht iets over de kredietaanvrager te weten te komen. Onevenredige inbreuken op privacy kunnen voorkomen worden door bij het conceptualiseren van de probleemruimte en de daaropvolgende dataverzameling na te gaan welke gegevens noodzakelijk zijn voor de kredietbeoordeling en hoe dit de kredietaanvrager raakt. Omdat meer gegevens waarschijnlijk tot een beter beeld van de kredietaanvrager leiden, dient bepaald te worden wanneer een model goed genoeg is en hoe omgegaan wordt met bijvoorbeeld *proxy* kenmerken.[76] Daarnaast moet de kredietaanvrager helder en uitgebreid geïnformeerd worden over de privacyaspecten van de kredietbeoordeling.

Vooringenomenheid en discriminatie

Een algoritmisch model kan bewust of onbewust vooringenomenheid (*bias*) vertonen bij het geven van kredietscores aan kredietaanvragers, bijvoorbeeld op basis van nationaliteit of etniciteit, hetgeen kan leiden tot discriminatie/ongelijke behandeling. Vooringenomenheid kan ontstaan door afhankelijkheid van gevoelige features, maar ook door proxies voor deze features (kenmerken die gebruikt worden in plaats van correlerende onzichtbare of onmeetbare kenmerken). Zo kan bijvoorbeeld een postcode of muzieksmaak als een proxy gebruikt worden voor de etniciteit van een persoon.[77]

De essentie is dat de voorspellende waarde van bepaalde kenmerken, bijvoorbeeld een postcode, gemiddeld groot genoeg is om in het model op te nemen. Dit betekent echter niet dat voor een individuele kredietaanvrager een postcode altijd een redelijke voorspeller is van kredietwaardigheid. Op individueel niveau lijkt er immers geen verband te zijn tussen postcode en kredietwaardigheid (verandert je kredietwaardigheid als je in een ander postcodegebied gaat wonen?).

Het gebruik van deze proxy kenmerken levert dus potentieel onvoorziene effecten op. Bij het gebruik van een postcode is de aanname dat kredietaanvragers met dezelfde postcode genoeg op elkaar lijken dat de kredietwaardigheid van de ene kredietaanvrager ook iets zegt

[76] Een proxy kenmerk is een 'neutraal' kenmerk dat sterk correleert met een gevoelig kenmerk dat niet gebruikt mag worden of verborgen wordt gehouden (denk bijvoorbeeld aan bijzondere persoonsgegevens zoals etniciteit of religie). Zo kan muzieksmaak bijvoorbeeld een proxy zijn voor iemands etniciteit, of dieetwensen voor iemands religie. Zie ook de volgende paragraaf.

[77] Zie bijvoorbeeld: Marshall, S. R., Naumann, L. R. (2018), What's Your Favorite Music? Music Preferences Cue Racial Identity, in: Journal of Research in Personality, 76 (2018) 74-91

over de kredietwaardigheid van de ander. Omdat een kenmerk, bijvoorbeeld een postcode, ook kan correleren met allerlei andere kenmerken van subjecten kunnen besluiten (schijnbaar) gebaseerd worden op bevooroordeelde of discriminerende kenmerken.

Hetzelfde geldt voor individuele kenmerken die geen relatie met kredietwaardigheid lijken te hebben. Er bestaan organisaties die op basis van locatiedata, social-mediagebruik, toetsaanslagen, het apparaat waarmee je een website bezoekt enzovoorts, kredietwaardigheid voorspellen. Naast de vraag of deze kenmerken überhaupt een causale relatie hebben met kredietwaardigheid en het gebruik van deze kenmerken wenselijk is, levert het onzorgvuldig gebruik van dit soort gegevens ook risico's op voor (onbewuste) discriminatie en vooringenomenheid.

Het is aan de ontwikkelaar en de gebruiker om bij het formuleren van de probleemruimte na te gaan of de gebruikte kenmerken geschikt zijn om besluiten op te baseren.

Gedifferentieerde toegang tot krediet

Het doel van algoritmen bij kredietbeoordelingen is om 'betrouwbare' kredietaanvragers wel krediet te verlenen en 'onbetrouwbare' aanvragers niet. De toevoeging van complexere modellen aan de besluitvorming heeft een impact op deze bestaande differentiatie. Als besluiten meer gebaseerd worden op individuele kenmerken zal de door de kredietgever gewenste differentiatie van kredietaanvragers beter lukken. De kredietgever bepaalt dus of de toegang tot haar diensten/producten gelijker wordt (of dat juist 'kostbare' subjecten buiten de deur worden gehouden). Toezicht en monitoring van de uitkomsten van het algoritme kunnen helpen bij het nagaan of de bedoelde differentiatie behaald wordt of dat er onvoorziene uitkomsten zijn.

Verlies begrip kredietbeoordeling

Dit effect houdt in dat kredietaanvragers niet weten hoe het resultaat of de beslissing tot stand is gekomen. Dat leidt ertoe dat deze aanvragers niet kunnen nagaan of desbetreffende beoordeling correct is. Dit effect heeft zowel invloed op de ontwikkelaar van het algoritme, de gebruiker die het model laat inzetten (de kredietgever) en de kredietaanvrager.

Voor de ontwikkelaar is dit effect van belang, omdat de ontwikkelaar de organisatie is die moet uitleggen aan de kredietgever hoe het model werkt en hoe de kredietbeoordeling tot stand is gekomen. Wanneer dit niet aan de gebruiker kan worden uitgelegd, dan neemt het vertrouwen in het model af en gaat de gebruiker een andere ontwikkelaar of een alternatieve en een betrouwbaardere oplossing zoeken.

Voor de kredietgever is dit effect van belang, omdat zij het model inzet of laat inzetten in haar systemen tegenover de kredietaanvragers. Als een kredietaanvrager klaagt over een kredietbeoordeling of wanneer een toezichthoudende organisatie wil weten hoe een bepaalde beoordeling tot stand is gekomen, dan moet een heldere en volledige uitleg gegeven kunnen worden. Wanneer dat niet mogelijk is, dan zorgt dit voor problemen voor de kredietgever (bijvoorbeeld reputatieschade en boetes). Ook kan de kredietgever moeilijk controleren of de beslissingen wel kloppen. Worden niet goede potentiële klanten geweerd door verkeerde beslissingen van het algoritme?

Voor de kredietaanvragers is dit effect van belang, omdat zij willen weten hoe een kredietbeoordeling tot stand is gekomen. Bij een negatieve kredietbeoordeling willen kredietaanvragers weten hoe deze beoordeling tot stand is gekomen, zodat bekend is op basis van welke aspecten negatief is geoordeeld. Met die informatie kunnen kredietaanvragers voor zichzelf beoordelen of zij desbetreffende aspecten kunnen veranderen, zodat zij wel een positieve kredietbeoordeling kunnen krijgen bij een volgende kredietaanvraag.

Onjuiste kredietbeoordelingen

De uitkomsten van kredietbeoordelingen geven een risico-indicatie. Dit is een voorspelling met een zekerheid die voor de kredietgever aanvaardbaar is. Of een kredietbeoordeling correct is geweest, is pas bij terugbetaling of bij betaalproblemen vast te stellen. Daar komt bij dat het voor de kredietgever belangrijk is dat *alle* kredietbeoordelingen *zo goed mogelijk* zijn. Vanuit het perspectief van de kredietaanvrager is het juist een *individueel* besluit dat impact heeft. Modellen die goed genoeg zijn om ingezet te worden door kredietgever, kunnen dus voor sommige individuen toch een verkeerde uitkomst hebben.

Uit de interviews blijkt dat een belangrijk probleem bij de inzet van algoritmische modellen is dat de data maanden of jaren oud kunnen zijn. Het ontwikkelproces van een model duurt namelijk lang en op het moment dat het model wordt geïmplementeerd zijn de data deels verouderd. Dat is onder normale omstandigheden geen groot probleem, vooral niet wanneer goed toezicht wordt gehouden op het model. Echter, wanneer situaties zich voordoen die extreem en onvoorzien zijn, zoals de crisis rondom Covid-19, dan zijn de resultaten die uit de modellen komen mogelijk ook extreem en onvoorzien. In het geval van kredietbeoordelingen leidt dit ertoe dat in dergelijke crises vrijwel geen kredietaanvragen positief worden beoordeeld door de kredietbeoordelaars. Met die negatieve beoordelingen is het lastig voor kredietgevers om de kredietaanvraag toch positief te beoordelen; de kredietgever zal over het algemeen het advies van het algoritme volgen.

Om de nauwkeurigheid van het model zo hoog mogelijk te hebben en te houden én om in te kunnen ingrijpen wanneer de nauwkeurigheid afneemt, moet constant toezicht plaatsvinden op het model.

> *"You need to have constant monitoring of the models. When you monitor your model closely, you can detect when things go wrong. […] Models that you forget about will deteriorate quite quickly."*
>
> Wereldwijde aanbieder van (krediet) informatiediensten en producten

8 Case study 3: HR Analytics

Het is voor bedrijven van essentieel belang om goede werknemers in dienst te hebben. Een werknemer die niet goed bij de organisatie past zorgt voor minder omzet of neemt na korte tijd afscheid van de organisatie.[78] Het is daarom niet verrassend dat uit onderzoek van de Boston Consulting Group blijkt dat de *recruitment* rollen binnen HR de grootste invloed hebben op de omzetgroei van de organisatie.[79] Daar komt bij dat het een grote uitdaging is om goede medewerkers te vinden.[80]

Om deze uitdaging beter aan te gaan, worden steeds vaker algoritmische modellen ingezet in het *Human Resource* domein. Deze toepassingen worden aangeduid met de brede term *HR analytics*. In het algemeen dienen HR analytics ertoe om personeelsbeleid te automatiseren, nieuwe inzichten te geven en medewerkers en teams te ondersteunen om beter te presteren.[81] Dit kan bijvoorbeeld worden gedaan door de samenstelling van teams te analyseren en suggesties voor verbeteringen te doen. Traditioneel worden dergelijke analyses door hoger management gedaan. De beschikbaarheid van meer data over medewerkers en operaties, en de capaciteit om deze (automatisch) te analyseren zorgt ervoor dat er meer en sneller kennis over de organisatie ontsloten kan worden.

Een van de meest in het oog springende toepassingen is de inzet van modellen om beter inzicht in de kandidaten voor een functie te krijgen.[82] De belofte van dergelijke toepassingen is om snel en objectief werknemers te kunnen beoordelen zonder dat de vooroordelen of persoonlijke voorkeuren een rol spelen.[83] Deze toepassing van modellen voor de beoordeling van nieuwe en bestaande bedewerkers is een specifieke toepassing van HR analytics waar we deze case mee illustreren.

HR analytics zijn om verschillende redenen relevant. Door het automatiseren van delen van het recruitment proces worden besluiten ondersteund of genomen over sollicitanten en de samenstelling van teams. Dit zijn besluiten die een potentieel grote impact kunnen hebben op

[78] Pessach et al. (2020). Employees Recruitment: A prescriptive analytics approach via machine learning and mathematical programming. Geraadpleegd via: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7252110/.

[79] Strack, R. et al. (2012). From Capability to Profitability: Realizing the Value of People Management. BCG. Geraadpleegd via: https://image-src.bcg.com/Images/BCG_From_Capability_to_Profitability_Jul_2012_tcm9-103684.pdf

[80] Pessach et al. (2020). Employees Recruitment: A prescriptive analytics approach via machine learning and mathematical programming. Geraadpleegd via: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7252110/.

[81] Paragraaf 8.1 gaat in op de doelen van HR analytics

[82] Sennaar, K. (2019). Machine Learning for Recruiting and Hiring – 6 current applications. Emerj. Geraadpleegd via: https://emerj.com/ai-sector-overviews/machine-learning-for-recruiting-and-hiring/

[83] Leicht-Deobald et al. (2019). The challenges of Algorithm-Based HR Decision-Making for Personal Integrity. Journal of Business Ethics 160:377-392.

subjecten en de organisatie in zowel positieve als negatieve zin. Omdat er bij recruitment een keuze wordt gemaakt tussen mensen is er een groot risico dat vooroordelen en discriminatie een rol spelen. Dat is het geval bij traditionele werving en selectie en mogelijk ook als algoritmen een deel van de besluitvorming informeren of overnemen. Aan de andere kant is de belofte van algoritmen voor recruitment dat op data gebaseerde besluiten minder vooringenomen zijn.

8.1 Doel HR analytics

HR analytics in bredere zin zijn toepassingen rond inzichten uit data gericht op de medewerkers van een organisatie. Bijvoorbeeld het monitoren van tevredenheid, opmerken van ongewenst gedrag, ondersteunen van talentontwikkeling, enzovoorts. HR analytics voor recruitment worden toegepast om op basis van publiek beschikbare data (bijvoorbeeld LinkedIn), kenmerken die door de kandidaat worden aangeleverd (bijvoorbeeld CV, assessments) en data beschikbaar binnen het eigen bedrijf, besluiten te nemen over potentiële kandidaten.

De belofte van een meer data-gedreven personeelsbeleid is efficiëntie, verbetering van werkomstandigheden en een eerlijkere meer diverse werkvloer.[84] Bijvoorbeeld bij inzet van algoritmen voor recruitment is het de bedoeling dat iedere kandidaat op een consistente wijze wordt beoordeeld door een model, in plaats van door mensen die vooringenomen en inconsistent kunnen zijn in hun beoordeling. Hierdoor hebben subjectiviteit en vooroordelen minder impact.[85]

8.2 Implementatie en inzet HR analytics

De implementatie van HR analytics hangt af van het doel van de toepassing, de organisatie en de toepassing zelf. Naast de automatisering van HR processen worden HR analytics met name ingezet ter ondersteuning van de werkzaamheden van HR-medewerkers, managers en de werknemer zelf.[86]

[84] Purtill, C. (2020). Algorithms learn our workplace biases. Can they help us unlearn them. New York Times. Geraadpleegd via: https://www.nytimes.com/2020/03/10/us/algorithms-learn-our-workplace-biases-can-they-help-us-unlearn-them.html

[85] PSI Testing Intelligence, (2015). 4 Reasons Why an Automated Hiring Process Will Help Your Company. Geraadpleegd via: https://blog.psionline.com/talent/4-reasons-why-an-automated-hiring-process-will-help-your-company.

[86] De toepassing van algoritmen voor de analyse en selectie van nieuwe medewerkers is daar een voorbeeld van. Deze algoritmen worden primair ingezet omdat gebruikers beter inzicht willen hebben welke sollicitanten het beste passen binnen een organisatie, team of functie.

Welke data worden gebruikt hangt af van de ontwikkelaar van het algoritme en de gebruiker van het algoritme (klant van de ontwikkelaar). Een ontwikkelaar kan, bijvoorbeeld, een advies verlenen puur op basis van op Internet beschikbare data (bijvoorbeeld LinkedIn profielen) of op basis van bijvoorbeeld personeelsdossiers van de organisatie van de gebruiker.

Een voorbeeld van een specifieke toepassing van HR analytics voor *recruitment* en *team performance* wordt aangeboden door het bedrijf Seedlink.[87] Het doel van de toepassing is het analyseren van persoonskenmerken van kandidaten om na te gaan of deze overeenkomen met kenmerken die in een baan tot succes leiden of bij de bedrijfscultuur passen. Seedlink modelleert met deep learning algoritmen de geschreven antwoorden op drie open vragen van medewerkers, hun team feedback op competenties en (jaarlijkse) uitkomsten op KPI's (prestatie). Seedlink kijkt dus naar de onbewuste informatie in taalgebruik die inzicht geeft in persoonlijkheidskenmerken in plaats van bijvoorbeeld kernwoorden in een CV, werk/studie achtergrond of gender.

Het taalgebruik en de prestaties van de huidige medewerkers van de klant worden door het model vergeleken met het taalgebruik van de sollicitant. Zo kan voor de sollicitant voorspeld worden hoe deze scoort op verschillende competenties. Vervolgens adviseert het model aan de eindgebruiker welke sollicitant het beste aansluit bij de vacature. Aan de sollicitant kan worden meegedeeld waarom ze wel of niet als een *match* gezien worden. Zo ontstaat een steeds verfijnder model waarin taalgebruik geassocieerd wordt met bepaalde eigenschappen. Na een keuze voor een kandidaat wordt geanalyseerd of de door het systeem voorspelde geschiktheid ook in de praktijk blijkt, waar het systeem weer van kan leren.

Om tot een succesvolle implementatie van deze toepassing te komen is er nauw contact tussen de gebruiker en de ontwikkelaar, bijvoorbeeld een adviseur van de ontwikkelaar die aan de gebruiker is gekoppeld. Voor een succesvolle samenwerking moet het, bijvoorbeeld, duidelijk zijn dat de klant begrijpt wat het model gaat doen en wat voor data nodig is voor een succesvol advies. Uit de interviews blijkt dat de ontwikkelaar en de gebruiker samen de stappen van data verzamelen, modelontwikkeling en validatie doorlopen om zo een goed begrip van de werking en impact van de uitkomsten van het algoritme te krijgen.

[87] Zie ook de website van Seedlink via: https://www.seedlinktech.com/,

8.3 Impact van HR analytics op recruitment en employee performance

De inzet van HR analytics heeft een aantal effecten op de wijze waarop *recruitment* en *employee performance* metingen worden verricht.

Automatisering

De inzet van HR analytics voor recruitment zorgt ervoor dat een deel van het werk dat voorheen door de eindgebruiker werd gedaan nu geautomatiseerd wordt uitgevoerd. Het gevolg hiervan is dat het werk van de eindgebruiker verandert; van beoordelen van een (potentiële) medewerker naar het interpreteren en beoordelen van de resultaten uit het model. Om deze nieuwe werkzaamheden goed uit te voeren moeten desbetreffende mensen worden getraind het model goed in te zetten en de resultaten te kunnen duiden.

Naast bedrijfseconomische voordelen (efficiëntie, lagere kosten) wordt algoritmische ondersteuning bij recruitment ook genoemd als manier om vooroordelen en discriminatie te voorkomen. Doordat de menselijke factor (voor een deel) weg te nemen uit het screeningproces is de belofte dat besluiten eerlijker worden genomen.

