KU LEUVEN

FACULTY OF
ENGINEERING SCIENCE

De zwarte doos openen

Het effect van SHAP en counterfactuals op het vertrouwen in en begrip van black-boxmodellen bij AI-novices

Jeffrey Quicken

Thesis voorgedragen tot het behalen
van de graad van Master of Science
in de ingenieurswetenschappen:
computerwetenschappen, hoofdoptie
Artificiële intelligentie

Promotoren:

Prof. dr. Katrien Verbert
Dr. ir. Robin De Croon

Assessoren:

ir. Thierry Deruyttere
Dr. ir. Koen Yskout

Begeleider:

Jeroen Ooge

Academiejaar 2021 – 2022

Voorwoord

Het schrijven van deze masterproef was met momenten een zeer zware, maar toch interessante opgave. Mijn dank gaat uit naar mijn begeleider, Jeroen, voor de manier waarop hij mij heeft bijgestaan met raad en daad om het beste uit mijn masterproef te halen. Zonder zijn aandacht voor detail en zijn open en transparante communicatie zou het eindresultaat niet hetzelfde zijn geweest. Ook bedankt aan Prof. Dr. Katrien Verbert en Dr. ir. Robin De Croon, en bij uitbreiding de leden van de HCI-onderzoeksgroep, voor de feedback gedurende het hele traject.

Ik wil ook van deze gelegenheid gebruik maken om mijn ouders te bedanken om me de kans te geven deze opleiding tot een goed einde te brengen en altijd in mij te blijven geloven. Ook wil ik mijn familie en vrienden te bedanken. Zonder hun steun en motivatie had ik de eindmeet niet gehaald. Speciale dank gaat uit naar mijn vriendin, Carolien, die een heel jaar lang haar best heeft gedaan om me te motiveren en ervoor te zorgen dat ik niets tekort kwam tijdens het werken aan deze masterproef.

Jeffrey Quicken

Inhoudsopgave

Samenvatting

De opkomst van Artificiële Intelligentie leidt steeds vaker tot het gebruik van black-boxalgoritmen. Dit zijn algoritmen met vaak veel voorspellingskracht, met als nadeel dat ze een verborgen interne werking hebben en bijgevolg vaak leiden tot onverklaarbare voorspellingen en beslissingen. Hoewel niet alle toepassingen van dit soort AI-systemen een grote impact hebben op ons dagelijks leven zijn er ook kritische toepassingen van deze systemen waarbij dit wel zo is. Hiervoor is er nood aan interpreteerbaarheid. Om deze interpreteerbaarheid te bekomen kan er gebruik worden gemaakt van explanationmethodes. Dit zijn technieken die verschillende soorten verklaringen genereren gebaseerd op het gebruikte AI-systeem. In deze masterproef worden twee state-of-the-art methodes voor lokale, post-hoc explanations geanalyseerd: Shapley additive explanations (*SHAP*) en counterfactual explanations (*CF*). Meer bepaald wordt het vertrouwen in en het begrip van een black-boxmodel bij AI-novices onderzocht aan de hand van een gebruikersstudie. Uit de kwantitatieve resultaten van deze studie blijkt dat het gebruik van SHAP niet leidt tot meer vertrouwen in of meer begrip van het AI-systeem bij AI-novices. Het toevoegen van counterfactual explanations als ondersteuning bij SHAP leidt wel tot meer vertrouwen in het AI-systeem bij AI-novices, maar niet tot een beter begrip van het AI-systeem. Uit een thematische analyse blijkt dat een groot deel van de AI-novices de SHAP-explanations niet begrijpt en onduidelijk vindt. De explanations gaan vaak in tegen de verwachtingen van de gebruikers, wat kan zorgen voor een gebrek aan vertrouwen. De counterfactual explanations blijken duidelijker te zijn voor AI-novices. De gebruikers geven ook aan dat ze inzichten opdoen dankzij de counterfactual explanations, hoewel dit niet blijkt uit de kwantitatieve resultaten.

Hoofdstuk 1

Introductie

Artificiële Intelligentie (AI) is een steeds belangrijker wordend thema in onze huidige maatschappij. Hoewel het AI-domein reeds vele jaren bestaat, is er recent een grote toename in de toepassing van AI-systemen dankzij o.a. de verhoogde toegankelijkheid van frameworks waarbij state-of-the-art algoritmen snel geïmplementeerd kunnen worden. Bedrijven proberen steeds vaker uit te pakken met de nieuwste technologie en integreren daarom ook AI-systemen in allerlei toepassingen, gaande van chatbots en aanbevelingssystemen tot systemen die autonoom beslissingen maken. Hoewel niet al deze toepassingen van AI-systemen een grote impact hebben op ons dagelijks leven zijn er ook situaties waarin dit wel zo is, denk bijvoorbeeld maar aan AI-systemen die bepaalde ziektes kunnen diagnosticeren. Belangrijk is om bij zulke systemen aandacht te geven aan de problemen en uitdagingen die gepaard gaan bij het gebruik van AI-algoritmen.

Een van de belangrijkste uitdagingen is het overkomen van het gebrek aan interpreteerbaarheid van AI-systemen. Dit is één van de uitdagingen waarvoor men oplossingen zoekt binnen het domein van de Explainable Artificial Intelligence (XAI). Hiervoor zijn er reeds vele methodes ontwikkeld, vaak met het oog op gebruikers die reeds ervaring hebben met artificiële intelligentie (*AI-experts*). Door de opkomst van artificiële intelligentie in het dagelijkse leven moet er echter ook aandacht besteed worden aan zulke methodes gericht op gebruikers zonder ervaring (*AI-novices*). In deze masterproef worden twee state-of-the-art methodes uitgelicht en geanalyseerd: Shapley additive explanations ($SHAP$) en counterfactual explanations (CF). Meer bepaald wordt het vertrouwen in en het begrip van een black-boxmodel bij AI-novices onderzocht. Dit gebeurde door een gebruikersstudie uit te voeren waarbij deze methodes toegepast werden op een black-boxsysteem dat de energieprijs in België voorspelt.

In hoofdstuk 2 wordt de relevante literatuur besproken. Het black-boxprobleem wordt toegelicht en belangrijke concepten zoals interpreteerbaarheid en uitlegbaarheid worden gedefinieerd. Ook de nood aan interpretatie en uitleg wordt behandeld, samen met de karakteristieken van een kwalitatieve verklaring. Vervolgens worden de

verschillende doelstellingen en evaluatiecriteria van verklaringen besproken en wordt er een classificatie gegeven van de bestaande verklaringsmethodes. SHAP en CF worden in meer detail toegelicht en de wiskundige fundamenten van beide methodes worden besproken. Er wordt ook een kort overzicht gegeven van de relevante literatuur binnen het energiedomein. Ten slotte wordt het gerelateerd werk besproken. In hoofdstuk 3 worden de onderzoeksvragen voorgesteld samen met de methodes die toegepast werden om het prototype dat gebruikt werd in de gebruikersstudie te ontwikkelen en te evalueren. Hoofdstuk 4 gaat dieper in op het ontwikkelingsproces van het prototype en de verschillende iteraties die doorlopen werden. De resultaten van het onderzoek worden gepresenteerd in hoofdstuk 5 en vervolgens besproken in hoofdstuk 6. Hier worden ook de limitaties van het onderzoek besproken alsook mogelijk toekomstig werk. In het laatste hoofdstuk wordt de conclusie van het onderzoek gevormd.

Hoofdstuk 2

Literatuurstudie

In dit hoofdstuk worden de belangrijkste concepten en bevindingen uit de relevante literatuur toegelicht. Eerst wordt het black-boxprobleem aangekaart. Hierbij wordt er dieper ingegaan op de definitie van interpreteerbaarheid, en waarom er een nood is aan interpretatie en uitleg. Vervolgens wordt er meer informatie gegeven over het domein van Explainable AI (*XAI*). Er wordt bekeken wat de criteria zijn voor een goede verklaring (*explanation*), welke explanationmethodes er bestaan, en wat de doelstellingen van een explanation zijn en hoe we ze kunnen evalueren. Ten slotte wordt er een overzicht gegeven van gerelateerd werk.

2.1 Black-boxprobleem

Een systeem wordt geclassificeerd als black-box wanneer de invoer en uitvoer van het systeem observeerbaar is, maar de interne werking verborgen blijft. Een black-boxsysteem wordt ook wel ondoorzichtig genoemd (*opaque*).

Paul Humphreys [31] definieert ondoorzichtigheid als volgt:

> *"... a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process."*

Hiermee wordt bedoeld dat een proces (of systeem) ondoorzichtig is wanneer het onmogelijk is voor een agent om de elementen in het proces te kennen en te begrijpen. Toegepast op Artificiële Intelligentie kan een AI-algoritme gezien worden als het proces en, afhankelijk van de context, de ontwerper en/of eindgebruiker van het AI-algoritme als de agent.

Een eenvoudig voorbeeld van een black-box is een gedachte-experiment bedacht door de psycholoog B.F. Skinner [64]. De gebruiker krijgt een doos met een reeks ingangen (schakelaars) en een reeks uitgangen (lampjes). Door de ingangen te manipuleren kan de mogelijke uitkomst geobserveerd worden, maar er kan niet in de

doos gekeken worden naar de interne werking. In het meest eenvoudige geval, zoals een rechtstreekse verbinding tussen schakelaar en lamp, kan er met grote zekerheid vastgesteld worden dat de schakelaar de lamp rechtstreeks bedient. Voor een complex systeem daarentegen, zoals de doos van Skinner, is het echter onmogelijk om de interne werking te bepalen door het uitproberen van verschillende combinaties. Zelfs wanneer er in de doos gekeken kan worden, en de interne structuur en componenten waarneembaar zijn kan de werking van de doos niet verklaard worden. De complexiteit ontstaat namelijk uit de interactie van vele eenvoudige componenten. Bij een complex systeem is het zeer onwaarschijnlijk dat de uitvoer van het systeem bepaald kan worden voor een gegeven invoer, zonder het experiment uit te voeren, of andere technieken toe te passen die het systeem uit kunnen leggen.