Maar vergaande automatisering van voor het subject belangrijke beslissingen heeft mogelijk ook een impact op ervaren 'menselijkheid' van het proces. Uit interne documenten van Amazon blijkt dat interne systemen in de Verenigde Staten de productiviteit van medewerkers nauwkeurig bijhouden en automatisch waarschuwingen geven en zelfs medewerkers ontslaan. Dergelijke automatisering van besluitvorming is voor de werkgever waarschijnlijk positief (want zeer efficiënt), werknemers zullen dit waarschijnlijk anders beoordelen. Dit effect hangt sterk af van het doel waarvoor de toepassing ingezet wordt. In het voorbeeld van Seedlink zorgt automatisering ervoor dat sollicitanten beter geïnformeerd worden waarom ze afgewezen zijn. Met een persoonlijke rapportage krijgen sollicitanten inzicht waarom ze afgewezen zijn en welke functies wel bij hun profiel passen. In traditionele recruitment procedures is er vaak weinig inzicht voor sollicitanten over de redenen waarom zij afgewezen zijn. De automatisering die plaatsvindt met de inzet van algoritmen in dit voorbeeld kan dus ook zorgen voor een meer persoonlijke behandeling.

Datagedreven advisering

Toepassingen op het gebied van HR analytics kunnen zorgen voor data-gedreven menselijke besluitvorming of zelfs volledige automatisering van besluitvorming. Traditioneel vindt besluitvorming in HR plaats op basis van de ervaring en de beoordeling van professionals. De toepassing van HR analytics zorgt ervoor dat HR-professionals nieuwe informatie kunnen

gebruiken in hun werk. Daarmee ontstaat een afhankelijkheid tussen de gebruiker en het model. Leicht-Deobald et al. beschrijven dat als volgt:

> *"Algorithm-based HR decision-making can shift the delicate balance between employees' personal integrity and compliance more toward the compliance side because it may evoke blind trust in processes and rules, which may ultimately marginalize human sense-making as part of the decision-making process."*[88]

Het hangt dus af van de specifieke implementatie, en gekozen waarborgen of de toepassing zorgt voor positieve of negatieve effecten.

Snelheid

Traditioneel kosten *recruitment* procedures veelt tijd. Er moeten bijvoorbeeld veel CV's en motivatiebrieven worden gelezen en veel gesprekken gevoerd. Ook andere HR-processen kunnen tijdrovend zijn: er moet veel bedrijfs- en/of personeelsdata gefilterd, geanalyseerd en geïnterpreteerd worden en daaruit beleidsconclusies worden getrokken. De inzet van HR analytics zorgt ervoor dat deze processen sneller en vaker uitgevoerd kunnen worden. Hierdoor is er meer informatie beschikbaar waarmee gestuurd kan worden. Het effect hiervan is moeilijk te voorspellen en hangt af van de toepassing.

8.4 Potentiële onvoorziene effecten van HR analytics

Bij de toepassing van algoritmen in de context van HR analytics kunnen onvoorziene effecten ontstaan die de ontwikkelaar en de gebruiker niet hebben voorzien of bedoeld. Bovendien kunnen deze effecten ongewenst zijn. In deze paragraaf beschrijven we welke onvoorziene effecten op basis van literatuuronderzoek en de interviews naar voren zijn gekomen. Uit de interviews blijkt dat de precieze inbedding en gebruik van de toepassingen varieert tussen organisaties, en dus mogelijke onvoorziene effecten ook. We noemen hier in deze paragraaf de meest zichtbare effecten.

Machtsverschil tussen werknemer en werkgever

Bij een sollicitatieprocedure is reeds sprake van een machtsverschil tussen sollicitant (het subject) en potentiële werkgever (de gebruiker). De sollicitant wil graag een vacature vervullen

[88] Leicht-Deobald et al. (2019). The challenges of Algorithm-Based HR Decision-Making for Personal Integrity. Journal of Business Ethics 160:377-392.

en de gebruiker bepaalt of dat gebeurt.[89] Door (een deel) van de sollicitatieprocedure te baseren op andere kenmerken dan die de kandidaat expliciet kiest om over te brengen heeft zij minder controle over de gegevens die de werkgever gebruikt om een besluit over haar te nemen. Dit is ook het geval voor de eindgebruiker die voorheen zelf de sollicitant analyseerde. Een deel van de sturing in het sollicitatieproces komt bij een algoritme te liggen, zoals al eerdergenoemd hangt het van de implementatie af wat dit betekent voor de 'eerlijkheid' van het proces.

Hetzelfde geldt voor het inzetten van HR analytics om de prestaties van medewerkers te meten. Tussen werknemer en werkgever is namelijk een inherent machtsverschil. Dit machtsverschil worden vergroot met de inzet van HR analytics, omdat de organisatie meer over een medewerker te weten komt. Er kan ook voor worden gekozen om de inzichten juist (alleen) beschikbaar te stellen aan de medewerker wanneer zijn/haar presteren door een systeem wordt gemeten. Hierdoor kan een scheefgroei in de machtsverhoudingen (deels) worden voorkomen of zelfs teruggedrongen.[90]

> *HR-afdelingen stelden traditioneel vooral data beschikbaar als stuurinformatie voor management. Wij zijn van dat idee af. De mindshift is: data inzetten om de medewerker informatie te geven zodat hij of zij betere beslissingen kan nemen.*
>
> Tertia Wiedenhof, Product Owner People Analytics & Insights Rabobank

Een ander mogelijk onvoorzien effect dat in deze categorie speelt is dat de gevoelsafstand tussen de medewerker en de werkgever wordt vergroot, omdat een deel van het menselijke aspect van de beoordeling en de meting hiermee verdwijnt.

Bias

De belofte van HR analytics gericht op recruitment, is de belofte dat het model helpt om vooroordelen en discriminatie in selectieprocedures weg te nemen. In de praktijk blijkt dat deze

[89] Het is mogelijk dat de werkgever een persoon heeft gevraagd voor een rol, waardoor het machtsverschil meer naar een gelijk niveau komt tussen deze twee personen. In deze situatie gaan we uit van een sollicitant die zelf solliciteert naar een rol bij een organisatie.

[90] Hierbij zijn de doelen van de gebruiker van doorslaggevende betekenis: wanneer de inzichten worden gebruikt voor ontwikkeling en empowerment van medewerkers zijn de effecten positief, wanneer de inzichten worden gebruikt om de medewerker te 'nudgen' in een richting die het bedrijf graag ziet zijn de effecten waarschijnlijk eerder negatief voor de medewerker.

belofte niet altijd wordt waargemaakt.[91] Algoritmen leren van bestaande voorbeelden, waardoor bestaande vooroordelen kunnen worden gerepliceerd. Afhankelijk van de toepassing en implementatie kan deze vooringenomenheid de volgende oorzaken hebben:

- het doel waarvoor de toepassing wordt ingezet is bevooroordeeld;
- de bestaande omgeving is bevooroordeeld en het model repliceert deze werkelijkheid;
- de data waarop modellen worden gebaseerd zijn vooringenomen (bijvoorbeeld omdat een bepaalde groep ontbreekt);
- de implementatie beperkt de werking van het model op verschillende manieren voor verschillende subjecten;
- de eindgebruiker gebruikt de uitkomsten van het model op een vooringenomen manier.

Een voorbeeld dat dit illustreert is een algoritme van Amazon dat als doel had om op basis van CV's van kandidaten de beste te selecteren voor een bepaalde functie. Het onvoorziene effect van het gebruik van het algoritme was dat de toepassing systematisch mannen als geschikter beoordeelde dan vrouwen.[92] De oorzaken van dit effect lijken te liggen in een combinatie van factoren. Hoewel het geslacht van kandidaten niet expliciet werd meegenomen, gebruikte het model toch proxies van dit kenmerk (zoals hobby's). In combinatie met de oververtegenwoordiging van mannen in de organisatie selecteerde het model CV's van meer mannen dan verwacht. Ook lijkt er bij het bepalen van de doelen van het algoritme niet expliciet te zijn bepaald dat man-vrouw verhouding een belangrijke factor was voor de uitkomsten.[93] Uiteindelijk besloot Amazon de toepassing niet meer te gebruiken.

Het is dus van belang om voor elke fase in het proces bewust te zijn van mogelijke vooroordelen, de gewenste uitkomsten en welke methode ingezet wordt. De toepassing van algoritmen kan ook vooringenomenheid in menselijke oordelen transparant maken. Een voorbeeld is de identificatie van vertekeningen in menselijke beoordelingen. Seedlink kwam bij een implementatietraject zoveel afwijkingen tegen in beoordelingen van medewerkers en de data over prestaties, dat de organisatie besloot op een andere (neutralere) manier medewerkers te beoordelen.

[91] Dastin, 'Amazon scraps secret AI recruiting tool that showed bias against women', 10-10-2018, Reuters, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[92] Dastin, 'Amazon scraps secret AI recruiting tool that showed bias against women', 10-10-2018, Reuters, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[93] Amazon's sexist AI recruiting tool: how did it go so wrong? Geraadpleegd op 04-09-2020 via: https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e

> *We besteden veel aandacht aan data verzamelen en de kwaliteit van de labels (input waarop je een model traint). Zo kregen we ooit van een klant de beoordelingen van hun medewerkers en zagen we dat verschillende managers verschillende oordelen hadden. We hebben toen geadviseerd om op een andere manier deze beoordelingen te doen.*
>
> Rina Joosten-Rabou, CEO Seedlink

(on)eerlijke uitkomsten / transparantie besluiten

Een aspect van het doel van algoritmen voor HR-toepassingen is het eerlijker maken van de bestaande situatie. Hiermee ontstaat echter ook de mogelijkheid dat de uitkomsten van algoritmen oneerlijke uitkomsten opleveren. Omdat HR-toepassingen over mensen gaan kan de impact van een dergelijk effect groot zijn. Het is dus van belang dat gebruikers van HR-toepassingen kritisch kijken naar de doelen, werking en effecten van hun algoritmen.

> *Het geweten van HR is belangrijk hierbij. Je kunt dat niet alleen bij de medewerker neerleggen, de medewerker kan namelijk niet alles weten van wat er speelt.*
>
> *Tertia Wiedenhof, Product Owner People Analytics & Insights Rabobank*

Het is van belang dat gebruikers, eindgebruikers en subjecten begrijpen hoe de uitkomsten van een model tot stand zijn gekomen. Uit de interviews blijkt dat alle betrokken partijen het begrip van de werking van het algoritme zeer belangrijk vinden. Naast de gekozen modellen en de inherente (in)transparantie van besluiten die daaruit volgen, is inzicht in de ontwikkeling van de toepassing, communicatie van onderliggende logica, en de redenen voor een specifieke beslissing van groot belang.

Bevriezen werkelijkheid / inclusiviteit en diversiteit

Omdat algoritmen voor HR-toepassingen vaak datagedreven zijn moet worden nagegaan of de huidige situatie een goed uitgangspunt is om besluiten te informeren. Omdat het doel van een recruitment algoritme is om op basis van bepaalde indicatoren (bijvoorbeeld organisatiecultuur) geschikte kandidaten te identificeren is het vaststellen van deze indicatoren bepalend voor de uitkomst. De keuze om de bestaande situatie te repliceren zonder na te gaan wat de wenselijkheid is van de oude situatie is, zorgt voor onvoorziene effecten. Het eerdergenoemde voorbeeld van algoritmen die vooringenomenheid ten opzichte van vrouwen voor bepaalde functies repliceerde laat dit zien. Omdat algoritmen niets anders doen dan

functioneren zoals de ontwikkelaar heeft bepaald, moeten een gewenste werking expliciet worden meegenomen. Als de (onuitgesproken) wens is dat een algoritme een gelijk aandeel mannen en vrouwen als kandidaten voorstelt, dan moet de ontwikkelaar daarvoor zorgen.

Aan de andere kant kunnen door de inzet van algoritmen veel meer gegevens geanalyseerd worden dan voorheen voor een individu of zelfs een organisatie haalbaar was. Zo kunnen inzichten ontstaan die voorheen niet toegankelijk waren. Een voorbeeld hiervan is de identificatie van de vertekening in beoordelingen door sommige managers in een organisatie.

Begrip van de gewenste de situatie of uitkomst is dus niet voldoende om gewenste effecten te bereiken. Begrip van de volledige probleemruimte, (oorzaken van) de bestaande situatie, latente verwachtingen en aannames dragen bij aan de succesvolle toepassing van HR analytics.

9 Case study 4: Fysieke & geestelijke gezondheid

Fysieke en mentale gezondheid is van groot belang voor de levenskwaliteit van individuen. Een aantal (wereldwijde) trends tonen aan dat op deze gebieden nog veel te winnen is. Het aantal mensen met overgewicht neemt toe.[94] Wereldwijd lijden honderden miljoenen mensen aan depressie.[95] Fysieke en mentale gezondheid staan ook onder druk de huidige Covid-19 pandemie.[96,97]

Ontwikkelingen in slimme sensoren (*wearables*), beschikbaarheid van meer data en mogelijkheden deze te analyseren maken nieuwe toepassingen mogelijk om geestelijke en fysieke gezondheid te bevorderen. Er lijkt ook een groeiende wil van consumenten te ontstaan om gezondheidsgerelateerde gegevens van zichzelf te willen meten. Deze laatste trend wordt ook wel de *quantified self* trend genoemd.

Zowel mobiele telefoons als *wearables* hebben allerlei sensoren die iets (kunnen) zeggen over de fysieke en/of mentale gesteldheid van het individu. De Apple Watch bijvoorbeeld bevat een stappenteller, hartslagmeter, geluidsensor, GPS en allerlei andere sensoren die gebruikt kunnen worden om gegevens gerelateerd aan de gezondheid te meten. Met deze gegevens kan een applicatie in het horloge zelf of met de gekoppelde telefoon inzicht geven in de fysieke en mentale gesteldheid.

Waar traditioneel de fysieke en mentale problemen met (medische) professionals werden aangepakt maken deze nieuwe toepassingen een nieuwe manieren van diagnosticeren en behandelen mogelijk. Deze nieuwe mogelijkheden kunnen zorgen voor laagdrempeligere toegang tot hulp en goedkopere en effectievere behandelingen.[98] Tegelijkertijd is de kwaliteit van de toepassing en de menselijkheid van de toepassing van groot belang. Onvoorziene effecten van toepassingen in het domein van de fysieke en mentale gezondheid hebben mogelijk vergaande consequenties.

In deze case study richten we ons op (niet-medische) toepassingen van algoritmen voor ondersteuning op het gebied van fysieke en mentale gezondheid met wearables. De relevantie

[94] Sung, H. et al. (2019). Global patterns in excess body weight and the associated cancer burden. CA: a cancer journal for clinicians, 69(2), 88-112.

[95] World Health Organization (2020). Depression. Geraadpleegd via: https://www.who.int/news-room/fact-sheets/detail/depression.

[96] Wan, 'The coronavirus pandemic is pushing America into a mental health crisis', The Washington Post, 04-05-2020, https://www.washingtonpost.com/health/2020/05/04/mental-health-coronavirus/.

[97] Schraer, R. (2020). Depression doubles during coronavirus pandemic. BBC. Geraadpleegd via: https://www.bbc.com/news/health-53820425.

van deze case volgt uit de grote rol van het subject bij de inzet en effectiviteit van de toepassing in combinatie met het domein van de toepassing.

9.1 Doel AI voor fysieke & geestelijke gezondheid

Voordat *wearables* en daarvoor ontwikkelde algoritmen op de markt waren, was het verzamelen en analyseren van gezondheidsdata het domein van medische professionals. Hartslag-, bloedsuiker-, thermometers bestonden wel als losstaande (digitale) apparaten maar boden weinig mogelijkheden tot interpretatie van de gemeten gegevens voor subjecten. Door gegevens te analyseren en inzichten aan te verbinden kunnen subjecten meer begrijpen over hun mentale en fysieke gezondheid. Denk bijvoorbeeld aan toepassingen gericht op het inschatten van de vruchtbaarheid van vrouwen. Deze toepassingen meten indicatoren van vruchtbaarheid en maken deze inzichtelijk.[99] Op het gebied van geestelijke gezondheid zijn diverse soorten toepassingen beschikbaar, bijvoorbeeld chatbots gericht op vermindering van angstige gedachten, meditatie, of het bijhouden van emoties.[100] Amazon kondigde recent een wearable aan die naast gangbare functies van fitness trackers ook de toon van je stem analyseert en op basis van toon en 'energie' een logboek van emoties maakt.[101] De toepassing van Mindstrong meet 'passief' smartphonegebruik van de smartphone: typen, *scrollen* en *swipen*. Op basis van deze data signaleert het algoritme of je stress of andere mentale problemen hebt.[102] De toepassing van Wysa wordt ingezet als chatbot, om op basis van tekst te signaleren met wat voor problemen het individu zit en om de persoon te adviseren om professionele hulp te zoeken.[103] De toepassingen lopen dus sterk uiteen en maken gebruik van gegevens gemeten door de sensoren van *wearables* en smartphones, maar ook van door het subject verstrekte informatie.

Wat deze case onderscheid van de vorige cases, is dat het subject ook de eindgebruiker is van het algoritme. Niet alleen wordt het subject onderworpen aan de beslissingen van het algoritme, het speelt ook een actieve rol bij het aanleveren van de relevante informatie en het opvolgen van de adviezen. De ontwikkelaar en de gebruiker hebben invloed op de inzet door middel van bijvoorbeeld uitleg over de applicatie en voorlichting over risico's en het betrekken

[99] Welke indicatoren verschilt sterk tussen de toepassingen: temperatuur, hormonen in urine, hartslag

[100] Bijvoorbeeld Woebot: https://woebothealth.com of Mindwell: https://www.mindwell.live

[101] The Verge (2020). Amazon announces halo, a fitness band and app that scans your body and voice https://www.theverge.com/2020/8/27/21402493/amazon-halo-band-health-fitness-body-scan-tone-emotion-activity-sleep

[102] Mindstrong Health (s.a.), How it works. Geraadpleegd via: https://mindstrong.com/how-it-works/.

[103] Wysa (s.a.). Question most asked by users. Geraadpleegd via: https://www.wysa.io/faq.

van experts/coaches. Uiteindelijk bepaalt het subject of het de applicatie gebruikt en zo ja, voor welke doelen.

9.2 Implementatie en inzet

De inzet van toepassingen voor fysieke en geestelijke gezondheid verschilt dus van voorgaande casestudies omdat het individu zowel eindgebruiker als subject is. Het individu past het algoritme toe op zichzelf, in plaats van bijvoorbeeld een huisarts, personal coach, diëtist of psychiater. Waar bij de andere cases de relatie tussen de ontwikkelaar en de (eind)gebruiker zeer nauw was, is dat hier minder het geval. Hierdoor hebben de ontwikkelaar en gebruiker van de toepassing veel minder inzicht in de effecten van hun toepassing.