2.1.1 Interpreteerbaarheid en uitlegbaarheid

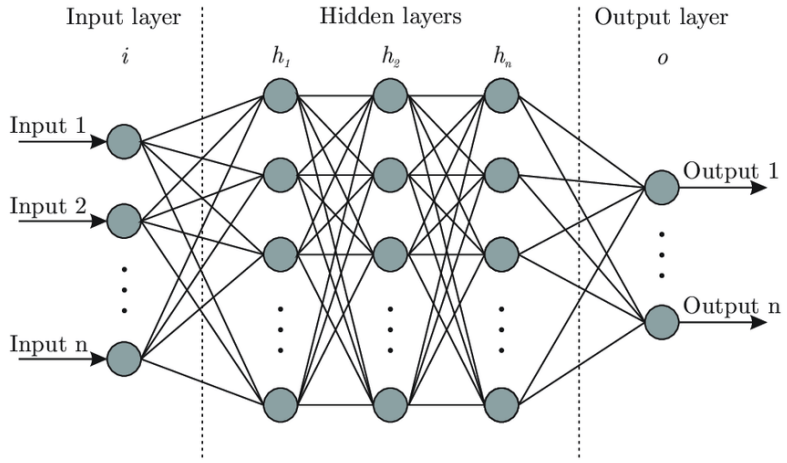
Er is geen consensus in de literatuur over wat de exacte definitie van interpreteerbaarheid nu net is. Doshi-Velez et al. [18] geven de volgende definitie:

> *" Interpret means to explain or to present in understandable terms. In the context of ML systems, we define interpretability as the ability to explain or to present in understandable terms to a human."*

Interpreteerbaarheid wordt hier dus gedefinieerd als het vermogen om in begrijpelijke termen aan de mens uit te leggen of te presenteren. Zhang et al. [74] breiden deze definitie nog verder uit door te stellen dat deze uitleg idealiter bestaat uit logische beslissingsregels (als-dan regels) of in een vorm die hiernaar gereduceerd kan worden. Ze stellen ook dat de begrijpelijke termen waaruit de uitleg moet bestaan moeten voortkomen uit de domeinkennis gerelateerd aan de taak die het algoritme moet vervullen.

In de literatuur valt er ook een subtiel, maar belangrijk verschil op te merken tussen interpreteerbaarheid (*interpretability*) en uitlegbaarheid (*explainability*). Gilpin et al. [24] zien uitlegbaarheid als interpreteerbaarheid gecombineerd met volledigheid (*completeness*). Hiermee wordt bedoeld dat uitlegbaarheid slechts bereikt wordt indien de explanations interpreteerbaar zijn én accuraat zijn met betrekking tot het model dat ze verklaren. De uitdaging bestaat erin de juiste balans tussen deze twee te vinden. Neem als voorbeeld een neuraal netwerk zoals in fig. 2.1. De structuur van een uitgebreid neuraal netwerk is zeer moeilijk te interpreteren, maar men kan wel altijd de uitvoer verklaren aan de hand van een wiskundige formule die het verband tussen de invoer en uitvoer beschrijft. Dit is dus een volledige uitleg, maar niet per se interpreteerbaar aangezien deze formules zeer uitgebreid en complex zijn.

In de huidige wereld van AI zijn er slechts een aantal algoritmen die beschouwd kunnen worden als inherent doorzichtig, en dus te interpreteren zijn voor de mens.



Figuur 2.1: Voorbeeldstructuur van een Artificiëel neuraal netwerk, uit Bre et al.[10].

Voorbeelden hiervan zijn lineaire regressie, beslissingsbomen (*decision trees*), regelgebaseerde machine learning (*rule-based machine learning*) en Bayesiaanse Netwerken (*Bayesian Networks*) [48, 40]. Het grootste deel van de huidige algoritmen is echter niet inherent doorzichtig. Men kan het hele algoritme inspecteren en de werking van ieder afzonderlijk element in isolatie begrijpen, zonder de werking van het hele algoritme te kunnen begrijpen en interpreteren. Men weet enkel dat men voor een bepaalde invoer een bepaalde uitvoer zal krijgen. Voorbeelden van zulke black-boxalgoritmen zijn neurale netwerken, support vector machines en gradient boosting methodes.

Wanneer we de werking van een AI-algoritme willen begrijpen, willen we zien hoe het algoritme redeneert voor een bepaalde invoer en waarom het tot een bepaalde uitvoer komt. De link tussen de invoer en de uitvoer van een AI-algoritme wordt ook wel het model genoemd. Wanneer we dit model begrijpen en kunnen verklaren zien we het algoritme niet meer als black-box.

Dat interpretatie van de interne werking van een AI-algoritme niet altijd mogelijk is kan verschillende oorzaken hebben. De interne werking kan verborgen zijn en dus niet door de gebruiker worden geobserveerd. Een andere mogelijkheid is dat de interne werking en structuur van het algoritme niet interpreteerbaar is. Dit is bijvoorbeeld het geval bij neurale netwerken, waarbij het wel mogelijk is om de interne werking te bekijken, maar de interne structuur zo uitgebreid en complex kan zijn dat de interpretatie ervan niet meer mogelijk is. In paragraaf 2.1.2 wordt er dieper ingegaan op waarom de interpretatie van het algoritme en het uitleggen van een beslissing of voorspelling van het algoritme net zo belangrijk is.