Hoe toepassingen voor fysieke en geestelijke gezondheid gebruikt worden verschilt per toepassing. Sommige toepassingen hoeven alleen geïnstalleerd te worden op een smartphone, anderen zijn onderdeel van een breder programma waar subjecten instructie en begeleiding krijgen.

Een voorbeeld van een bredere toepassing is Clear. Een Nederlandse startup gericht op gepersonaliseerd voedingsadvies. Het programma gebruikt een wearable van een derde partij. Dit apparaat meet de bloedsuikerspiegel. Subjecten houden in een applicatie bij wat ze eten, wat ze doen, en hoe ze zich voelen. Door een aantal weken data te verzamelen met de wearable en de app, krijgen de subjecten inzicht in de effecten van etenswaren op bloedsuiker. Zo kan een dieet worden ontwikkeld gebaseerd op de individuele biologie van een subject.

Om het programma succesvol te laten verlopen zijn er verschillende startbijeenkomsten (kick-off meetings en onboarding) waarin gebruikers instructies krijgen hoe ze de toepassing goed kunnen gebruiken en hoe met de resultaten moet worden omgegaan.

> *In de kick-off leggen we de deelnemers uit hoe je biologie werkt, bijvoorbeeld wat de rol insuline is […] Zodat mensen begrijpen wat ze zien en dat ze snappen hoe ze de resultaten moeten interpreteren.*
>
> Piet Hein van Dam, CEO Clear.

9.3 Impact van algoritmen op gezondheidsadvies

De inzet van algoritmen om fysieke en mentale gezondheidsproblemen te signaleren en te verhelpen heeft een aantal effecten op de manier waarop subjecten met hun gezondheid omgaan.

Zelfredzaamheid/veranderende relatie met behandelaars

De grootste impact van de algoritmen in deze casus is dat het subject toepassingen tot haar beschikking heeft die voorheen alleen indirect (via een arts, diëtist, coach) beschikbaar waren. Wat het gevolg hiervan is, is dat door de diversiteit aan toepassingen, subjecten, en aanbiedingsvormen moeilijk generiek te bepalen. Dat toepassingen direct beschikbaar zijn voor individuen kan leiden tot laagdrempeligere toegang tot hulp en zorg. De stap naar een behandelaar kan groot, spannend of duur gevonden worden, maar installeren van een app niet. Een gerelateerde impact is dat individuen mogelijk minder aanspraak maken op behandelaars, dat is kwalijk als daardoor minder goede zorg wordt genoten, maar kan juist ook heel goed zijn als daarmee onnodige aanspraken op schaarse zorgcapaciteit worden voorkomen.

Ook kunnen algoritmen in een eerder stadium medische problemen herkennen, waardoor op tijd een medisch specialist wordt geraadpleegd. Een voorbeeld van een dergelijke toepassing is de bloedsuikeranalyse van Clear, die het mogelijk maakt om mogelijke pre-diabetes te herkennen bij subjecten:

> *We sturen 5-10% van onze deelnemers door naar de huisarts omdat uit de data blijkt dat ze mogelijk pre-diabetes hebben. Dat zijn mensen die anders veel later in beeld zouden komen.*
>
> *Piet Hein van Dam, CEO Clear.*

Nieuwe inzichten / dataficering

Het doel van veel toepassingen in het gezondheidsdomein is meer inzicht in geestelijke en lichamelijke gezondheid op makkelijke, snelle en relatief goedkope manier. Hoewel *wearables* nog niet alomtegenwoordig zijn, zijn *smartphones* dat wel. Deze telefoons hebben veelal dezelfde sensoren als de *wearables*, waardoor een groot aantal mensen bij dagelijks gebruik van hun telefoon inzicht kan krijgen in allerlei gezondheidsdata.

De impact van deze nieuwe informatie is veelvormig. Subjecten komen meer te weten over zichzelf, daar zitten allerlei voordelen aan zoals meer inzicht, zekerheid en controle. Uit onderzoek blijkt dat er ook (tijdelijke) nadelige effecten kunnen zijn zoals ongemak of angstige gevoelens.[104]

[104] Kaziunas, E., Ackerman, M. S., Lindtner, S., & Lee, J. M. (2017, February). Caring through data: Attending to the social and emotional experiences of health datafication. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 2260-2272).

9.4 Potentiële onvoorziene effecten op fysieke & geestelijke gezondheid

Bij de toepassing van algoritmische modellen voor fysieke en geestelijke gezondheid kunnen allerlei effecten ontstaan die voor één of meerdere actoren onvoorzien zijn.

Autonomie en beïnvloeding

De toepassing van algoritmen door het subject heeft geen eenduidige impact op de autonomie. Het subject kiest er zelf voor om bepaalde analyses te laten uitvoeren die die invloed hebben op hem of haar. Wat die invloed is hangt af van het subject, de manier waarop uitkomsten worden gepresenteerd en de wijze waarop de gebruiker het subject inzicht geeft in de toepassing (de werking, correct gebruik, risico's, et cetera).

Subjecten kunnen een te groot vertrouwen hebben in de uitkomst van een algoritmisch advies. Zeker wanneer niet duidelijk is op basis van welke criteria een advies tot stand is gekomen kan dit tot negatieve effecten leiden wanneer dit advies verkeerd of onvolledig blijkt.

Het is aan ontwikkelaars en gebruikers om de subjecten te informeren en te ondersteunen bij het behoud van persoonlijke autonomie. Clear besteedt bij de startbijeenkomst bijvoorbeeld aandacht aan de juiste interpretatie van de uitkomsten:

> *We leggen aan de deelnemers uit dat ze het advies niet blind moeten volgen. Als je een glutenintolerantie hebt, maar het effect van brood op je bloedsuikerspiegel is goed, dan moet je dat niet gaan eten. Dat bedoelen we met je gezonde verstand: begrijp wat je aan het doen bent.*
>
> *Piet Hein van Dam, CEO Clear*

Om goed de effecten op het individu te voorzien is het van belang dat er begrip is van de impact van een toepassing op subjecten, zowel op de korte als langere termijn. Hiertoe kan gebruik gemaakt worden van algoritme impact assessments, (extern) advies rond ethische aspecten, inzichten (van experts) uit de gedragswetenschap, psychologie en andere relevante expertises. Gebruikersonderzoek is een andere manier om inzicht te krijgen in de impact van een toepassing op subjecten.

Verkeerd (opgevolgd) advies

Het doel van de toepassingen in deze case is het subject inzicht geven en eventueel adviseren. Het uiteindelijke effect van de toepassing is dus afhankelijk van de kwaliteit van het advies en wat het subject hiermee doet.

Het advies van het model kan onjuist zijn om meerdere redenen. Het model zelf kan niet goed functioneren, of de inputdata is verkeerd. Wanneer het subject bijvoorbeeld de verkeerde

voedselhoeveelheid heeft ingevuld in een dieet toepassing, dan zal het model mogelijk een verkeerd advies geven. Ook andere databronnen kunnen onverwachte effecten veroorzaken. Sensoren kunnen niet de juiste data doorgeven door een technische storing, of gegevens over telefoongebruik kunnen onbetrouwbaar zijn als anderen (bijvoorbeeld kinderen) het apparaat in handen krijgen. De ontwikkelaar kan dit effect dus voorkomen door extreme afwijkingen te detecteren (in data en uitkomsten) en het subject te informeren over juist gebruik van de toepassing.

Zoals bij autonomie beschreven is verkeerde omgang van het subject met de uitkomsten van de toepassing op verschillende manier te voorkomen. Het subject heeft natuurlijk zelf ook zelf de verantwoordelijkheid om na te gaan waar een toepassing voor bedoeld is (en waarvoor niet) en hoe deze gebruikt dient te worden.

Registratie van gezondheidsgegevens

De registratie van gezondheidsgegevens heeft impact op privacy en gegevensbescherming. De sensordata, ingevoerde gegevens en de inferenties en adviezen die de toepassing genereren zijn mogelijk bijzonder gevoelig. Verlies van gegevens, toegang door onbevoegden, of onverwacht datagebruik kunnen een grote, langdurige impact hebben op het subject. De gevoeligheid van gegevens is ook niet altijd meteen duidelijk. Een voorbeeld is Strava, een hardlooptracker, die een overzichtskaart gaf van de route die gebruikers renden, op basis van deze gegevens kon de locatie van militaire bases worden afgeleid en de patrouilleroutes van soldaten.[105]

Het is aan de ontwikkelaar/gebruiker om de veilige omgang met (persoons)gegevens te waarborgen en duidelijk te maken aan het subject welke gegevens verwerkt worden voor welke doeleinden. In elke stap van ontwikkeling en inzet van de toepassing moet gegeven de gevoeligheid van de gegevens ook aandacht worden besteed aan privacy en gegevensbescherming (van een data protection impact assessment, privacy by design, tot een helder privacy statement),

Het individu dat de toepassing inzet kan onvoorziene effecten rond privacy en gegevensbescherming voorkomen door zich vooraf goed te informeren hoe en waarom haar gegevens gebruikt worden, en zich bewust te zijn van haar rechten.

[105] BBC (2018) Fitness app Strava lights up staff at military bases. Geraadpleegd op 02/09/2020 via: https://www.bbc.com/news/technology-42853072

10 Analyse

In dit hoofdstuk brengen wij de belangrijkste inzichten uit de casestudies, literatuurstudie en interviews bij elkaar.

10.1 Grondoorzaken onvoorziene effecten

Op basis van de literatuurstudie en de observaties uit de casestudies kunnen we stellen dat er drie 'grondoorzaken' zijn voor het optreden van onvoorziene effecten in de context van de toepassing van (zelf)lerende algoritmen.

1) Er is een onvolledig of verkeerd begrip van de probleemruimte.
2) Het zelflerende algoritme is niet goed toegerust om om te gaan met de complexe omgeving waarbinnen het wordt ingezet.
3) Het zelflerende algoritme wordt niet goed ingepast in een bredere (socio-technische) context.

Uiteraard zijn ook combinaties van deze grondoorzaken denkbaar die leiden tot onvoorziene effecten.

De hierboven genoemde grondoorzaken spelen op de korte, middellange en lange termijn een rol. Mogelijk kan voor de middellange tot lange termijn het vraagstuk van emergent gedrag door de interactie tussen verschillende algoritmen ook als grondoorzaak relevant worden, maar op dit moment lijken er weinig situaties te zijn waar het vraagstuk van emergent gedrag naar voren komt. In de case studies hebben wij in ieder geval geen voorbeelden gevonden waar emergent gedrag een rol speelt.

10.1.1 Ad 1) Onvolledig of verkeerd begrip van de probleemruimte

Een beslismodel neemt beslissingen op basis van het begrip dat het heeft van de probleemruimte. Dit begrip krijgt het op basis van de beschikbare data. Deze data worden geselecteerd en aangeleverd door de mens.

Wanneer de mens bewust, of waarschijnlijker, onbewust ontoereikende of verkeerde data gebruikt voor het modelleren van de werkelijkheid, dan is de kans dat het uiteindelijke beslismodel verkeerde beslissingen neemt aanwezig. Dit kan leiden tot onvoorziene en vaak ongewenste effecten.

Een inmiddels klassiek voorbeeld is discriminatie door algoritmen. Algoritmen gebruiken bijvoorbeeld attributen als etniciteit als belangrijkste voorspellers voor crimineel gedrag. Een

achterliggende verklaring voor deze discriminatie kan *societal bias* zijn.[106] Wanneer door (on)bewuste discriminatie een bepaalde bevolkingsgroep in het verleden scherp onder de loep genomen is, is het logisch dat deze groep oververtegenwoordigd is in de misdaadstatistieken die gebruikt worden als trainingsdata voor een beslismodel voor het voorspellen van criminaliteit. Het algoritme kent deze context evenwel niet en zal etniciteit zien als de belangrijkste voorspeller voor crimineel gedrag: een incorrecte aanname. Een andere verklaring kan zijn dat etniciteit sterk correleert met sociaal-economische of andere omstandigheden die leiden tot criminaliteit. De verkeerde aanname is dat het model goed zijn werk doet: het voorspelt immers (zij het via een *proxy*) crimineel gedrag. Maar een dergelijke conclusie is kortzichtig. Wat er daadwerkelijk aan de hand is, is dat het model de maatschappelijke status quo bevestigt. Een ongewenst effect is daarmee dat het model bestaande maatschappelijke verhoudingen en eventueel racisme verder institutionaliseert. Dit risico is met name in de cases betreffende kredietwaardigheid en HR analytics aan de orde.

De data die worden gegenereerd en vervolgens gebruikt voor het trainen van modellen zijn menselijke constructen en kunnen dus bewust of onbewust allerlei verkeerde aannames bevatten die het gevolg zijn van een verkeerd begrip van de situatie of voortvloeien uit de huidige inrichting van onze maatschappij.

Dit betekent dat er bij de selectie van de data ook impliciet keuzes worden gemaakt. Bijvoorbeeld door te kiezen om een fenomeen te beschrijven met bepaalde data *(fraude voorspellen op basis van gekende eerdere fraude waardoor nieuwe vormen van fraude niet herkend worden)*, of door voor een bepaalde indeling in de data te kiezen waardoor andere opties vervallen *(sekse registreren als alleen mannelijk óf vrouwelijk)*.[107]

Het bovenstaande wil geenszins zeggen dat modellen nooit goed kunnen voorspellen of altijd discrimineren: het kan goed zijn dat een menselijke beslissing vele malen inaccurater of meer discriminerend is. Wat het bovenstaande wel zegt is dat een model nooit neutraal is: het is gebaseerd op data die zijn gegenereerd door mensen en waarin bewust of onbewust al keuzes zijn gemaakt.

Binnen het CRISP-DM model gaat het bij dit probleem met name om de stappen *business understanding* en *data understanding*.

[106] Mitchell, S., Potash, E., Barocas, S., D'Amour, A. (2020), Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions

107 Sedee, M. (2020). Derde Nederlander krijgt 'X' in paspoort. NRC.nl. Geraadpleegd via: https://www.nrc.nl/nieuws/2020/02/28/derde-nederlander-krijgt-x-in-paspoort-a3992149.

10.1.2 Ad 2) Algoritme niet toegerust op de complexiteit van de omgeving

Wanneer er sprake is van een complexe omgeving (veel variabelen met interdependenties, een dynamische omgeving, of beide) en een algoritme niet de capaciteit heeft om goed om te gaan met deze complexe omgeving, dan valt te voorzien dat het algoritme verkeerde beslissingen neemt, omdat het niet adequaat reageert op de omgeving.

Dit probleem speelt in het bijzonder bij gesitueerde intelligenties (algoritmen die een directe interactie hebben met de fysieke wereld). Een treffend voorbeeld is een zelfrijdende auto die in een relatief overzichtelijke omgeving zoals een snelweg goed functioneert, maar in de bebouwde kom ongelukken veroorzaakt, omdat de omgeving complexer is.

Wanneer de omgeving heel dynamisch is, dan zal het model snel verouderen (*model degradation*). Het valt dan te voorzien dat het model verkeerde beslissingen gaat nemen omdat het niet alle relevante informatie meer in ogenschouw kan nemen. Wat de gevolgen zijn van deze verkeerde beslissingen is sterk afhankelijk van de specifieke context waarbinnen het model wordt toegepast.

Wanneer de probleemruimte heel groot is (veel variabelen en afhankelijkheden) en ook nog eens snel verandert, dan is de kans op onvoorziene effecten uiteraard het grootst.

Het model zelf tenslotte heeft ook invloed op de werkelijkheid. Het probleem daarbij is dat de uitkomsten van het model kunnen leiden tot ongewenste *feedback loops*. Zo kan een *predictive policing* model bijvoorbeeld voorspellen in welke buurten criminaliteit plaatsvindt en de politie daarheen sturen. Actuele criminaliteitscijfers (zoals bijvoorbeeld arrestaties) worden vervolgens gebruikt om het model te updaten. Omdat de politie in deze buurten daadwerkelijk criminaliteit heeft gevonden en dit weerspiegeld is in de data, worden deze buurten zwaarder gewogen door het geüpdate model en stuurt het de politie opnieuw die kant op, waardoor er een '*runaway feedback loop*' ontstaat.[108] Omdat de inzet van de politie in andere buurten minder is geweest wordt de criminaliteit in deze buurten waarschijnlijk onderschat, waardoor het model objectief gezien door het initiële succes in de tijd verkeerde voorspellingen gaat doen.

Binnen het CRISP-DM model gaat het bij het adresseren van dit probleem niet alleen om interventies in de stappen *business understanding* en *data understanding*, maar ook in de stappen *modeling*, *evaluation* en *deployment*.

[108] Ensign, D. et al. (2018), Runaway Feedback Loops in Predictive Policing, Proceedings of Machine Learning Research 81:1–12, 2018 Conference on Fairness, Accountability, and Transparency

10.1.3 Ad 3) Toepassing binnen een (socio-technische) context

De laatste grondoorzaak voor mogelijke onvoorziene effecten is wanneer zelflerende algoritmen binnen de context van een organisatie (en breder de maatschappij) worden ingebed.

Wanneer bijvoorbeeld een model getraind, gevalideerd en getest is en goed werkt, dan kan een verkeerde inbedding en koppeling met andere systemen alsnog tot problemen en daarmee onvoorziene effecten leiden. Een voorbeeld hiervan is dat een model dat gegevens ophaalt uit een ander systeem via een API en daarbij de verkeerde gegevens gebruikt, omdat de *API calls* (het ophalen van gegevens) niet goed geconfigureerd zijn.[109] Doordat het model de verkeerde gegevens gevoed krijgt, neemt het verkeerde beslissingen (*garbage in, garbage out*) die kunnen leiden tot onvoorziene effecten. Dit is evenwel niet toe te schrijven aan de specifieke aspecten van het model, maar aan de (verkeerde) implementatie binnen de organisatie.

Een ander voorbeeld is een arts die niet goed getraind is in het gebruiken en interpreteren van adviezen van een algoritme. De arts kan zich bijvoorbeeld te veel verlaten op de uitkomsten van het algoritme en minder op het eigen kritische denkvermogen. Wanneer het model dan een verkeerd of incompleet antwoord geeft wordt dit niet gecorrigeerd of aangevuld door de arts, hetgeen tot onvoorziene effecten kan leiden.