2.1.2 De nood aan interpretatie en uitleg

Er zijn meerdere redenen waarom er nood is aan interpretatie en uitleg in de context van black-box AI-algoritmes. Lipton [39] stelt de volgende vijf doelen voor

interpreteerbaarheid voor: vertrouwen creëren, correlaties tot causaties generaliseren, de kwaliteit van het algoritme evalueren, informatie verschaffen en rechtvaardigheid en ethische correctheid afdwingen. Adadi et al. [2] brengen de verschillende motivaties voor de nood aan interpretatie (en dus aan XAI) onder in 4 domeinen: uitleggen om te rechtvaardigen, uitleggen om te controleren, uitleggen om te verbeteren en uitleggen om te ontdekken. In wat volgt gaan we dieper in op de doelen van interpreteerbaarheid zoals gedefinieerd door Adadi et al. [2].

Uitleggen om te rechtvaardigen

Tegenwoordig maken veel bedrijven gebruik van Artificiële Intelligentie voor beslissingsprocessen. Denk bijvoorbeeld maar aan Amerikaanse banken die AI-algoritmes gebruiken om de kredietscore van hun (potentiële) klanten te berekenen en op basis van het resultaat al dan niet een lening verlenen [20]. Het delegeren van dit soort beslissingen aan black-boxalgoritmen zonder de mogelijkheid tot het verkrijgen van een interpreteerbare verklaring kan leiden tot verregaande gevolgen zoals discriminatie, gebrek aan vertrouwen en gebrek aan verantwoording. De voorbije jaren zijn er reeds meerdere controversiële toepassingen van AI de revue gepasseerd.

Een recent geval van geautomatiseerde discriminatie door gebruik van een AI-algoritme werd ontdekt door journalisten van propublica.org [37]. De COMPAS-score, een voorspellend model dat het risico op het hervallen naar criminele activiteiten weergeeft, classificeert mensen met Afrikaanse roots twee keer zoveel als blanken als een hoog-risicogeval, terwijl COMPAS de tegenovergestelde fout maakt bij blanken.

Een onderzoek aan Princeton [13] toont aan hoe teksten en webpagina's menselijke vooroordelen (*bias*) bevatten: namen typisch geassocieerd met personen van Afrikaanse roots komen vaker voor in combinatie met woorden die een negatieve connotatie oproepen in vergelijking met blanke personen. Dit heeft als resultaat dat modellen die men traint op data afkomstig uit deze teksten de vooroordelen mogelijk overnemen. Het interpreteerbaar zijn van een model kan helpen om te evalueren of het model al dan niet uitgaat van vooroordelen, en kan er dus, indien correct gebruikt, voor zorgen dat onder meer etnische vooroordelen geen invloed hebben op het beslissingsproces.

Deze voorbeelden maken duidelijk dat er steeds meer nood is aan verklaringen binnen het domein van AI. Wanneer het in deze context gaat over verklaringen van een beslissing genomen door een algoritme, gaat het om het verklaren en rechtvaardigen van een bepaalde uitkomst van dat algoritme, en niet om een interpretatie of beschrijving van de innerlijke werking van het algoritme.

Om te evalueren of beslissingen van een geautomatiseerd systeem op een eerlijke en ethische manier genomen werden, zonder de hierboven vernoemde vooroordelen, kan er gebruik worden gemaakt van Explainable AI. In deze context kan XAI zorgen voor voldoende informatie om de beslissing van het algoritme mee te rechtvaardigen en om inzicht te geven in het beslissingsproces wanneer er onverwachte beslissingen genomen worden.

Ook wetgeving speelt steeds vaker in op het uitleggen om te rechtvaardigen. De Algemene Verordening Gegevensbescherming (*Engels: General Data Protection*

Regulation (GDPR)) is de geldende wetgeving binnen de Europese Unie die bepaalt hoe de verwerking van persoonsgegevens door bedrijven en organisaties binnen de EU mag plaatsvinden. Artikel 22 en overweging 71 van deze verordening stellen dat het data-subject (de persoon wiens data gebruikt wordt door het algoritme en waarop de beslissing betrekking heeft) recht op uitleg heeft wanneer er door middel van een geautomatiseerd proces beslissingen worden genomen [17], al is er enige discussie over de effectiviteit en toepasbaarheid van deze wet in de context van Artificiële Intelligentie [69].

Uitleggen om te controleren

Het kunnen uitleggen en interpreteren van een algoritme kan ook leiden tot het ontdekken van problemen en kwetsbaarheden in het algoritme. Interpretatie is dus noodzakelijk om de controle over het algoritme en de kwaliteit van de gegenereerde voorspellingen of beslissingen niet te verliezen.

Uitleggen om te verbeteren

Niet enkel rechtvaardiging speelt een belangrijke rol bij de nood aan interpretatie. Ook het continu streven naar verbeteringen van een algoritme is een van de redenen waarom er nood is aan interpretatie en uitleg. Een algoritme dat interpreteerbaar en verklaarbaar is, is voor de ontwerper makkelijker te verbeteren. Ook voor eindgebruikers kan dit leiden tot verbeteringen. Gebreken en vooroordelen in data komen naar boven door interpretatie en dat leidt uiteindelijk tot het verbeteren van het model.