Binnen het CRISP-DM model gaat het bij dit probleem met name om de stappen *business understanding* en *deployment*.

Binnen de case betreffende fysieke en geestelijke gezondheid heeft dit vraagstuk nog een extra dimensie, omdat het daar het subject is dat interacteert met het model van een ontwikkelaar / gebruiker. De ontwikkelaar en de gebruiker hebben hierbij minder grip op de data die het subject gebruikt en de manier waarop hij of zij de uitkomsten van het model interpreteert.

10.2 Ondoorgrondelijkheid van zelflerende algoritmen

De ondoorgrondelijkheid van zelflerende algoritmen (het *black box* probleem) versterkt de kans op onvoorziene effecten en met name op het voortduren daarvan. Ondoorgrondelijkheid van een machine learning model maakt het onmogelijk om te beoordelen wat het model doet en met name of het wat het doet, goed doet.

[109] API staat voor Application Programming Interface

10.2.1 Onjuiste beslissingen

Het voornaamste probleem van *black box* modellen is dat hun werking niet gecontroleerd kan worden. Dit zorgt ervoor dat de kans dat verkeerde beslissingen of inaccurate voorspellingen tijdig herkend worden kleiner wordt. Dit vergroot vervolgens de kans op onvoorziene effecten.

Een klassiek voorbeeld is dat van *de Husky v. Wolf classifier*. Een black box algoritme was getraind om plaatjes van huskies en wolven van elkaar onderscheiden. Omdat het algoritme alleen de uitkomsten gaf (wolf of husky) kon niet worden vastgesteld welke kenmerken (*features)* het model gebruikte om onderscheid te maken. Door toepassing van een lokale interpretatiemethode (zie paragraaf 11.2.3) werd duidelijk dat het belangrijkste onderscheidende kenmerk niet een kenmerk van de dieren zelf was, maar de omgeving waarin zij gefotografeerd waren. Het model classificeerde consequent een dier als wolf wanneer er sneeuw in het plaatje aanwezig was. Dit omdat in de oorspronkelijke dataset de gelabelde plaatjes van wolven grotendeels afbeeldingen waren van wolven in de sneeuw.

Je kan daarmee stellen dat de ondoorgrondelijkheid van zelflerende algoritmen niet zozeer zorgt voor onvoorziene effecten op zichzelf, maar dat dit het zicht op de grondoorzaken voor het optreden van onvoorziene effecten zoals hierboven omschreven ontneemt. Dit betekent dat het idee kan ontstaan bij gebruikers dat hun algoritmen goed functioneren, terwijl er in werkelijkheid verkeerde beslissingen worden genomen of voorspellingen worden gedaan die niet accuraat zijn. Dit vergroot niet alleen de kans op het ontstaan en voortbestaan van onvoorziene effecten, maar ook de impact daarvan.

10.2.2 Vertrouwen in geautomatiseerde besluitvorming

Een bijkomend effect van *black box* modellen is dat het vertrouwen in deze modellen minder kan worden naarmate hun beperkingen en risico's meer algemeen bekend worden. Een gebrek aan vertrouwen in geautomatiseerde besluitvorming kan een negatief effect hebben op de toepassing en adoptie van kunstmatige intelligentie.

In een experiment gerelateerd aan de husky versus wolf classifier werd gevraagd aan een groep van 27 personen werden de resultaten van de husky versus wolf classifier voorgelegd. 10 van de 27 personen vertrouwde de uitkomsten van het algoritme. Toen werd getoond dat sneeuw de belangrijkste feature voor de herkenning was vertrouwde nog maar 3 van de 27 personen de uitkomsten.[110]

[110] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

10.3 Epistemische en normatieve zorgen bij algoritmische besluitvorming

De bovenstaande grondoorzaken leiden tot 'ethische zorgen' over de totstandkoming van de besluiten en voorspellingen van algoritmen en de effecten van algoritmische besluitvorming. Mittelstadt et al. onderscheiden *epistemische zorgen* met betrekking tot algoritmische besluitvorming en *normatieve zorgen*:[111]

Epistemische zorgen	
Niet overtuigend bewijs	Algoritmische systemen geven geen sluitende reden voor een bepaalde uitkomst.
Misplaatst bewijs	Algoritmische systemen baseren zich op een inaccuraat of onvolledig beeld van de werkelijkheid.
Ondoorgrondelijk bewijs	Algoritmische systemen produceren uitkomsten die niet uitgelegd of onderzocht worden omdat het algoritme een 'black-box' is.

Normatieve zorgen	
Oneerlijke uitkomsten	Algoritmische systemen produceren uitkomsten die als oneerlijk of onethisch beschouwd worden.
Transformatieve effecten	Algoritmische systemen veranderen de blik op, en de interactie met de wereld. Dit kan ongewenste en onvoorziene effecten opleveren.

De epistemische zorgen zien meer op de totstandkoming van de besluitvorming en of deze wel 'correct' is, de normatieve zorgen zien meer op daadwerkelijke toepassing en de (onvoorziene) effecten die dat heeft. Hieronder lichten wij de verschillende zorgen nader toe.

Niet overtuigend bewijs

Mittelstadt et al. stellen dat (geavanceerde) algoritmen meestal ingezet worden in gevallen waarbij meer betrouwbare technieken niet beschikbaar of te duur zijn.[112] Bijvoorbeeld, de moderatie van sociale media content is te duur om met menselijke moderators te doen en een algoritme waarin expliciet geprogrammeerd is wat wel en niet mag is (nog) niet beschikbaar.[113]

[111] Mittelstadt, B. D. et al. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).

[112] Mittelstadt, B. D. et al., 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*.

[113] Ibid.

De consequentie van deze strategie is dat bijvoorbeeld op schaal moderatie kan worden gedaan, maar dat niet alle uitkomsten van het algoritme gebaseerd zijn op overtuigend bewijs.[114] Immers, als er een precieze grens getrokken kan worden tussen wat wel en niet overtuigend bewijs is, dan kunnen deze regels expliciet geprogrammeerd worden en is waarschijnlijk geen machine learning nodig.

Misplaatst bewijs

Uitkomsten van algoritmen zijn zo betrouwbaar (en neutraal) als de data waarop ze gebaseerd zijn. Als trainingsdata waarmee een model getraind worden een vooringenomenheid kennen of onvolledig zijn, kunnen de uitkomsten gebaseerd zijn op misplaatst bewijs. Mittelstadt *et al.* stellen verder dat zelfs als de trainingsdata neutraal zijn er nog veel andere factoren zijn die tot uitkomsten gebaseerd op misplaatst bewijs leiden, zoals overtuigingen van de ontwikkelaar bij het ontwerp, keuzes voor bepaalde functionaliteiten en technische beperkingen die hun weerslag hebben het functioneren van het algoritme.[115]

Ondoorgrondelijk bewijs

De werking en uitkomsten van algoritmen kunnen ondoorgrondelijk zijn (het *black box* probleem). Hierdoor is het niet vast te stellen hoe een resultaat tot stand is gekomen en of het betrouwbaar is. Deze ondoorgrondelijkheid heeft een effect op de twee bovenstaande zorgen; als de besluitvorming ondoorgrondelijk is, dan is ook niet na te gaan of die tot stand gekomen is op basis van overtuigend en rechtmatig bewijs.

Oneerlijke uitkomsten

Uitkomsten van algoritmen kunnen overtuigend, transparant en terecht zijn, en toch zorgen oproepen. Bijvoorbeeld omdat het *doel* van de handeling onrechtvaardig, onrechtmatig of risicovol is (niet de manier waarop dat doel wordt bereikt). Een voorbeeld is een gezichtsherkenningsalgoritme dat getraind is op een representatieve groep gezichten (en dus in de werking niet discrimineert), dat vervolgens wordt ingezet om een bepaalde etnische groep te identificeren zodat deze uitgesloten kan worden. Hierbij wordt een adequaat werkend gereedschap ingezet om een oneerlijke uitkomst te bewerkstelligen.

Transformatieve effecten

[114] Oftewel, dat er geen 'betere' oplossing bestaat (met betrekking tot schaal en kosten) betekent niet dat elke uitkomst correct zal zijn.

[115] Mittelstadt, B. D. et al., 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*.

De inzet van algoritmen kan ook een veranderend beeld van de werkelijkheid opleveren, niet alleen vanwege de uitkomsten van een algoritme, maar ook vanwege het feit dat een bepaald effect met een algoritme bereikt wordt. Algoritmen kunnen bijvoorbeeld opties presenteren die een gebruiker nooit had overwogen (omdat deze bijvoorbeeld onbekend bleven). Het gebruikt van een algoritme zorgt dan voor een andere blik op de werkelijkheid. Dit kan positief zijn, maar ook negatief. Een voorbeeld hiervan zijn filterbubbels die zorgen voor een eenzijdige ervaring van een bepaald onderwerp op social media. Het systematisch tonen van gelijkaardige zaken geeft de gebruiker een onrealistisch beeld van de wereld, dat in sommige gevallen problematisch is. Het paradoxale is dat deze personalisatie in sommige gevallen juist een 'beter' beeld van de wereld geeft omdat alleen 'relevante' zaken getoond worden. Een ander voorbeeld is een te groot vertrouwen in algoritmen, waardoor het eigen kritische denkvermogen wordt uitgeschakeld.

Onder transformatieve effecten verstaan wij ook die effecten die zorgen voor een significante verandering in de wijze waarop de betrokken actoren zich verhouden tot de omgeving en elkaar. Een toename in toezicht is bijvoorbeeld een transformatief effect dat ontstaat door de mogelijkheden van algoritmen. De gevolgen van dergelijke effecten kunnen zowel positief als negatief zijn.

10.4 Overzicht onvoorziene effecten

De potentiële onvoorziene effecten van een algoritmisch besluit zijn inherent afhankelijk van een specifieke casus, context en situatie. In deze paragraaf geven wij een algemeen overzicht van onvoorziene (en doorgaans ook ongewenste) effecten van de toepassing van zelflerende algoritmen. Het gaat om effecten die in de casestudies en het literatuuronderzoek zijn gevonden. Het betreft hier uiteraard niet een uitputtende opsomming van alle mogelijke onvoorziene effecten van algoritmen.

De door ons geïdentificeerde onvoorziene effecten zijn overwegend negatief van aard. Dit komt doordat onvoorziene uitkomsten door hun onvoorspelbaarheid doorgaans negatief zijn. Dit kan de indruk wekken dat de toepassing van zelflerende algoritmen overwegend negatieve effecten heeft. Dat is evenwel niet het geval en ook niet het doel van deze rapportage. De bedoelde effecten zullen doorgaans positief zijn. Of de positieve effecten zwaarder wegen dan de (negatieve) onvoorziene effecten is afhankelijk van de omstandigheden van het geval.

Onvoorzien effect	Toelichting
(Extreme) uitkomsten	In tegenstelling tot mensen hebben algoritmen geen gevoel voor een context en 'herkennen' dus niet zelfstandig dat een uitkomst extreem is. Wanneer er geen grenswaarden of andere randvoorwaarden zijn vastgesteld dan kunnen de uitkomsten van een algoritme extreem of zelfs 'onethisch' zijn. Dit is doorgaans een onvoorzien (en ongewenst) effect. Een voorbeeld is de extreem hoge ritprijs voor een taxi tijdens rampen en aanslagen.[116]
Algoritmische beïnvloeding	Individuen en groepen kunnen door algoritmen beïnvloed worden om bepaalde acties te ondernemen (bijvoorbeeld het doen van aankopen) zonder dat zij hier zich echt bewust van zijn. Voorbeelden hiervan zijn micro-targeting van advertenties/berichten om gebruikers van gedrag te laten veranderen, of adviezen van lifestyle coaches.
Machtsverschillen en informatieasymmetrie	Door inzet van algoritmen kan de macht en/of informatiepositie van de gebruiker van het algoritme sterk groter worden, zonder dat dit het specifieke doel is. De analyse van gegevens van een subject kan informatie opleveren waarvan het subject zich niet bewust is. Deze informatie kan gebruikt worden voor oneerlijke handelspraktijken, manipulatie, of andere ongewenste doelen.
(Onbewuste) discriminatie	De uitkomsten van algoritmen kunnen bewust of onbewust vooringenomen zijn ten opzichte van groepen mensen wat ervoor kan zorgen dan individuen anders behandeld worden.
Gedifferentieerde toegang tot goederen en diensten	Profilering van subjecten biedt de mogelijkheid om bepaalde goederen en diensten anders (of niet) aan te bieden. Bijvoorbeeld het uitsluiten van een 'hoog risico' groep van klanten voor bepaalde (abonnements)diensten. Dit kan een bewuste keuze zijn, maar het kan ook een onvoorzien effect zijn van het algoritmisch prijzen of aanbieden van producten.

[116] Evolutionaire algoritmen kunnen in simulaties tot hele originele oplossingen komen om een probleem op te lossen. Sommige 'oplossingen' zijn echter dusdanig onorthodox dat het niet perse valide oplossingen voor het probleem zijn en in sommige gevallen zelfs zeer onwenselijk. Zie: Lehman, J. et al. (2019), The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities

Chilling effects	Een chilling effect is een zelfopgelegde beperking die voortvloeit uit de wetenschap dat je gedrag gemonitord wordt of kan worden.[117] Bijvoorbeeld bepaalde elementen van je leven verbergen op sociale media omdat een (toekomstige) werkgever deze mogelijk niet kan waarderen. De toepassing van algoritmen kan bijvoorbeeld een chilling effect hebben op de samenleving, omdat het monitoring op grote schaal mogelijk maakt.
Verlies begrip besluiten	De inzet van ondoorzichtige of onbegrijpelijke algoritmen kan ervoor zorgen dat betrokkenen geen inzicht meer hebben op hoe een individueel besluit tot stand gekomen is. Dit zorgt ervoor dat niet kan worden nagegaan of het besluit juist is, of welke actie tot een ander besluit kan leiden. Ook ondermijnt dit het vertrouwen in kunstmatige intelligentie.
Oneerlijke uitkomsten	Met de inzet van algoritmen bestaat de kans, net zoals bij menselijke besluitvorming, dat een oneerlijk besluit genomen wordt. Afhankelijk van de taak kan dit grote impact hebben op individuen en groepen. Bijvoorbeeld, wanneer iemand onterecht wordt bestempeld als fraudeur of potentiële terrorist.
Versmald begrip van de toekomst	Door in een algoritme vast te leggen hoe bepaalde besluiten genomen moeten worden verandert de logica van die besluitvorming niet meer. In tegenstelling tot menselijke besluitvorming kunnen algoritmische besluiten niet zelf 'meebewegen'. Een algoritme kent geen 'geest van de wet'. Elke mogelijkheid tot verandering (andere logica, data, doel) moet expliciet ontworpen worden. Zo loopt algoritmische besluitvorming in sommige gevallen het risico de bestaande werkelijkheid te reproduceren omdat er geen andere optie is.[118] Afhankelijk van de toepassing is deze bevriezing van mogelijke uitkomsten meer of minder problematisch. Logischerwijs speelt dit in gebieden waar er een verschil is tussen wat is, en hoe het zou moeten zijn (bijvoorbeeld rond diversiteit), een grote(re) rol.

[117] Büchi, M. et al. (2020). The chilling effects of algorithmic profiling: Mapping the issues. Computer Law & Security Review, 36, 105367.Malik, Farhad.,' Neural Network Layers. Understanding How Neural Network Layers Work', 18-05-2019, https://medium.com/fintechexplained/neural-network-layers-75e48d71f392.min

[118] Wikipedia (s.a.), 'Naturalistische Dwaling'. Geraadpleegd via: https://nl.wikipedia.org/wiki/Naturalistische_dwaling.

Toename surveillance	Algoritmen kunnen, al dan niet in combinatie met sensoren, op veel grotere schaal data genereren en analyseren. Daarnaast kan de analyse inzichten opleveren die voor het subject onbekend zijn, of waarvan het subject zou willen dat die niet bekend zijn.[119] Deze surveillance kan plaatsvinden in de fysieke of de digitale wereld, horizontaal of verticaal zijn.[120]
Toename afhankelijkheid	Het overhevelen van besluitvorming naar algoritmen kan ertoe leiden dat subjecten de oorspronkelijke taak niet meer zelf vervullen en afhankelijk worden van een algoritme. Het is afhankelijk van de toepassing of dit een probleem is en zo ja, hoe groot dat probleem is.
Individualisering/ ontcollectivisering	Personalisering van producten en diensten gebeurt om de gebruikerservaring (door de ogen van de gebruiker van het algoritme) te verhogen. Vaak is personalisatie alleen (economisch) mogelijk door het algoritmisch op maat maken van een product of dienst. De mogelijkheid om individuele besluiten te nemen waar dat voorheen niet kon, kan het draagvlak voor collectieve/publieke oplossingen wegnemen. Bijvoorbeeld, het draagvlak voor collectieve verzekeringen kan afnemen als per individu relevant gedrag gemeten kan worden (bijvoorbeeld voor zorgverzekeringen bij een ongezonde levensstijl).[121]
Stigmatisering en stereotypering	Algoritmische profilering kan ervoor zorgen dat bestaande maatschappelijke stigma's en stereotypes vastgelegd worden in algoritmische besluiten. Een voorbeeld is het classificeren van een man met een witte jas als 'dokter', en een vrouw als 'verpleger'. Wanneer een individu gereduceerd wordt tot data of een profiel, ontstaat het risico op stigmatisering en stereotypering.

[119] Zie ook: Machtsverschillen en informatieasymmetrie

[120] Zie ook: Chilling effects

[121] Van Dijck, J. (2019). Digitale personalisatie mag solidariteit en sociale zekerheid niet ondermijnen. FD. Geraadpleegd via: https://fd.nl/opinie/1311018/digitale-personalisatie-mag-solidariteit-en-sociale-zekerheid-niet-ondermijnen

10.5 Effecten op de korte-, middellange en lange termijn

Bij het onderscheiden van de onvoorziene effecten moeten wij een onderscheid maken tussen effecten die primair toe te schrijven zijn aan de schaal van de toepassing van (zelflerende) algoritmen (*kwantitatieve effecten*) en effecten die primair toe te schrijven zijn aan de specifieke eigenschappen van (zelflerende) algoritmen (*kwalitatieve effecten*).

Op de korte termijn zullen de meeste onvoorziene effecten voortvloeien uit de grondoorzaken zoals hierboven beschreven. Overig is ons beeld op basis van de cases studies, de literatuur en de interviews dat in de praktijk er nog weinig zelflerende algoritmen worden toegepast. Hoewel toepassingen zoals *AlphaStar* van Deepmind en *Hide and Seek* van OpenAI bijzonder indrukwekkend zijn en mogelijk een glimp bieden van de toekomst van kunstmatige intelligentie in onze samenleving, zijn zij niet representatief voor de toepassing van (zelflerende) algoritmen in onze samenleving. De meeste *machine learning* toepassingen die worden toegepast binnen bedrijven zijn een stuk minder complex.

Naarmate meer en meer machine learning toepassingen hun weg vinden in ons dagelijks leven, kunnen transformatieve effecten optreden. Denk aan een toename in toezicht (bijvoorbeeld het gebruik van fraude detectie of *HR analytics*) en mogelijke *chilling effects* die daaruit voortvloeien. Uit angst voor wat een werkgever door middel van algoritmen kan vinden en analyseren gaan potentiële werknemers wellicht hun gedrag aanpassen. Het meest extreme voorbeeld van deze ontwikkeling zien we in China met het '*social credit system*'.[122]

De verwachting is dat kwalitatieve effecten van zelflerende algoritmen met name op de middellange tot lange termijn zullen plaatsvinden. Door verder ontwikkeling en innovatie op het gebied van machine learning worden algoritmen krachtiger en complexer. Dit vergroot de kans op onvoorziene effecten. Voor de lange(re) termijn lijken met name de interacties tussen algoritmische systemen tot onvoorziene effecten te kunnen gaan leiden. Omdat agenten die met elkaar interacteren veel sneller beslissingen nemen dan mensen kan eventuele interactie en competitie tussen hen leiden tot *emergent* gedrag en daarmee tot onvoorziene effecten.

10.6 Afsluitende beschouwing

Het is moeilijk om onvoorziene effecten van zelflerende algoritmen te voorspellen, juist om de reden dat ze onvoorzien zijn. Het effect van een (zelflerend) algoritmen is sterk afhankelijk van de context waarbinnen de algoritmen worden toegepast.

[122] Zie bijvoorbeeld: https://nhglobalpartners.com/chinas-social-credit-system-explained/

Hoewel de effecten uiteenlopend kunnen zijn, lijken de grondoorzaken die leiden tot onvoorziene effecten min of meer gelijk te zijn. Op basis daarvan kunnen we maatregelen nemen om deze oorzaken te adresseren.

Sommige onvoorziene effecten zijn primair toe te schrijven aan de algoritmen zelf en de manier waarop zij functioneren, andere onvoorziene effecten, zoals bijvoorbeeld een toename van surveillance, of afhankelijkheid van algoritmen, vloeien eerder voort uit de toepassing van algoritmen binnen een bredere (socio-technische) context.

11 Inzicht en mitigatie ongewenste effecten

In het vorige hoofdstuk hebben we verschillende onvoorziene effecten in kaart gebracht en de grondoorzaken voor het ontstaan van deze effecten geïdentificeerd. Uit deze oorzaken valt af te leiden dat doorgaans een onvolledig of verkeerd begrip van het probleem (en de daarbij behorende data), een gebrekkige implementatie en gebrekkige monitoring van de zelflerende algoritmen ten grondslag liggen aan onvoorziene effecten, in ieder geval op de korte tot middellange termijn.

Om de kans op onvoorziene effecten zo klein mogelijk te maken en de kans op bedoelde en gewenste effecten zo groot mogelijk, dient de ontwikkelaar en/of de gebruiker van een algoritmisch model een zo goed mogelijk inzicht te hebben in het ontwikkelings- en implementatieproces.

In dit hoofdstuk beschrijven we de maatregelen die organisaties kunnen treffen, om de kans op onvoorziene effecten te verkleinen. Hierin maken we een onderscheid tussen 1) organisatorische maatregelen, 2) technische maatregelen, 3) toezicht op de toepassing van algoritmen en 4) het verbeteren van de positie van het subject.

11.1 Organisatorische maatregelen / governance

In deze paragraaf beschrijven we welke maatregelen organisaties zelf kunnen nemen om inzicht te krijgen in de risico's en mogelijke effecten van de inzet van algoritmische modellen. Het kan zijn dat deze maatregelen nu of in de toekomst van overheidswege worden afgedwongen om uitgevoerd te worden. Het grootste deel van de maatregelen die hieronder worden beschreven zijn aan de organisatie zelf om goed in te regelen en uit te voeren.

Door processen in te richten om de besluitvorming, ontwikkeling, inzet en evaluatie van algoritmen in goede banen te leiden, kunnen onvoorziene effecten mogelijk voorkomen worden.

Er zijn diverse raamwerken voor de *governance* van AI-processen ontwikkeld door organisaties, adviesbureaus en toezichthouders.[123] De precieze uitwerking varieert, maar er zijn een aantal elementen die regelmatig terugkomen:

- het benoemen en beleggen van rollen en verantwoordelijkheden bij de verantwoorde ontwikkeling van modellen;

[123] Bijvoorbeeld: PDPC (privacytoezichthouder Singapore):https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework. ICO (privacytoezichthouder VK): SAS, https://blogs.sas.com/content/hiddeninsights/2017/11/16/model-governance-framework-mrm/

- training en bewustzijn;
- effectbeoordelingen (*impact assessments*);
- documentatie en verslaglegging; en
- interne en externe audits.

Hieronder lichten wij deze mitigerende maatregelen toe.

11.1.1 Rolverdeling binnen de organisatie

Om ervoor te zorgen dat algoritmen en modellen op een zorgvuldige wijze worden ontwikkeld en geïmplementeerd, is het van belang dat er mensen binnen de organisatie zijn die de rol en verantwoordelijkheid toegewezen krijgen om dit te borgen. Het *AI Governance Framework* van de PDPC stelt bijvoorbeeld:

> *"Responsibility for and oversight for the various stages and activities involved in AI deployment should be allocated to the appropriate personnel and/or departments. If necessary and possible, consider establishing a coordinating body, having relevant expertise and proper representation from across the organisation."*[124]

Het is hierbij van belang dat de personen die deze rollen bekleden, voldoende tijd, ruimte en middelen krijgen om deze verantwoordelijkheden op zich te nemen. Verder is het essentieel dat zij goed getraind worden, zodat zij voldoende kennis hebben van de ontwikkeling van algoritmische modellen en de mogelijke effecten die kunnen ontstaan gedurende de ontwikkeling en de inzet van het model.

Op basis van het *AI Governance Framework* van de PDPC hebben dergelijke rollen de volgende verantwoordelijkheden:

- bepalen in welke mate natuurlijke personen betrokken zij bij de mogelijke geautomatiseerde besluitvorming op basis van het algoritmische model;
- overzien van het dataselectieproces en het modeltrainingsproces;
- beheren, monitoren, documenteren en beoordelen van algoritmische modellen die zijn ingezet;
- beoordelen van communicatiekanalen met de relevante *stakeholders* die gebruik maken van de modellen, met als doel op effectieve wijze *feedback* ontvangen; en

[124] PDPC (2019), Model AI Governance Framework.Geraadpleegd via: https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework

- garanderen dat relevante medewerkers die in het ontwikkelingsproces zitten, behoorlijk zijn getraind.[125]

Bij veel organisaties zijn soortgelijke functies reeds ingevuld voor onderwerpen als privacy en informatiebeveiliging (denk aan *privacy* officers en *security officers*). Verder hebben veel organisaties ook een functionaris gegevensbescherming (FG of DPO) aangesteld om intern toezicht te houden op privacy gerelateerde zaken. Soortgelijke rollen kunnen ook worden ontwikkeld voor het toezicht op de ontwikkeling en inzet van algoritmische modellen.

11.1.2 Review boards

Wanneer onderzoekers experimenteel onderzoek doen, dan is het goed mogelijk dat onvoorziene effecten optreden. Bij onderzoek is het de gewoonte dat voorstellen voor (gevoelig) onderzoek door een *review board* beoordeeld worden. Binnen een organisatie kan een soortgelijk orgaan worden ingesteld voor de ontwikkeling van modellen, zodat de stappen binnen het ontwikkelingsproces door een onafhankelijk orgaan kritisch beoordeeld worden.

Het IEEE schetst deze review boards als volgt:

> *"Review boards can provide valuable additional oversight by fielding a diversity of disciplines and deliberating without direct investment in the advancement of research goals."*

> *"Review boards should be composed of impartial experts with a diversity of relevant knowledge and experience. These boards should be continually engaged from the inception of the relevant project, and events during the course of the project that trigger special review should be determined ahead of time."*[126]

Om *review boards* in te richten moeten onder andere de volgende stappen worden genomen:

- Er moet een gevarieerde groep personen met verschillende achtergronden worden geselecteerd voor de review board;
- Deze personen moeten getraind worden zodat zij inzicht hebben in het ontwikkelproces;
- De review board moet in overleg met het management een mandaat vaststellen waarin is beschreven hoe en wanneer zij betrokken worden in het ontwikkelproces.

[125] Ibid.

[126] Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794.

De Rabobank heeft voor HR analytics bijvoorbeeld een dergelijk review proces. De Rabobank heeft een comité ingesteld bestaande uit verschillende rollen en disciplines binnen de organisatie. Het model gaat langs dit comité ter beoordeling, waarbij vragen langskomen als: moeten we dit wel willen gebruiken? Draagt het model daadwerkelijk bij aan de dagelijkse werkzaamheden? Willen we wel deze data hiervoor gebruiken? Feedback vanuit het comité komt terug bij de proceseigenaren van het HR analytics model, waarna aanpassingen plaatsvinden of zelfs besloten kan worden dat het model niet ingezet mag worden.

11.1.3 Training en bewustzijn

Om ervoor te zorgen dat het ontwikkel- en implementatieproces overeenkomt met het projectplan en deze processen conform ethische en juridische normen plaatsvinden, is training en bewustzijn nodig van alle actoren (ontwikkelaar, gebruiker en eindgebruiker).

Over het algemeen moet training en bewustzijn geconcentreerd worden op de volgende doelgroepen: beslissers, ontwikkelaars en eindgebruikers.

Voor beslissers is het belangrijk dat zij weten wat de mogelijkheden en beperkingen van kunstmatige intelligentie en algoritmische modellen zijn. Verder moeten zij weten wat de ethische en juridische risico's zijn bij het implementeren van algoritmische modellen.

Ontwikkelaars dienen bewust te zijn dat de implementatie van algoritmische modellen en daaruit voortkomende besluitvorming kan leiden tot ethische en juridische vraagstukken. De training dient ervoor te zorgen dat een ontwikkelaar bijvoorbeeld begrijpt wat voor data-elementen en modeleigenschappen kunnen leiden tot *bias* in het model.

Het businessmodel van de ontwikkelaar is gewoonlijk dat óf het model door de ontwikkelaar wordt toegepast in de systemen van de gebruiker op basis van een abonnementsvorm (SaaS-oplossing), óf dat het model wordt ontwikkeld voor de systemen van de klant en alleen op die systemen draait. In beide gevallen is het zo dat de ontwikkelaar ervoor moet zorgen dat de gebruiker begrijpt wat het model gaat doen. Wanneer de gebruiker niet begrijpt wat het model doet, dan zal het model niet succesvol functioneren. Dit is zowel in het nadeel van de ontwikkelaar, als in het nadeel van de gebruiker. Daarom is het van belang dat de ontwikkelaar aan de gebruiker goed uitlegt hoe het model werkt en wat nodig is binnen de organisatie van de gebruiker om het model succesvol te implementeren.

> *Daar proberen we heel duidelijk in te zijn. Als de klant niet voldoende klaar is voor onze toepassing, dan wordt het niet succesvol.*
>
> *Ontwikkelaar pricing algoritmen*

Niet alleen moet de ontwikkelaar aan de gebruiker duidelijk uitleggen hoe het model werkt, de gebruiker moet op haar beurt tijd en middelen inzetten om eigen medewerkers te trainen in het gebruik van het model. Deze medewerkers moeten bijvoorbeeld, snappen wat voor data wordt gebruikt voor het model, waarom deze data wordt gebruikt, hoe het model gemonitord moet worden en welke risico's aanwezig zijn bij het gebruik van het model.

Eindgebruikers zijn de medewerkers van de gebruiker die het model zelf direct inzetten. Zij moeten zich bewust zijn van de interactie die zij hebben met het ontwikkelde model. Voornamelijk wat de reikwijdte is van de implementatie van het algoritmische model en wat de beperkingen zijn. Daarnaast moet de eindgebruiker de werking van het model tot op zekere hoogte kunnen monitoren en evalueren.

11.1.4 Effectbeoordelingen AIIA en DPIA

Binnen de bedrijfsvoering is het essentieel dat voor, tijdens en na de ontwikkeling van een algoritmisch model interventies plaatsvinden die ertoe leiden dat de kans op onvoorziene effecten zoveel mogelijk afneemt.

Dergelijke interventies passen goed in het CRISP-DM model (zie hoofdstuk 3). Bij iedere fase van het concept is namelijk een of meerdere momenten ingebouwd om de ontwikkeling en situatie van het model op dat moment te evalueren en van een beoordeling te voorzien. Wanneer de beoordeling op dat moment als onvoldoende wordt beschouwd, dan moet de ontwikkeling van het model één of meerdere stappen terug.

Er zijn verschillende manieren om potentiële negatieve effecten, die ontstaan door het inzetten van algoritmische modellen, te identificeren en te mitigeren. The Center for Data Innovation benoemt bijvoorbeeld "*impact assessments, error analysis and bias testing*."[127] In deze paragraaf concentreren wij ons op de *impact assessments*, oftewel effectbeoordelingen. Deze effectbeoordelingen kunnen van overheidswege verplicht worden gesteld door middel van wetgeving.

Effectbeoordelingen bevatten systematische beschrijvingen van de doelen, de betrokkenen, de risico's en andere relevante aspecten die de voorgenomen toepassing met zich mee brengt. In deze paragraaf behandelen we twee varianten van effectbeoordelingen, namelijk de DPIA en de AIIA.

[127] Center for Data Innovation (2019), 'RE:Competition and Consumer Protection in the 21st Century Hearings, Project Number P181201'.

DPIA

De DPIA is de meest bekende risicobeoordeling en staat voor *Data Protection Impact Assessment*. De DPIA, of gegevensbeschermingseffectbeoordeling is verplicht gesteld in de Algemene Verordening Gegevensbescherming (AVG) voor verwerkingen van persoonsgegevens met een hoog risico.[128] De DPIA is een wettelijke verplichting wanneer de verwerking van persoonsgegevens "waarschijnlijk een hoog risico inhoudt voor de rechten en vrijheden van natuurlijke personen" wiens persoonsgegevens worden verwerkt.[129]

De DPIA moet, op basis van artikel 35 lid 1 AVG, worden uitgevoerd voordat de verwerking van persoonsgegevens plaatsvindt. Wanneer we naar het CRISP-DM model kijken (zie hoofdstuk 3.1), dan dient deze risicobeoordeling dus reeds in de fase van de *Business Understanding* plaats te vinden. Wanneer het projectplan vorm heeft gekregen, de ontwikkeling van het algoritmisch model een specifiek doel kent en duidelijk is wat voor type gegevens nodig zijn voor de volgende fase, dan dienen de potentiële risico's middels de DPIA systematisch op een rij te worden gezet en te worden bedacht welke mitigerende maatregelen genomen moeten worden om deze risico's zoveel mogelijk te mitigeren (ervan uitgaande dat er persoonsgegevens worden verwerkt en er sprake is vaneen hoog risico).

AIIA

De AIIA is een effectbeoordeling, ontwikkeld door het ECP (Electronic Commerce Platform), die er specifiek op gericht is om de juridische en ethische normen die een rol spelen bij de ontwikkeling en inzet van algoritmische modellen inzichtelijk te maken en voor bedrijven aantoonbaar te maken welke afwegingen ten grondslag liggen aan keuzes en besluiten.[130] Deze effectbeoordeling is geen wettelijke verplichting voor organisaties.

De AIIA bestaat uit de volgende 8 stappen:

1. *Bepaal de noodzaak* voor het doen van een AIIA

In deze stap moeten vragen worden gesteld als: heeft het model een hoge mate van autonomie, wordt het model toegepast in een complexe omgeving of is de besluitvorming door het model complex? Wanneer op een van dergelijke vragen een positief antwoord gegeven wordt, dan dient een AIIA te worden uitgevoerd.

[128] Persoonsgegevens zijn gegevens die direct of indirect een natuurlijk persoon kunnen identificeren, zoals een naam, telefoonnummer, emailadres, BSN of bankrekeningnummer.

[129] Artikel 35 lid 1 Algemene Verordening Gegevensbescherming

[130] ECP (2018), 'Artificial Intelligence Impact Assessment. Geraadpleegd via: https://ecp.nl/actueel/artificial-intelligence-impact-assessment/

2. *Beschrijf* de toepassing van het algoritmisch model

In deze stap moet worden beschreven wat voor type algoritmisch model toegepast gaat worden, wat de doelen zijn die met de inzet van het model worden beoogd, welke data worden gebruikt en welke actoren en belanghebbenden relevant zijn voor de toepassing.

3. *Beschrijf de* baten van de inzet van het algoritmisch model

In deze stap wordt geformuleerd welke baten de inzet van het algoritmisch model heeft voor de organisatie, voor het individu waar het model betrekking op heeft en voor de maatschappij als geheel.

4. *Analyseer* of het doel en de wijze waarop dat wordt bereikt ethisch en juridisch verantwoord is

Vervolgens moet de gebruiker beoordelen hoe subjecten geraakt worden door de inzet van het algoritmisch model, welke waarden en belangen van deze subjecten geraakt worden en in hoeverre deze waarden en belangen in wet- en regelgeving geconcretiseerd zijn.

5. *Analyseer* of de toepassing van het algoritmisch model betrouwbaar, veilig en transparant is.

Hier moeten vragen worden gesteld als: welke maatregelen zijn genomen om de betrouwbaarheid, de veiligheid en de transparantie van het handelen van het model te borgen.

6. *Afweging en beoordeling*

In deze stap moet de gebruiker de verschillende belangen wegen en onderbouwen waarom de toepassing van het model rechtmatig en verantwoord is.

7. *Documenteer*

Deze stap is bedoeld om te garanderen dat de bevindingen en daaruit voortkomende afweging en beoordeling goed worden gedocumenteerd.