Uitleggen om te ontdekken

In 1996 heeft Deep Blue, een Artificiëel Intelligent systeem van IBM, schaakgrootmeester Garry Kasparov verslagen [14]. Dit was het eerste geval van een autonoom artificieel intelligent systeem dat een menselijke speler heeft kunnen verslaan in een schaaktoernooi. Meer recent heeft AlphaGo, een AI die het complexe spel Go speelt, de Europese Go kampioen verslagen [63]. XAI kan de mens helpen inzicht te krijgen in verbanden en strategieën die deze spelsystemen autonoom hebben ontdekt. Het is niet ondenkbaar dat er in de nabije toekomst AI-systemen kunnen zijn die nog voor ons onbekende verbanden en wetmatigheden in de natuurwetenschappen vinden. Interpretatie en uitleg van deze modellen kan dan leiden tot een beter begrip van de wereld om ons heen.

Er kan dus besloten worden dat XAI een belangrijk onderzoeksdomein is binnen het overkoepelende domein van Artificiële Intelligentie. Het kan helpen zorgen voor rechtvaardigheid in beslissingen, verbetering van algoritmes en het verkrijgen van nieuwe inzichten in de problemen die ons dagdagelijks bezig houden. Dit alles op zijn beurt leidt dan weer tot AI-systemen die betrouwbaarder zijn en meer vertrouwen krijgen van gebruikers. Hoe het vertrouwen van gebruikers kan gemeten worden en hoe het zich verhoudt tot XAI wordt behandeld in paragraaf 2.2.2.

2.2 Explanations

Om onderzoek te verrichten binnen het domein van XAI wordt er eerst gedefinieerd wat er net verstaan wordt onder explanations en aan welke eigenschappen ze moeten voldoen. Ook de manier waarop explanations kunnen geëvalueerd worden wordt toegelicht. Vervolgens worden de verschillende soorten explanations en hoe ze geclassificeerd kunnen worden toegelicht en wordt er dieper in gegaan op de twee explanationmethodes die gebruikt zullen worden in deze masterproef: SHapley additive explanations *SHAP* en counterfactual explanations *CF*.

2.2.1 Waaruit bestaat een goede explanation

Binnen de onderzoeksgemeenschap bestaat er geen consensus over wat gezien wordt als een effectieve verklaring. Er zijn wel een aantal karakteristieken die wenselijk zijn en algemeen aanvaard worden als noodzakelijk voor een goede uitleg. Deze karakteristieken komen niet louter uit het veld van de computerwetenschappen, maar komen vaak ook voort uit andere onderzoeksgebieden zoals de sociale wetenschappen. Miller [44] stelt dat onderzoek binnen de XAI-gemeenschap zich nog te vaak beperkt tot technische concepten en definities om explanations op te baseren. Het onderzoek naar explanations zou volgens Miller [44] moeten vertrekken vanuit de sociale karakteristieken van explanations, en zich dus baseren op hoe mensen onderling beslissingen aan elkaar uitleggen.

Dit hoofdstuk baseert zich op Schneider et al. [61], die zeven criteria voorstellen waarmee rekening gehouden moet worden om een goede explanation te verkrijgen: Betrouwbaarheid, generaliseerbaarheid, verklarende kracht, interpreteerbaarheid, inspanning, privacy en eerlijkheid.

Betrouwbaarheid

Zoals eerder behandeld in paragraaf 2.1.2 is het nodig om een AI-systeem te kunnen interpreteren wanneer het gebruikt wordt om beslissingen te maken die verregaande gevolgen hebben. Neem terug het voorbeeld van de COMPAS score. Het is van belang dat het model hierachter interpreteerbaar is, maar ook dat de explanation die zorgt voor de interpreteerbaarheid betrouwbaar is. Aangezien een van de doelen van interpreteerbaarheid is om het model te kunnen controleren en te zien of het uit gaat van vooroordelen is deze betrouwbaarheid een noodzakelijkheid. Indien de betrouwbaarheid van de explanation onvoldoende groot is kan dit leiden tot incorrecte conclusies over het systeem en een gebrek aan vertrouwen in het systeem.

Wanneer de explanation uitgaat van een benadering of vereenvoudiging van de situatie zal de betrouwbaarheid hier vaak onder lijden. De afweging tussen betrouwbaarheid en vereenvoudiging moet altijd gemaakt worden door rekening te houden met het doelpubliek. Wanneer het doelpubliek van de explanation geen of weinig domeinkennis heeft is het vaak wenselijk dat de explanation eenvoudig is.

Die eenvoud komt vaak met als keerzijde een verminderde betrouwbaarheid. Er moet geprobeerd worden de balans tussen de eenvoud en betrouwbaarheid van een explanation af te stemmen op het doelpubliek en het doel van de explanation [45].