8. *Evalueer periodiek*

De ontwikkeling van het algoritmisch model, ook na inzet daarvan, houdt natuurlijk niet op. Daarom is het ook van belang dat deze stappen periodiek opnieuw worden gezet.

11.1.4.1 Monitoring en evaluatie

Na implementatie van het model moeten de prestaties en de nauwkeurigheid van het model in de gaten gehouden worden. Een algoritmisch model wordt namelijk getraind met specifiek gekozen trainingsdata. Ondanks het feit dat de trainingsdata beoordeeld kunnen zijn op toepasbaarheid en vooringenomenheid, is de trainingsdata onvermijdelijk data gebaseerd op het verleden. Nu het model wordt geïmplementeerd, worden data als *input* genomen die af kunnen wijken van de trainingsdata. Daarom moeten de werking en de resultaten van het algoritme gemonitord worden na implementatie.

11.2 Technische maatregelen

Technische maatregelen ter voorkoming van onvoorziene effecten richten zich op het verzamelen en gereed maken van data, de ontwikkeling van het model en de inzet in de praktijk.[131] In het algemeen zijn de volgende maatregelen te identificeren die het onvoorziene effecten helpen voorkomen.

11.2.1 Data (management)

De centrale rol van data bij de inzet en ontwikkeling van algoritmen betekent dat datamanagement essentieel is om algoritmen in te zetten. Datamanagement is een wereld op zichzelf en beslaat een breed palet aan processen en onderwerpen zoals: data governance, architectuur, database en opslagbeheer, datakwaliteit, security, privacy, enzovoorts.

In deze paragraaf bespreken we een aantal concepten die bijdragen aan het terugdringen van onvoorziene effecten.

Bias assessments

Bias die voortkomt uit de data waarmee een model getraind wordt kan voor uitkomsten zorgen die niet aansluiten bij het doel waarvoor het algoritme wordt ingezet. Door na te gaan of de gebruikte data geen (verborgen) vooringenomenheid kent, kan een deel van het risico op bias weggenomen worden. Een aantal grote organisaties heeft *toolkits* ontwikkeld om na te gaan hoe vooringenomen de gebruikte data en uitkomsten van het model zijn. Bijvoorbeeld de door IBM ontwikkelde 'AI Fairness 360' toolkit[132] voor detectie en mitigatie van *bias* en *fairness* of de

[131] Uiteraard moeten deze technische maatregelen ondersteund worden door organistorische maatregelen (governance) om ze effectief te implementeren.

[132] https://github.com/IBM/AIF360

'ML Fairness gym'[133] van Google waarmee de evolutie van lerende modellen in verschillende omgevingen gesimuleerd kan worden.

Voor de beoordeling van de representativiteit van datasets, steekproeven en onderliggende verdelingen, zijn diverse technieken breed beschikbaar en bekend voor data scientists.

Data lineage

Voor de ontwikkeling en toepassing van een algoritmisch model worden veel data verzameld en gebruikt. Deze data komen ergens vandaan, worden op een bepaalde manier in het model gestopt en ondergaan verandering door de ontwikkeling en toepassing van het model.

Data lineage (dataoorsprong) helpt bij het mitigeren van risico's die ontstaan door (verkeerd) data gebruik. Data lineage houdt in dat er controle is over de hele levenscyclus van gegevens, van verzameling tot en met gebruik. Dit maakt het voor organisaties mogelijk om na te gaan welke gegevens op welke manier verwerkt en gebruikt worden.[134]

Data lineage betekent dat onder meer wordt bijgehouden waar de data vandaan komen, wat de relatie is tussen data, welke bewegingen de data maken, hoe data getransformeerd worden en welke gebruikers toegang hebben tot de data.[135]

Zonder *data lineage* kan het volgende voorbeeld zich voordoen: iemand exporteert trainingsdata die voor het model wordt gebruikt, deze persoon wijzigt de data en plaatst de data terug in het trainingsdatabase. Zonder monitoring is het onbekend wat er met de data is gebeurd en waar de (nieuwe) data vandaan komen. Het is onbekend of de bron (nog) betrouwbaar is en of de wijzigingen wel kloppen.[136] Door de herkomst, het gebruik en de transformatie van data bij te houden, kunnen onvoorziene effecten bij de ontwikkeling, implementatie en inzet van het model voorkomen worden.

11.2.2 Feature selection en feature engineering

In de casestudies is een aantal keren naar voren gekomen dat de kenmerken op basis waarvan besluiten worden genomen onvoorziene effecten kunnen veroorzaken. Het ontwikkelen van een model zorgt inherent voor een simplificatie van de werkelijkheid. In dit proces worden keuzes gemaakt over wat wel en niet relevant is voor het model. Omdat niet alle kenmerken die

[133] https://github.com/google/ml-fairness-gym

[134] Leclercq, F. et al. (2020). Perfectly parallel cosmological simulations using spatial comoving Lagrangian acceleration. arXiv preprint arXiv:2003.04925.

[135] Harris, J. (s.a.). Data lineage: Making artificial intelligence smarter. SAS. Geraadpleegd via: https://www.sas.com/en_us/insights/articles/data-management/data-lineage--making-artificial-intelligence-smarter.html.

[136] De Hoon, P. (2020). Data Lineage: The What, Why & How (Part 1). Geraadpleegd via: https://home.kpmg/nl/nl/home/social/2020/01/data-lineage-the-what-why-and-how-part-1.html.

een rol spelen in de werkelijkheid meegenomen kunnen worden, moeten er keuzes worden gemaakt wat wel en niet relevant (genoeg) is.

Met de keuze voor een kenmerk ontstaat een nieuwe uitdaging. Het kan namelijk zijn dat een kenmerk in het algemeen (voor het hele model) sterk bijdraagt aan de kwaliteit van de uitkomsten, maar dat dit bij een individuele uitkomst tot onvoorziene effecten leidt. Het gebruik van een postcode is hier een voorbeeld van. Omdat binnen postcodegebieden mensen gelijkenissen kunnen vertonen (gezinssamenstelling, inkomen, afkomst) kunnen modellen betere voorspellingen doen als dit kenmerk wordt meegenomen. Maar het feit dat mensen gemiddeld op elkaar lijken, betekent niet dat iedereen op elkaar lijkt.[137] Bijvoorbeeld, in een wijk met veel jonge gezinnen, wonen niet *alleen* jonge gezinnen. Een model dat een dergelijke inferentie (statistische generalisatie) gebruikt zal in sommige gevallen onjuiste conclusies trekken. Het is belangrijk te begrijpen dat onjuiste uitkomsten inherent zijn aan modelleren. Het doel van *data scientists* is om een zo goed mogelijk model te maken; foute uitkomsten minimaliseren is daar een groot onderdeel van.

Het is dus van belang dat de keuze van kenmerken (*feature selection,* welke kenmerken zijn van belang?) en de ontwikkeling van kenmerken (*feature engineering,* op welke manier worden deze beschreven?) weloverwogen te nemen. Wachter en Mittelstadt pleiten voor een 'recht op redelijke inferenties' om gebruikers te beschermen tegen onredelijke generalisaties.[138] De voorgestelde verplichte *disclosure* van informatie over gebruikte inferenties geeft een indruk welke overwegingen relevant zijn:

- Is het gebruik van de data voor generalisaties normatief te verantwoorden?
- Is het gebruik van deze generalisaties relevant voor het specifieke doel en normatief te verantwoorden?
- Zijn de data en de methoden van analyse accuraat en statistisch betrouwbaar?

[137] De zogenaamde ecologische fout, zie ook: https://nl.wikipedia.org/wiki/Ecologische_fout

[138] Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Colum. Bus. L. Rev., 494.

11.2.3 Vergroten transparantie

Transparantie is een veelgebruikte term in het maatschappelijk debat als het gaat om oplossingen voor de onvoorziene effecten van (zelflerende) algoritmen.[139],[140] Transparantie wordt vaak in een adem genoemd met termen als interpreteerbaarheid, uitlegbaarheid, openheid of traceerbaarheid. De European Parliamentary Research Service definieert transparantie als volgt:

> *"code, logic, model, goals, decision variables, or some other aspect that is considered to provide insight into the way the algorithm performs. Algorithmic system transparency can be global, seeking insight into the system behaviour for any kind of input, or local, seeking to explain a specific input-output relationship."*[141]

Transparantie van algoritmen kan dus begrepen worden als alles dat bijdraagt aan een begrip over de uitkomst van algoritme. De ICO maakt een helder onderscheid tussen zes verschillende vormen van transparantie van algoritmen vanuit verschillende perspectieven:

Ratio

Een opsomming van redenen die leiden tot het resultaat van de toepassing van het algoritme. Deze redenen worden gepresenteerd in toegankelijke en niet-technisch taalgebruik.

Verantwoording

Benoeming van de organisaties die onderdeel zijn van de ontwikkeling en implementatie van het algoritme en met wie contact opgenomen kan worden om menselijke besluitvorming te verzoeken.

Data

De data die is gebruikt voor het ontstaan van het algoritmisch model.

Fairness

De stappen die zijn genomen door de ontwikkelaar en de gebruiker om vooringenomenheid in het algoritmisch model te voorkomen en eerlijke resultaten te garanderen.

---

[139] Kist, R. (2020). 'We moeten sociale media dwingen transparant te zijn'. NRC Handelsblad. Geraadpleegd via: https://www.nrc.nl/nieuws/2020/06/25/we-moeten-sociale-media-dwingen-transparant-te-zijn-a4003962.

[140] Steen, M. (2020). Discussie over de transparantie van algoritmen blijft nodig. Het Parool. Geraadpleegd via: https://www.parool.nl/columns-opinie/discussie-over-de-transparantie-van-algoritmen-blijft-nodig~b882ed5f/.

[141] EPRS, 'A governance framework for algorithmic accountability and transparency', Panel for the Future of Science and Technology, 04-2019, https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf.

Betrouwbaarheid en nauwkeurigheid

De stappen die zijn genomen die garanderen dat de toepassing van het algoritmisch model zo nauwkeurig, betrouwbaar en veilig mogelijk is.

Impact

De stappen die zijn genomen om te ontdekken welke impact het model heeft op het subject, welke belangenafweging is genomen om het model toe te passen en op welke wijze de impact wordt gemonitord.[142]

De ICO legt hier dus de nadruk op de verantwoording richting de betrokkene. In deze sectie leggen wij de nadruk op mogelijkheden om de 'technische' transparantie van modellen te vergroten. Deze technische transparantie is relevant omdat modellen die ontwikkeld zijn met machine learning veelal niet direct te interpreteren zijn. Modellen die *theoretisch* direct te interpreteren zijn, dat wil zeggen de complete werking van het algoritme is te begrijpen, verliezen deze transparantie snel als ze 'groter' worden. Het is dus belangrijk te begrijpen dat transparantie niet hetzelfde is als uitlegbaarheid of direct leidt tot begrip over hoe een besluit tot stand is gekomen.

Voor modellen die niet direct te interpreteren zijn moet een aanvullende strategie worden gekozen om de werking van het model in het algemeen (globale interpretatie), of voor een specifiek besluit (lokale interpretatie) beter te begrijpen. De technieken om de 'black box' van het model te openen zijn volop in ontwikkeling er kunnen een aantal technieken worden onderscheiden die informatie opleveren over de werking.[143]

Transparantie betekent overigens niet dat ieder algoritme voor iedere actor volledig open en begrijpelijk moet zijn. Het draait om aspecten van transparantie die ervoor zorgen dat het algoritme voor desbetreffende actor voldoende transparant is. Mensen gebruiken bijvoorbeeld *smartphones* en *laptops* voor al hun dagelijkse werkzaamheden. De meeste mensen hebben echter geen idee hoe deze apparaten precies werken en dat hoeft ook niet. Echter, wanneer mensen dat wel willen weten, dan is het mogelijk om deze uitleg te vinden. Fabrikanten hebben handleidingen beschikbaar, er is een reparatieservice en/of een klantenservice en er zijn doorgaan onafhankelijke experts waar de consument terecht kan. Iets dergelijks is er niet, of maar zeer beperkt voor algoritmische besluitvorming.

[142] Zie: Information Commissioner's Office (2020), Explaining decisions made with AI. Via: https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf

[143] Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Geraadpleegd via: https://christophm.github.io/interpretable-ml-book/

De onderstaande maatregelen kunnen bijdragen aan het verhogen van de transparantie.

Transparantie door gebruik interpreteerbare modellen

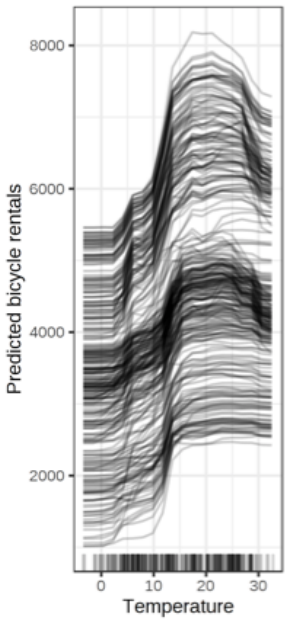
Wanneer gebruik wordt gemaakt van algoritmische modellen die interpreteerbaar zijn voor mensen (begrijpelijk, doorgrondelijk) kunnen onvoorziene effecten worden voorkomen.[144] De mate van openheid kan verschillen, maar het doel is dat het resultaat van de werking van het algoritme geïnterpreteerd kan worden. Dit zorgt ervoor dat wanneer onvoorziene effecten plaatsvinden, met een groter gemak kan worden bepaald waar het probleem zit. Het sneller vinden van het probleem zorgt voor een snellere oplossing en een kortere duur van desbetreffende effect. Een goed uitgangspunt is dus om het meest simpele model te gebruiken om een probleem op te lossen. Wanneer een interpreteerbaar model niet volstaat, dan pas moet worden gekeken naar complexere (machine learning) modellen.

Visualisatie

Door het model systematisch veranderende inputdata te geven kan uit de uitkomsten afgeleid worden welke rol verschillende kenmerken spelen in het model, hoe deze interacteren, en wat de aard van de relatie tussen de input en de output is. Onder deze methoden vallen bijvoorbeeld: *partial dependence plots, acumulated local effects* en *individual conditional expectations (ICE).*

Figuur 5 laat een model zien dat fietsverhuur voorspelt aan de hand van een aantal kenmerken en hoe de voorspellingen veranderen als de temperatuur varieert (elke lijn is een voorspelling). Op basis hiervan is te concluderen dat het kenmerk temperatuur (vrijwel) dezelfde rol speelt bij alle voorspellingen.

Deze technieken zijn gangbare statische methoden en breed beschikbaar in diverse softwarepakketten.



Figuur 5 ICE Plot

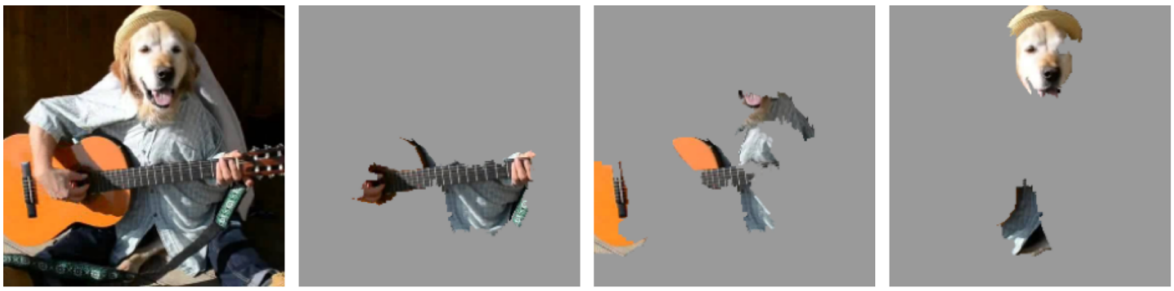
Surrogaatmodellen

Een andere strategie is om plaats van het proberen in de black box te kijken, een model te maken dat de werking van het originele model benaderd maar wel te interpreteren is. Hierdoor

[144] Sciforce (2020). Introduction to the White-Box AI: The Concept of Interpretability'. Geraadpleegd via: https://medium.com/sciforce/introduction-to-the-white-box-ai-the-concept-of-interpretability-5a31e1058611.

kan voor een specifieke uitkomst benaderd worden welke kenmerken op welke manier een rol spelen. Technieken die in deze categorie vallen zijn bijvoorbeeld LIME, Anchors en SHAP. Surrogaatmodellen geven géén inkijk in de werking van een model, maar zijn daar benaderingen van. Als het surrogaatmodel net zo goed presteerde als het origineel dan is er geen reden het origineel nog te gebruiken. Surrogaatmodellen zijn dus nuttig om een inzicht te krijgen in hoe een model ongeveer werkt, maar ze bewijzen niet volledig hoe in het originele model een besluit tot stand is gekomen.

Figuur 6 laat zien hoe met LIME onderdelen van een foto uitgelicht kunnen worden die invloedrijk zijn voor een bepaalde classificatie.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

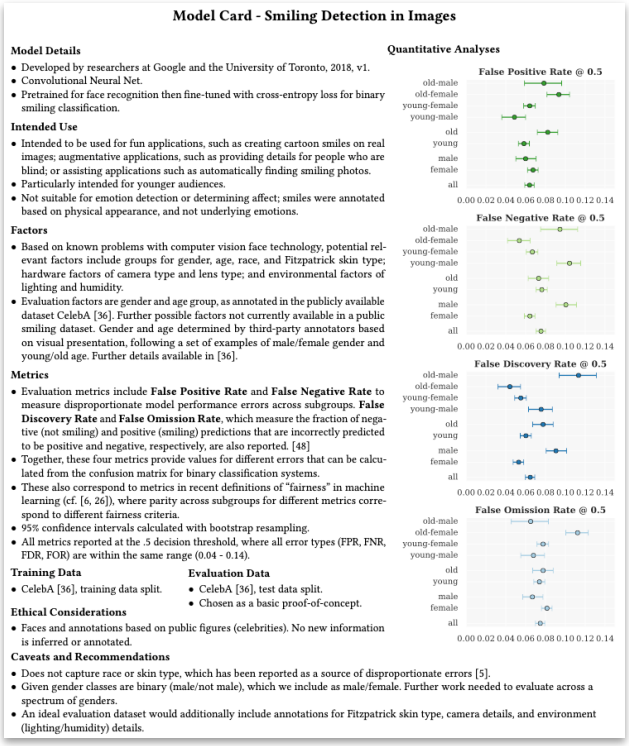
Figuur 6 Voorbeeld van een 'uitleg' van een geclassificeerde foto met LIME[145]

Model reporting cards

Een manier om het ontstaan en de werking van het algoritme transparant te maken is door middel van een *model reporting card*. De *model reporting card* is als een lijst met ingrediënten, zoals die ook voor levensproducten wordt gebruikt.

[145] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

"Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups […] that are relevant to the intended application domains."[146]



Model Card - Smiling Detection in Images

Model Details
- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use
- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors
- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data
- CelebA [36], training data split.

Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

Figuur 7 Model Card voor een algoritmisch model voor het detecteren van een glimlach in afbeeldingen

Deze kaart is bijvoorbeeld zeer geschikt om de door ICO genoemde zes categorieën te presenteren. In het figuur is een voorbeeld van een dergelijke kaart te zien.[147]

Deze 'ingrediëntenlijst' draagt op verschillende manieren bij aan het mitigeren van onvoorziene effecten. Om een dergelijke lijst te kunnen maken, moet de ontwikkelaar het ontstaan en de werking van het model goed documenteren. Dat betekent ook dat de ontwikkelaar het algoritme uitlegbaar moet maken. De uitlegbaarheid moet aangepast zijn op de doelgroep voor wie de uitleg is bestemd.

Daarnaast kan de *model reporting card* voor transparantie zorgen ten opzichte van de gebruiker, de eindgebruiker en het subject, waardoor deze actoren met een groter bewustzijn gebruik kunnen maken van het algoritme. Hierbij is het wel van belang dat deze actoren de uitlegbaarheid publiceren en toegankelijk maken.

11.2.4 Fairness assessments

Een algoritmisch model gebruikt data om een berekening/analyse te maken op basis van het vastgestelde doel. Voor het succes van het model is het logisch dat nauwkeurigheid van het resultaat essentieel is. Het is echter mogelijk dat het resultaat nauwkeurig is, maar dat het resultaat oneerlijk is of op oneerlijke wijze tot stand komt.

[146] Mitchell, M. et al. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

[147] Ibid.

Bijvoorbeeld de *recruiting tool* van Amazon die een vooringenomenheid vertoonde ten aanzien van geslacht: vrouwen werden minder geschikt geacht dan mannen.[148] Wanneer deze tool wordt ingezet om een individu te beoordelen op geschiktheid, dan kan het resultaat (bijvoorbeeld niet aannemen) kloppen. Misschien past deze persoon inderdaad niet bij de organisatie. Echter, het feit dat het resultaat is gebaseerd op een vooringenomenheid ten aanzien van geslacht leidt ertoe dat dit als oneerlijk moet worden beschouwd.

Daarom is het belangrijk dat een *fairness assessment* wordt uitgevoerd op het algoritmisch model. Een *fairness assessment* onderzoekt het algoritmisch model op zichzelf en de resultaten die het model als *output* heeft. Het doel van dit assessment is het uitsluiten van vooringenomenheid in de data die het model gebruikt en de vooringenomenheid in de *output* van het model. Uitsluiten van vooringenomenheid houdt bijvoorbeeld in dat mensen niet direct of indirect gegroepeerd worden op basis van eigenschappen als geslacht en migratie-achtergrond.[149]

11.3 (Extern) toezicht

Naast intern genomen maatregelen door de gebruiker zelf, kunnen ook externe maatregelen ervoor zorgen dat het bedrijf beter inzicht krijgt in het ontwikkelingsproces van algoritmische modellen. In deze paragraaf beschrijven we de invloed van externe organisaties, zoals toezichthouders, het uitvoeren van externe audits en het volgen van *best practices*.

11.3.1 Toezichthoudende organisaties

Het is noodzakelijk dat organisaties zelf middelen inzetten om toezicht te houden op de ontwikkeling en inzet van eigen algoritmische modellen. Echter, bij de ontwikkeling en inzet van algoritmische modellen is veel tijd, moeite en geld gemoeid van de organisatie. Verschillende en (schijnbaar) conflicterende belangen kunnen de aandacht voor de preventie van onvoorziene effecten beïnvloeden. Het verplichten van bepaalde maatregelen en het belasten van onafhankelijke organen met toezicht houden op de ontwikkeling van algoritmen kan een oplossing zijn.

De Raad van Europa stelt in *Unboxing Artificial Intelligence* voor dat lidstaten een juridisch kader moeten vaststellen zodat op onafhankelijke en effectieve wijze toezicht gehouden kan

[148] Lauret, J. (2019). Amazon's sexist AI recruiting tool: how did it go so wrong?. Geraadpleegd via: https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e.

[149] Koshiyama, A. & Engin, Z. (2019), 'Algorithm Impact Assessment: Fairness, Robustness and Explainability in Automated Decision-Making', 10-12-2019, https://zenodo.org/record/3241980.

worden op de ontwikkeling, inzet en gebruik van algoritmische modellen bij zowel overheidsorganisaties als private partijen. Een dergelijke organisatie dient actief organisaties en de ingezette algoritmische modellen te onderzoeken en te monitoren. Verder dient zij klachten af te handelen van personen die het leidend voorwerp zijn geweest van een algoritmisch model.[150]

Een dergelijk orgaan kan volledig extern en onafhankelijk zijn, maar kan ook deels bestaan uit relevante experts uit het bedrijfsleven. Op die manier kunnen private partijen gezamenlijk met de onafhankelijke experts toezicht houden op het ontwikkelproces. Wellicht worden voorbeelden met elkaar gedeeld, beoordeeld en komen daar *best practices* uit voort die gevolgd worden door de bedrijven die een algoritmisch model willen gaan ontwikkelen.

11.3.2 Externe audits

Een tweede vorm van toezicht is de inzet van externe auditors. Audits worden uitgevoerd om te beoordelen of de organisatie handelt in lijn met wet- en regelgeving, *best practices* en eigen beleid en procedures.

Audits kunnen zowel intern als extern worden uitgevoerd. Het voordeel van externe audits is dat deze worden uitgevoerd door organisaties die gespecialiseerd zijn in het uitvoeren van audits en de externe auditors hebben een onafhankelijke relatie tegenover de organisatie vergeleken met medewerkers die een interne audit uitvoeren.[151] Met de externe audit kan de ontwikkelaar of gebruiker de totstandkoming en de implementatie van het algoritmische model toetsen.

ISACA en ICO geven voorbeelden van hoe externe audits op algoritmische modellen uitgevoerd kunnen worden en criteria op basis waarvan een externe auditor het algoritme dient te toetsen. Beide organisaties benaderen de audit vanuit een breed perspectief. Aspecten die getoetst moeten worden tijdens de audit zijn onder meer:

- het data selectie proces;
- het modeltrainingsproces;
- het model validatieproces;
- de uitlegbaarheid van de uitkomsten van het model;[152]

[150] Council of Europe (2019). Unboxing Artificial Intelligence: 10 steps to protect Human Rights., p. 10.

[151] Wilson, G. (2017). The Intersection of Internal and External Audit. Workiva. Geraadpleegd via: https://www.workiva.com/sites/workiva/files/pdfs/thought-leadership/intersection-of-internal-and-external-audit-greg-wilson-white-paper-20170619-j5998.pdf.

[152] ISACA (2018). Auditing Artificial Intelligence.

- fairness en transparantie;
- nauwkeurigheid;
- dataminimalisatie;
- uitvoerbaarheid van rechten van betrokkenen; en
- impact op de betrokkenen in bredere zin.[153]

De aspecten die getoetst worden via een externe audit komen overeen met maatregelen die we eerder hebben genoemd ter mitigatie van onvoorziene effecten en zijn te plotten op het CRISP-DM model. Bij iedere stap in het proces moet namelijk een evaluatie plaatsvinden. Verder komen deze audit aspecten naar voren bij de uitvoering van DPIA's en AIIA's en het mogelijk maken van *data lineage* en transparantie in het algoritmisch model.

11.4 Bescherming van het subject (de consument)

Consumenten zullen doorgaans de subjecten zijn die aan algoritmische besluitvorming worden onderworpen (bijvoorbeeld fraude detectie of *credit scoring*), of van de algoritmen van aanbieders gebruik maken om zelf beslissingen te nemen (bijvoorbeeld over een gezonde levensstijl). Dit betekent dat hun mogelijkheden om algoritmen te analyseren of beïnvloeden doorgaans zeer beperkt tot niet aanwezig zijn.[154]

Om consumenten te beschermen tegen de onvoorziene effecten van algoritmen lijkt daarom het beschermen van de rechtspositie van het subject (de consument) de meest kansrijke route. Enerzijds betekent dit het stellen van regels met betrekking tot de ontwikkeling en toepassing van (zelf)lerende algoritmen en anderzijds het versterken van de rechten van subjecten.

Daar waar het gaat om de ontwikkeling en toepassing van algoritmen kunnen allereerst kwaliteitseisen worden gesteld aan de ontwikkeling en inzet van algoritmen. Dit kan gepaard gaan met de (verplichte) inzet van risicomanagement middelen en processen zoals hierboven beschreven. Ook zouden voor (zeer) risicovolle toepassingen extra eisen aan de begrijpelijkheid en transparantie van algoritmen kunnen worden gesteld.

Verder kan gedacht worden aan informatie- en notificatieplichten richting consumenten. Zo moet niet alleen het gebruik van algoritmische besluitvorming kenbaar zijn, maar ook hoe een besluit tot stand is gekomen. Dit betekent niet alleen een algemene uitleg van de logica van de

[153] ICO (2019). An overview of the Auditing Framework for Artificial Intelligence and its core components. Geraadpleegd via: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

[154] Op de langere termijn zou het wellicht mogelijk zijn voor consumenten om een contra-expertise te doen met behulp van alternatieve modellen, maar vooralsnog lijkt dit geen reële optie.

besluitvorming, maar ook hoe een beslissing in het concrete geval tot stand is gekomen. Bijvoorbeeld, welke *features* speelden een belangrijke rol in de totstandkoming van het besluit en wat had de consument bijvoorbeeld anders moeten doen om wél een positieve uitkomst te krijgen (*counterfactual explanations*).[155] Deze verplichting is in het bijzonder relevant daar waar het gaat om zelflerende algoritmen waarvan de werking ondoorgrondelijk is.

Als sluitstuk kan het (product)aansprakelijkheidsregime uit Burgerlijk Wetboek dienen. Wanneer een ontwikkelaar of een gebruiker een algoritme implementeert en dit leidt tot schade bij de consument, dan moet de ontwikkelaar en/of de gebruiker daarvoor aansprakelijk kunnen worden gesteld.

Naast het stellen van eisen aan ontwikkelaars en gebruikers van algoritmen kunnen consumenten ook rechten krijgen die zij zelf kunnen inroepen of uitoefenen.[156] Bijvoorbeeld het recht op een verklaring of een recht op een *second opinion*, al dan niet met een menselijke toets.

Of voor het implementeren van deze beschermingsmaatregelen nieuwe wet- en regelgeving nodig is, vormt niet het voorwerp van dit onderzoek. Wel hebben wij de indruk dat binnen het bestaande recht (bijvoorbeeld het consumentenrecht, wetgeving op het gebied van productaansprakelijkheid en het gegevensbeschermingsrecht) al veel mogelijkheden zijn voor het reguleren van algoritmische besluitvorming.

Naast het versterken van de rechtspositie van subjecten is het ook verstandig om de bewustwording bij consumenten met betrekking tot algoritmische besluitvorming en sturing te vergroten. Hierbij is het niet alleen van belang dat de consument zich beseft dat er sprake is van algoritmische besluitvorming, maar ook dat deze niet onfeilbaar is (net als menselijke besluitvorming). Het is van belang dat de consument zelf alert blijft en kritisch blijft denken. Een voorbeeld hiervan is om als bestuurder van een autonome auto nog steeds zelf alert te blijven.

[155] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.

[156] Daar waar iets als een recht voor de consument geformuleerd is zou je ook kunnen stellen dat dit daarmee een plicht schept voor de ontwikkelaar en/of gebruiker. Wel kan worden nagedacht over rechten die de consument actief moet inroepen (zoals bijvoorbeeld het recht om een verklaring te eisen) versus directe verplichtingen (de gebruiker moet altijd een verklaring geven).

12 Samenvatting en conclusies

De probleemstelling van dit onderzoek luidde:

> *Wat zijn mogelijke onvoorziene effecten van de inzet van (zelflerende) algoritmen door bedrijven en consumenten waarvan niet duidelijk is hoe zij tot een besluit komen en hoe kunnen deze effecten geïdentificeerd, gewogen en indien ongewenst gemitigeerd worden?*

Uit de literatuur en de casestudies blijkt dat onvoorziene effecten sterk afhankelijk zijn van de context waarbinnen algoritmen worden toegepast. Een sluitend overzicht van potentiële onvoorziene effecten is dan ook niet te geven, niet alleen omdat er zoveel mogelijke contexten zijn, maar ook omdat het onmogelijk is om (alle) onvoorziene effecten te voorzien. Het is natuurlijk überhaupt lastig om onvoorziene effecten te voorspellen, precies om de reden dat ze onvoorzien zijn.

Wat verder naar voren komt is dat onvoorziene effecten doorgaans negatief zijn. Dit heeft evenwel minder te maken met algoritmen, als met het feit dat onvoorziene effecten door hun onvoorspelbaarheid doorgaans eerder negatief dan positief zijn. Het is dan ook van belang om te onderstrepen dat dit niet een onderzoek betreft naar de positieve of negatieve effecten van algoritmen. In zoverre negatieve effecten van algoritmen de boventoon voeren in dit rapport, is dit omdat we ons enkel gericht hebben op onvoorziene effecten. De bedoelde effecten van algoritmen zullen doorgaans positief van aard zijn, dit is immers waarom men ze wil inzetten.

Het is ook relevant om ons te beseffen dat 'onvoorzien' een subjectief en relatief begrip is. In hoeverre een effect onvoorzien is hangt in sterke mate af van hoe goed je de toepassing en de mogelijke effecten daarvan doordacht hebt. Wat extreem gesteld: voor een roekeloos persoon zijn de gevolgen van het eigen handelen doorgaans onvoorzien, maar voor een voorzichtig iemand zijn diezelfde gevolgen wel voorzien.

Om de (negatieve) gevolgen van onvoorziene effecten van zelflerende algoritmen te beperken is het van belang de oorzaken die ten grondslag liggen aan het ontstaan van onvoorziene effecten nader te duiden. We kunnen stellen dat er drie 'grondoorzaken' zijn voor het optreden van onvoorziene effecten in de context van de toepassing van (zelf)lerende algoritmen.

1. Er is een onvolledig of verkeerd begrip van de probleemruimte.
2. Het zelflerende algoritme is niet goed toegerust om om te gaan met de complexe omgeving waarbinnen het wordt ingezet.
3. Het zelflerende algoritme wordt niet goed ingepast in een bredere (socio-technische) context.

Uiteraard kunnen ook combinaties van deze grondoorzaken leiden tot onvoorziene effecten.

De hierboven genoemde grondoorzaken spelen nu reeds een rol en zullen dat op de middellange tot lange termijn blijven doen. Mogelijk kan voor de middellange tot lange termijn het vraagstuk van emergent gedrag door de interactie tussen verschillende algoritmen ook als grondoorzaak relevant worden, maar op dit moment lijken er weinig situaties te zijn waar het vraagstuk van emergent gedrag naar voren komt.

Wat de onvoorziene gevolgen van algoritmen zijn en hoe groot de impact daarvan is, is in belangrijke mate afhankelijk van hoe ontwikkelaars, gebruikers en de maatschappij als geheel omgaan met de bovenstaande grondoorzaken.

De ondoorgrondelijkheid van zelflerende algoritmen (het *black box* probleem) vormt een complicerende factor bij de toepassing van zelflerende algoritmen. We kunnen stellen dat de ondoorgrondelijkheid van zelflerende algoritmen niet zozeer zorgt voor onvoorziene effecten op zichzelf, maar dat het zicht op de grondoorzaken voor het optreden van onvoorziene effecten ontneemt. Bij ontwikkelaars en gebruikers kan het idee bestaan dat hun modellen goed functioneren, terwijl er in werkelijkheid verkeerde beslissingen worden genomen of voorspellingen worden gedaan die niet accuraat zijn. Dit vergroot de kans op onvoorziene effecten.

Om de kans op onvoorziene effecten van (zelflerende) algoritmen te verkleinen is het zaak om de grondoorzaken weg te nemen die leiden tot onvoorziene effecten. Het gaat om maatregelen die in de verschillende stadia van het ontwikkelproces van modellen relevant zijn. Het gaat om:

1. organisatorische maatregelen;
2. technische maatregelen;
3. (extern) toezicht;
4. het beschermen van de rechtspositie van subjecten (consumenten).

Ad 1) Organisatorische maatregelen

Organisatorische maatregelen zijn met name gericht op het adresseren van de epistemische en normatieve zorgen rondom de toepassing van algoritmen.

Veel van de epistemische zorgen (niet overtuigend bewijs, misplaatst bewijs en ondoorgrondelijk bewijs) kunnen worden weggenomen door het goed definiëren en begrijpen van het probleem dat het model moet oplossen en de data die nodig zijn om dit probleem goed te modelleren (de *business* en *data understanding* fasen in het CRISP-DM model). Verder moet de toepassing van het model in de praktijk nauwkeurig gemonitord worden. Dit vereist

maatregelen op het gebied van *governance* en *risk management* voor het inzetten van modellen.

Om de normatieve zorgen te adresseren (oneerlijke beslissingen, transformatieve effecten) is een besef van de mogelijkheden, onmogelijkheden en risico's van de toepassing van algoritmische modellen noodzakelijk. Dit betekent allereerst bewustwording binnen de organisatie, maar daarnaast ook een duidelijk proces voor het inschatten van de risico's. Door middel van impact assessments kan bijvoorbeeld een beeld worden gekregen van de mogelijke risico's van de toepassing, waardoor onvoorziene effecten hopelijk reeds in de ontwikkelfase worden voorzien. Hierbij is het van belang dat niet alleen wordt gekeken naar het model zelf en de correcte werking daarvan, maar ook naar de bredere inbedding als socio-technisch systeem. Door naar dit laatste te kijken kunnen mogelijke negatieve transformatieve effecten eerder worden voorzien.