Generaliseerbaarheid

Naast betrouwbaarheid is de generaliseerbaarheid van een explanation [56] ook een gewenste karakteristiek. De generaliseerbaarheid van een explanationmethode geeft aan in hoeverre de methode bruikbaar is voor meerdere modellen en systemen. Explanationmethodes kunnen opgedeeld worden in verschillende categorieën die betrekking hebben op voor welke modellen de methode toepasbaar is: modelagnostische methodes [58] en modelspecifieke methodes. Modelagnostiche methodes zijn het meest generaliseerbaar omdat ze toepasbaar zijn op eender welk AI-model, terwijl modelspecifieke methodes slechts voor één of enkele modellen werken. In paragraaf 2.2.3 wordt er dieper ingegaan op het verschil tussen de verschillende categorieën van explanationmethodes.

Verklarende kracht

In de context van XAI bedoelen we met verklarende kracht hoeveel verschijnselen de explanation kan verklaren. Dit komt overeen met het aantal verschillende vragen dat de explanation kan beantwoorden. Voorbeelden hiervan zijn: waarom heeft het model deze voorspelling gemaakt?, Welke factoren dragen het meeste bij tot de beslissing van het model?, Is het model rechtvaardig?, Waarom is de beslissing van het systeem anders dan we verwachten?, Wanneer faalt het model en waarom?, ... [56].

Interpreteerbaarheid

De interpreteerbaarheid van een explanation is een belangrijk element dat leidt tot een kwaliteitsvolle explanation. Zoals eerder besproken staat interpreteerbaarheid voor de mate waarin iets begrijpbaar is voor mensen [26, 22]. Niet enkel systemen kunnen dus interpreteerbaar zijn, ook de explanations die voor deze interpreteerbaarheid zorgen moeten zelf in zeker mate interpreteerbaar zijn. Een explanation die te abstract is, en daardoor niet te interpreteren is, verliest een groot deel van zijn nut. Ras et al. [56] splitst de interpreteerbaarheid van explanations nog verder op in twee subcomponenten: duidelijkheid en complexiteit. Een explanation moet zo duidelijk en ondubbelzinnig mogelijk zijn. Om dat te bereiken moet de complexiteit van de explanation zo laag mogelijk gehouden worden en afgestemd worden op het doelpubliek. Net zoals de afweging tussen betrouwbaarheid en complexiteit in paragraaf 2.2.1 moet er hier ook een afweging gemaakt worden tussen interpreteerbaarheid en complexiteit.

Inspanning

Gerelateerd aan de complexiteit en de interpreteerbaarheid van een explanation is er ook nog de inspanning die een gebruiker moet leveren om de explanation te interpreteren. Hier geldt vaak dat hoe complexer de explanation is, hoe meer inspanning de gebruiker moet leveren. Recent onderzoek van Hudon et al [30] heeft uitgewezen dat een te grote inspanning leidt tot een vermindering in transparantie en vertrouwen in het systeem. Ook hier moet er dus voor gezorgd worden dat de explanation afgestemd wordt op het doelpubliek.

Privacy

Privacy speelt voornamelijk een belangrijke rol wanneer het gaat over gepersonaliseerde explanations. Dit zijn explanations waarbij er gebruik wordt gemaakt van data van de gebruiker, of data van andere gebruikers, om de explanation aan te passen. Neem terug het voorbeeld van de kredietscore. Indien explanations van dit systeem gegeven worden door bijvoorbeeld data van andere gebruikers te tonen die behoren tot dezelfde categorie als de gebruiker moet er voorzichtig omgegaan worden met deze persoonlijke data. Het zou kunnen dat indien er te veel of te specifieke informatie gegeven wordt, de betrokken gebruikers kunnen geïdentificeerd worden. Een kwaliteitsvolle explanation houdt zich dus aan het wettelijk kader en zorgt ervoor dat de veiligheid van persoonlijke data niet in het gedrang komt.

Rechtvaardigheid

Net zoals het noodzakelijk is dat AI-systemen rechtvaardig zijn in het nemen van beslissingen moeten ook explanations een vorm van rechtvaardigheid kunnen bieden. Het streefdoel is dat de explanations volledig egalitair zijn [35, 8]. Dit betekent dat twee verschillende gebruikers van hetzelfde systeem dezelfde kwaliteit van explanation moeten ontvangen met betrekking tot de hierboven gedefinieerde karakteristieken.

De zeven karakteristieken hierboven gedefinieerd zijn niet altijd duidelijk af te lijnen. De kwaliteit van een explanation is een samenspel van meerdere factoren die onderling ook nog eens een invloed hebben op elkaar. Belangrijk om te onthouden is dat de explanation altijd afgestemd moet zijn op het doel en het doelpubliek.

2.2.2 Doelstellingen en evaluatiecriteria

In een recent onderzoek naar explanations en de kwaliteit ervan komen Nunes et al. [52] tot de volgende bemerking:

> *"... There is no standard design for user studies that evaluate forms of explanations."*

Inderdaad, net zoals er geen consensus is over wat een explanation nu exact moet bevatten, is er ook geen standaard die toelaat om de kwaliteit van een explanation te evalueren. Waar onderzoekers zoals Miller [44], Chromik et al. [16], Lage et al. [36],