Tenslotte speelt *accountability* een belangrijke rol als organisatorische maatregel. Ontwikkelaars en gebruikers moeten verantwoording kunnen afleggen over hun gemaakte keuzes. Waarom is gekozen voor een bepaald type model? Hoe is het probleem beschreven? Welke datasets zijn geselecteerd en waarom?

Ad 2) Technische maatregelen

Technische maatregelen zijn met name gericht op het wegnemen van epistemische zorgen door te zorgen voor de juiste samenstelling van (trainings)data, het testen en valideren van modellen en het inzichtelijk/transparant maken van uitkomsten van black box modellen. Daar waar het gaat om het inzichtelijk maken van black box modellen, dragen technische maatregelen ook bij aan het kunnen identificeren van oneerlijke uitkomsten (een normatieve zorg). Technische maatregelen dienen hand in hand te gaan met de organisatorische maatregelen.

Ad 3) (Extern) toezicht

Intern en extern toezicht op de ontwikkeling en toepassing van zelflerende algoritmen is in het bijzonder van belang bij toepassingen waar een potentieel grote impact op de rechten en vrijheden van personen, organisaties of groepen valt te verwachten. Om goed toezicht te kunnen houden is de verantwoording van de ontwikkeling en het gebruik van algoritmen door de ontwikkelaar en de gebruiker noodzakelijk.

Ad 4) Het beschermen van de rechtspositie van subjecten

Omdat het subject van een algoritmisch model (een werknemer, een consument of een bedrijf) doorgaans geen invloed kan uitoefenen op de werking van een model, lijken met name maatregelen die zijn gericht op het beschermen of versterken van de rechtspositie van het subject relevant. Hierbij kan gedacht worden aan juridische bescherming via het gegevensbeschermingsrecht, het aansprakelijkheidsrecht het consumentenrecht en meer indirect het mededingingsrecht. Bij het beschermen van de rechtspositie van subjecten kan enerzijds gedacht worden aan het toekennen van rechten die het subject zelf moet inroepen, maar ook aan verplichtingen of verboden voor de gebruikers van algoritmen. Daarnaast is bewustwording over algoritmische besluitvorming en de effecten daarvan voor subjecten ook noodzakelijk.

We kunnen concluderen dat de toepassing van (zelflerende) algoritmen in alle door ons onderzochte cases van groot belang is en kan bijdragen aan de accuraatheid, efficiëntie en accuraatheid van besluitvorming. Tegelijkertijd zorgt een verkeerde toepassing ook voor onvoorziene effecten die door hun onvoorspelbaarheid doorgaans negatief van aard zijn. De onvoorziene effecten kunnen echter door een goed doordachte en zorgvuldige toepassing van kunstmatige intelligentie beperkt worden.

13 Bibliografie

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160.

Aggarwal, C. C. (2018), Neural Networks and deep learning. Springer International Publishing AG.

Akerlof, G. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. In: *The Quarterly Journal of Economics,* 84(3), 488–500.

Alberdingk Thijm, C. (2000). *Privacy versus auteursrecht in een digitale omgeving*. ITER reeks

Autoriteit Consument en Markt (2020). Leidraad Bescherming online consument. Geraadpleegd via: https://www.acm.nl/nl/publicaties/leidraad-bescherming-online-consument

Baker, B. et al. (2019). Emergent tool use from multi-agent autocurricula. arXiv preprint arXiv:1909.07528.

Center for Data Innovation (2019), 'RE:Competition and Consumer Protection in the 21st Century Hearings, Project Number P181201'.

Chapman, P. et al. (2000). CRISP-DM 1.0. CRISP-DM Consortium, 76(3).

Council of Europe (2019). Unboxing Artificial Intelligence: 10 steps to protect Human Rights.

Dawkins, R., & Krebs, J. R. (1979). Arms races between and within species. Proceedings of the Royal Society of London. Series B. Biological Sciences, 205(1161), 489-511.

De Hoon, P. (2020). Data Lineage: The What, Why & How (Part 1). Geraadpleegd via: https://home.kpmg/nl/nl/home/social/2020/01/data-lineage-the-what-why-and-how-part-1.html.

Dörner, D. (1997). The logic of failure: Recognizing and avoiding error in complex situations. Basic Books.

Downing, D. A et al. (2000). Dictionary of computer and Internet terms. Barron's Educational Series Inc..

ECP (2018), 'Artificial Intelligence Impact Assessment. Geraadpleegd via: https://ecp.nl/actueel/artificial-intelligence-impact-assessment/

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. Management science, 49(10), 1287-1309.

Ensign, D. et al. (2018), Runaway Feedback Loops in Predictive Policing, Proceedings of Machine Learning Research 81:1–12, 2018 Conference on Fairness, Accountability, and Transparency

Koene, A. et al. (2019). A governance framework for algorithmic accountability and transparency.

Harris, J. (s.a.). Data lineage: Making artificial intelligence smarter. SAS. Geraadpleegd via: https://www.sas.com/en_us/insights/articles/data-management/data-lineage--making-artificial-intelligence-smarter.html.

Healy, T. (2012). The unanticipated consequences of technology. Nanotechnology: ethical and social Implications, 155-173.

Hijink, M. (2018). Hoe wordt je kredietscore berekend? NRC Handelsblad. Geraadpleegd via: https://www.nrc.nl/nieuws/2018/12/27/hoe-wordt-je-kredietscore-berekend-a3127138.

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. Yale Journal of Law and Technology, 18(148).

ICO (2019). An overview of the Auditing Framework for Artificial Intelligence and its core components. Geraadpleegd via: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794.

ISACA (2018). Auditing Artificial Intelligence

Annoni, A. et al. (2018) . Artificial Intelligence: A European Perspective. JRC Working Papers JRC113826. Joint Research Centre

Kaziunas, E., Ackerman, M. S., Lindtner, S., & Lee, J. M. (2017, February). Caring through data: Attending to the social and emotional experiences of health datafication. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 2260-2272).

Kist, R. (2020). 'We moeten sociale media dwingen transparant te zijn'. NRC Handelsblad. Geraadpleegd via: https://www.nrc.nl/nieuws/2020/06/25/we-moeten-sociale-media-dwingen-transparant-te-zijn-a4003962.

Kollmeyer, B. (2013). Lucky travelers score $6.99 tickets to Hawaii after Delta glitch. MarketWatch. Geraadpleegd via: https://www.marketwatch.com/story/lucky-travelers-score-699-tickets-to-hawaii-after-delta-glitch-1388138286.

Koshiyama, A. & Engin, Z. (2019), 'Algorithm Impact Assessment: Fairness, Robustness and Explainability in Automated Decision-Making', 10-12-2019, https://zenodo.org/record/3241980.

Lauret, J. (2019). Amazon's sexist AI recruiting tool: how did it go so wrong?. Geraadpleegd via: https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e.

Leicht-Deobald et al. (2019). The challenges of Algorithm-Based HR Decision-Making for Personal Integrity. *Journal of Business Ethics* 160:377-392.

Lőrincz, Z. (2019). A brief overview of Imitation Learning. Geraadpleegd via: https://medium.com/@SmartLabAI/a-brief-overview-of-imitation-learning-8a8a75c44a9c.

Büchi, M. et al. (2020). The chilling effects of algorithmic profiling: Mapping the issues. Computer Law & Security Review, 36, 105367.Malik, Farhad.,' Neural Network Layers. Understanding How Neural Network Layers Work', 18-05-2019, https://medium.com/fintechexplained/neural-network-layers-75e48d71f392.

Mindstrong Health (s.a.), How it works. Geraadpleegd via: https://mindstrong.com/how-it-works/.

Mitchell, M. et al. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

Mitchell, S., Potash, E., Barocas, S., D'Amour, A. (2020), Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions

Mittelstadt, B. D. et al. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).

Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Geraadpleegd via: https://christophm.github.io/interpretable-ml-book/

OESO (2018), Personalised Pricing in the Digital Era. DAF/COMP(2018)13, p. 9.

PDPC (2019), Model AI Governance Framework. Geraadpleegd via: https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework

PDPC (2018). Discussion Paper on Artificial Intelligence (AI) and Personal Data - Fostering Responsible Development and Adoption of AI.

Pessach et al. (2020). Employees Recruitment: A prescriptive analytics approach via machine learning and mathematical programming. Geraadpleegd via: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7252110/.

Fountain, N. (producer). (2015). Planet money Episode 633: The Birth and Death of The Price Tag [Podcast]. Geraadpleegd via: https://www.npr.org/sections/money/2015/06/17/415287577/episode-633-the-birth-and-death-of-the-price-tag?t=1599817295236

Poort, J. & Zuiderveen Borgesius, F. J. (2019). Does everyone have a price? Understanding people's attitude towards online and offline price discrimination. Internet Policy Review, 8(1). DOI: 10.14763/2019.1.1383

Pringle, R., Michael, K., & Michael, M. G. (2016). Unintended Consequences of Living with AI: The Paradox of Technological Potential? Part II [Guest Editorial]. IEEE Technology and Society Magazine, 35(4), 17-21.

PSI Testing Intelligence, (2015). 4 Reasons Why an Automated Hiring Process Will Help Your Company. Geraadpleegd via: https://blog.psionline.com/talent/4-reasons-why-an-automated-hiring-process-will-help-your-company.

Purtill, C. (2020). Algorithms learn our workplace biases. Can they help us unlearn them. New York Times. Geraadpleegd via: https://www.nytimes.com/2020/03/10/us/algorithms-learn-our-workplace-biases-can-they-help-us-unlearn-them.html

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Riley, C. (2017). Uber criticized for surge pricing after London terror attack . CNN. Geraadpleegd via: https://money.cnn.com/2017/06/04/technology/uber-london-attack-surge-pricing/index.html.

Schraer, R. (2020). Depression doubles during coronavirus pandemic. BBC. Geraadpleegd via: https://www.bbc.com/news/health-53820425.

Sciforce (2020). Introduction to the White-Box AI: The Concept of Interpretability'. Geraadpleegd via: https://medium.com/sciforce/introduction-to-the-white-box-ai-the-concept-of-interpretability-5a31e1058611.

Sedee, M. (2020). Derde Nederlander krijgt 'X' in paspoort. NRC.nl. Geraadpleegd via: https://www.nrc.nl/nieuws/2020/02/28/derde-nederlander-krijgt-x-in-paspoort-a3992149.

Seele, P. et al. (2019). Mapping the Ethicality of Algorithmic Pricing: A Review of Dynamic and Personalized Pricing. Journal of Business Ethics, 1-23.

Sennaar, K. (2019). Machine Learning for Recruiting and Hiring – 6 current applications. Emerj. Geraadpleegd via: https://emerj.com/ai-sector-overviews/machine-learning-for-recruiting-and-hiring/

Steen, M. (2020). Discussie over de transparantie van algoritmen blijft nodig. Het Parool. Geraadpleegd via: https://www.parool.nl/columns-opinie/discussie-over-de-transparantie-van-algoritmen-blijft-nodig~b882ed5f/.

Strack, R. et al. (2012). From Capability to Profitability: Realizing the Value of People Management. BCG. Geraadpleegd via: https://image-src.bcg.com/Images/BCG_From_Capability_to_Profitability_Jul_2012_tcm9-103684.pdf

Sung, H. et al. (2019). Global patterns in excess body weight and the associated cancer burden. CA: a cancer journal for clinicians, 69(2), 88-112.

The Alphastar Team (2019). AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. Deepmind. Geraadpleegd via: https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii.

The Verge (2020). Amazon announces halo, a fitness band and app that scans your body and voice https://www.theverge.com/2020/8/27/21402493/amazon-halo-band-health-fitness-body-scan-tone-emotion-activity-sleep

Leclercq, F. et al. (2020). Perfectly parallel cosmological simulations using spatial comoving Lagrangian acceleration. arXiv preprint arXiv:2003.04925.

Van Dijck, J. (2019). Digitale personalisatie mag solidariteit en sociale zekerheid niet ondermijnen. FD. Geraadpleegd via: https://fd.nl/opinie/1311018/digitale-personalisatie-mag-solidariteit-en-sociale-zekerheid-niet-ondermijnen

Voss, P. (2016). Why Machine Learning won't cut it'. Geraadpleegd via: https://medium.com/@petervoss/why-machine-learning-wont-cut-it-f523dd2b20e3#.wifeugkuq.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Colum. Bus. L. Rev., 494.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.

Wan, W. (2020). The coronavirus pandemic is pushing America into a mental health crisis. The Washington Post. Geraadpleegd via: https://www.washingtonpost.com/health/2020/05/04/mental-health-coronavirus/.

Wechsler, D. (1958). The Measurement and Appraisal of Adult Intelligence. Baltimore, MD: The Williams & Wilkins Company

Whitworth, B., & Ahmed, A. (2020). Socio-technical system design. The Encyclopedia of Human-Computer Interaction, 2nd Ed.

Wiggers, K. (2020). Yann LeCun and Yoshua Bengio: Self-supervised learning is the key to human-level intelligence. Geraadpleegd via: https://venturebeat.com/2020/05/02/yann-lecun-and-yoshua-bengio-self-supervised-learning-is-the-key-to-human-level-intelligence/.

Wikipedia (s.a.), 'Naturalistische Dwaling'. Geraadpleegd via: https://nl.wikipedia.org/wiki/Naturalistische_dwaling.

Wilson, G. (2017). The Intersection of Internal and External Audit. Workiva. Geraadpleegd via: https://www.workiva.com/sites/workiva/files/pdfs/thought-leadership/intersection-of-internal-and-external-audit-greg-wilson-white-paper-20170619-j5998.pdf.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). London, UK: Springer-Verlag.

World Bank Group. (2019). Credit Reporting Knowledge Guide 2019. Washington: World Bank Group.

World Health Organization (2020). Depression. Geraadpleegd via: https://www.who.int/news-room/fact-sheets/detail/depression.

Wysa (s.a.). Question most asked by users. Geraadpleegd via: https://www.wysa.io/faq.

Yin, R. K. (2009). Case study research, design and methods, fourth edition. Sage publishing

14 Bijlagen

Bijlage 1: Overzicht geïnterviewden

In het kader van dit onderzoek zijn de volgende personen geïnterviewd. Enkele van hen hebben er voor gekozen niet met naam en toenaam genoemd te worden in dit overzicht.

- Piet Hein van Dam, CEO, Clear.
- Sven Schagen, Director Data Science, Albert Heijn
- Jurriaan Nagelkerke, Principal Consultant Analytics, Cmotions
- Tertia Wiedenhof, Product Owner People Analytics & Insights, Rabobank
- Anne Doeser, Lead HR Innovation Hub, Rabobank
- Ruud Schmeink, CEO MarketRedesign
- Paul de Bijl, Owner & Principal, Radicand Economics
- CEO, Ontwikkelaar pricing algoritmen
- Director Analytics, Wereldwijde aanbieder van (krediet) informatie diensten en producten
- Director Communications, Wereldwijde aanbieder betaaldiensten

De inzichten, meningen en uitspraken van de voor dit onderzoek geïnterviewde experts geven niet noodzakelijk de standpunten weer van de organisatie waar ze werkzaam zijn.

Bijlage 2: Leidraad interviews

In het kader van dit onderzoek zijn semi-gestructureerde interviews gehouden met experts en belanghebbenden. Onderstaand is de leidraad die wij hebben gehanteerd bij deze interviews.

Inzicht in casus

Omschrijving van de casus

- Wat is [casus]?
- Wat is het doel van [casus]?
- Wat is (grofweg) de werking van [casus]?
- Wat is de meerwaarde van AI voor [casus]?

Technische aspecten en ontwikkeling

- Wat is de operationele context van [casus]?

Technische Aspecten

- Welke typen soort modellen/algoritmen worden gebruikt?
- Hoe worden modellen toegepast in bedrijfscontext?
    - Applicaties
    - Rol van operators

Onvoorziene effecten

- Welke onvoorziene/onverwachte effecten kunnen optreden bij de toepassing
    - Voor de verantwoordelijke organisatie
    - voor eindgebruiker/operator
    - voor subject
    - voor de maatschappij
- Wat is de oorzaak van deze effecten?
- Wat is het effect van tijd op deze effecten?
    - Op de korte, middellange, lange termijn

Inzicht in proces

Ontwikkeling en test

- Welke actoren zijn betrokken bij de ontwikkeling van de toepassing?
- Welke processen/structuren worden bij ontwikkeling gevolgd
    - Agile
    - Crisp-dm
    - …
- Welk beleid/regels/procedures spelen een rol?
    - (pre-) DPIA/GEB

    - AI Impactassessments
    - Human Rights Impactassessments
    - ….
- Hoe verhoudt het proces van model/algoritme ontwikkeling zich tot andere elementen in het software ontwikkelingsproces?

Acceptatie

- Wat is het proces van acceptatie?
- Wat wordt er getest in de acceptatiefase?

Inzet en onderhoud

- Welke actoren zijn betrokken bij het in productie nemen van de toepassing?
- Hoe verloopt de inzet van de toepassing in de praktijk?
- Welke maatregelen/beheer zijn van toepassing?
- Hoe wordt de beschikbaarheid, betrouwbaarheid, en continuïteit van de toepassing geborgd?
- Wat is de levensloop van het model?
    - Monitoring
    - Audits
    - End-of-life

Inzicht in risicomanagement

Identificatie risico's

- Hoe wordt tijdens de ontwikkeling nagedacht en omgegaan met mogelijke risico's van de toepassing?
- Welke procedures/beleid/methoden spelen een rol?

Risico inschatting/identificatie onvoorziene effecten

- Op welke manier worden tijdens de ontwikkeling met potentiele risico's geïdentificeerd?
- In welke stappen in het proces van ontwikkelen en inzetten van treden risico's op? (Data, modelleren, inzet, gebruik
- Op welke manier worden de perspectieven (eind-)gebruikers/subjecten onderzocht en meegenomen?
- Op welke manier worden (eind-)gebruikers/subjecten geïnformeerd over risico's, maatregelen van de toepassing?

Omgaan met mogelijke risico's onvoorziene effecten

- Op welke manier wordt er omgegaan met risico's?
    - Risico Assessments,
    - Interface voor verhaal/Terugdraaien
    - Intern toezicht/Audit

- Extern toezicht

Omgang met restrisico's

- Niet elk risico (op een onvoorzien effect) is altijd volledig weg te nemen, hoe wordt omgegaan met de restrisico's?

Toekomstige ontwikkelingen

- op de korte, middellange, en lange termijn voor [casus]

CONSIDERATI