Mohseni et al. [47] en Hoffman et al.[28] het wel over eens zijn, is dat de kwaliteit van een explanation (en de bijhorende explanationmethode) niet enkel geëvalueerd kan worden op basis van de technische karakteristieken van de explanationmethode. Er moet voor de evaluatie ook gekeken worden naar het gedrag van de gebruiker: stelt de explanation de gebruiker in staat om bepaalde taken beter uit te voeren, begrijpen de gebruikers dankzij de explanation waarom het model zich op een bepaalde manier gedraagt, creëert de explanation vertrouwen in het model. Deze masterproef baseert zich op het framework voorgesteld door Mohseni et al. [47]. Dit framework deelt de evaluatie op in een aantal dimensies. Eerst moet het doelpubliek bepaald worden. Mohseni et al. [47] delen gebruikers op in drie verschillende categorieën: AI-novices (eindgebruikers van AI-systemen die nog geen of zeer beperkte ervaring hebben met Artificiële intelligentie), data-experts (gebruikers die regelmatig gebruik maken van AI-systemen voor analyse en onderzoek) en AI-experts (onderzoekers en AI-wetenschappers die AI-systemen en algoritmen ontwerpen).

Elke categorie van gebruikers heeft zijn eigen doelstellingen en evaluatiecriteria voor explanationmethodes zoals te zien in fig. 2.2. In wat volgt wordt er dieper ingegaan op de doelstellingen en evaluatiecriteria van AI-novices aangezien dit de doelgroep is van het onderzoek in deze masterproef. De doelstellingen en evaluatiecriteria van data-experts en AI-experts worden verder buiten beschouwing gelaten.

Doelstellingen

Mohseni et al. [47] stellen vier verschillende doelstellingen voorop voor explanationmethodes gericht op AI-novices: Algoritmische Transparantie (de eindgebruiker helpen begrijpen hoe het systeem werkt), Vertrouwen en Betrouwbaarheid (het vertrouwen van de eindgebruiker in het systeem verhogen), Beperking van vooroordelen (de eindgebruiker helpen bepalen of het systeem bevooroordeeld is) en Privacy (eindgebruikers in staat stellen de data-privacy van het algoritme te beoordelen)

Deze doelstellingen komen overeen met een subset van de criteria voor een goede explanation, vooropgesteld door Schneider et al. [61] (respectievelijk interpreteerbaarheid, betrouwbaarheid, eerlijkheid en privacy). Ook hier geldt dat deze doelstellingen enigszins overlappen en een invloed hebben op elkaar.

Evaluatiecriteria

Gekoppeld aan de doelstelling van de explanationmethodes is de manier waarop het behalen van deze doelstelling kan geëvalueerd worden. Mohseni et al. [47] geven de volgende vier evaluatiecriteria:

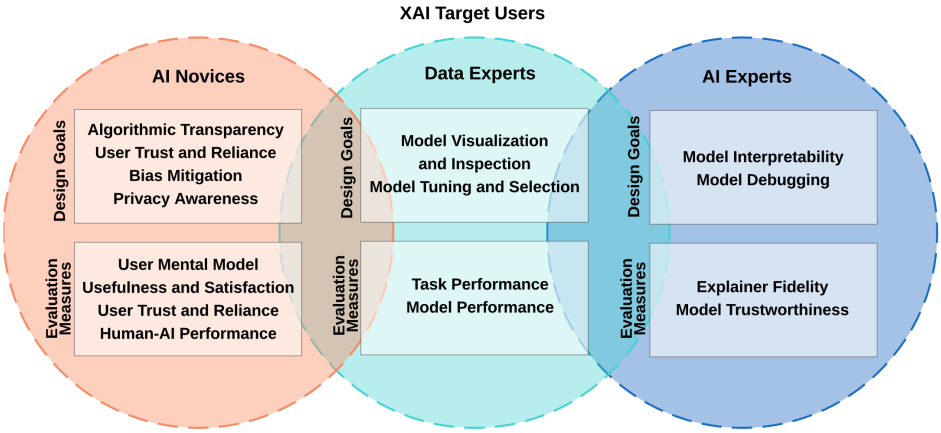
Mentaal model Het mentaal model van een gebruiker is een concept dat voortkomt uit de cognitieve psychologie. Het is de representatie van hoe gebruikers een systeem begrijpen. Evaluatie van het mentaal model laat toe te begrijpen in hoeverre de eindgebruiker op de hoogte is van de werking van het AI-systeem.

Bruikbaarheid en tevredenheid De tevredenheid van eindgebruikers en de bruikbaarheid van de explanation zijn twee belangrijke factoren bij het evalueren

van explanations. Onderzoek van Gedikli et al. [23] heeft bijvoorbeeld aangetoond dat er een relatie is tussen de tevredenheid van de eindgebruiker en de waargenomen transparantie van het algoritme.

Vertrouwen en betrouwbaarheid Het vertrouwen van de eindgebruiker is een belangrijk concept. Er moet hierbij rekening gehouden worden dat gebruikers al een initieel (gebrek aan) vertrouwen hebben voor ze in aanraking komen met een AI-systeem.

Mens-AI taakperformantie AI-systemen zijn vaak ontworpen om gebruikers bij te staan bij het vervullen van bepaalde taken of het maken van beslissingen. Mens-AI taakperformantie is het evaluatiecriterium waarbij er gekeken wordt naar hoe de explanation de gebruiker ondersteunt bij het vervolledigen van taken en maken van beslissingen.



Figuur 2.2: Gebruikerscategorieën en hun bijhorende doelstellingen en evaluatiecriteria binnen het domein van XAI, uit Mohseni et al. [47].

2.2.3 Classificatie van explanationmethodes

Wanneer we het hebben over explanationmethodes binnen XAI zijn in de huidige state-of-the-art opdelingen in meerdere dimensies mogelijk. Figuur 2.3 geeft een schematische weergave van deze opdeling.

Modelagnostische en modelspecifieke methodes

Modelagnostiche explanationmethodes zijn methodes die toegepast kunnen worden op elk AI-model. Bij dit soort methodes wordt het oorspronkelijke model gezien als een black-box, en zal de explanationmethode post-hoc (na de voorspelling) explanations genereren gebaseerd op de voorspellingen van het black-boxmodel. Hier zijn verschillende methodes voor [57]: een inherent interpreteerbaar model trainen

op de predicties van het oorspronkelijke black-boxmodel [7], het perturberen van de verschillende inputs van het model en analyseren hoe het hierop reageert [34], of een combinatie van de twee zoals toegepast in LIME (*Local Interpretable Model-agnostic Explanations*) [58].

Het grootste voordeel aan deze modelagnostische methodes is dat ze flexibeler zijn in gebruik. Ontwikkelaars en/of gebruikers kunnen gebruikmaken van het gewenste AI-model en vervolgens modelagnostische methodes toepassen om explanations te genereren. Een alternatief voor modelagnostische methodes is gebruik maken van een inherent interpreteerbaar model zoals beslissingsbomen. Bij gebruik van dit soort methodes is er dan geen nood aan explanations die leiden tot interpreteerbaarheid, aangezien de interpreteerbaarheid volgt uit het algoritme zelf. Er wordt dan wel vaak ingeboet aan de accuraatheid van het model, aangezien de meeste van deze methodes minder voorspellende kracht hebben. Ook zijn deze specifiek inherent interpreteerbare modellen niet altijd het gepaste model voor de taak die het moet vervullen.

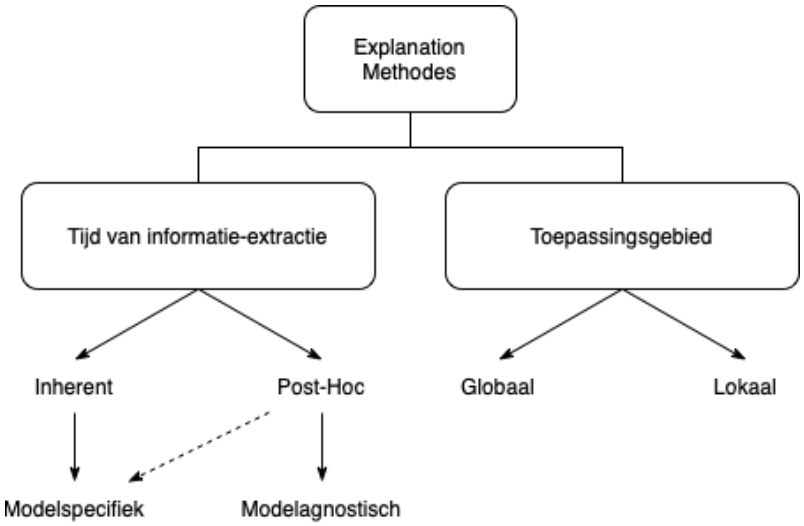
Aan de andere kant van het spectrum bevinden zich de modelspecifieke methodes. In tegenstelling tot de modelagnostiche methodes zijn deze slechts toepasbaar op één enkel AI-model (of een groepering van gelijkaardige modellen). Voorbeelden hiervan zijn methodes die enkel werken voor convolutionele neurale netwerken [60] of enkel werken op gradient boosted tree ensembles, zoals gbt-hips [27].

Globale en lokale methodes

Globale explanationmethodes proberen het hele AI-model dat de voorspelling maakt te beschrijven. De bedoeling van deze methodes is om de volledige logica en werking van het model te duiden en inzicht te geven in de redenering die het algoritme maakt voor alle mogelijke voorspellingen die het model kan maken. Voorbeelden van globale methodes zijn partial dependence plots [21] en accumulated local effects [6].

Lokale explanationmethodes daarentegen zijn methodes die één specifieke voorspelling van een AI-model proberen te verklaren. Deze technieken beantwoorden vragen zoals: waarom heeft het model deze specifieke voorspelling gemaakt?, welk effect heeft deze specifieke input op de voorspelling? en wat zou de voorspelling zijn als ik een kleine aanpassing doe aan de input?. De explanationmethodes die binnen deze categorie vallen zijn onder andere LIME [58], SHAP [42] en counterfactual explanations (CF) [68, 49].

Ook hier geldt weer dat de keuze van explanationmethode afhangt van het doel en het doelpubliek. Een modelontwerper zal vaak gebruik willen maken van globale methodes om het model te evalueren en te begrijpen. Een eindgebruiker daarentegen is in de meeste gevallen meer geïnteresseerd in lokale methodes die de specifieke voorspelling kunnen verklaren die een impact heeft op die eindgebruiker.



Figuur 2.3: Classificatie van explanationmethodes, naar Puiutta et al. [55].

2.2.4 Shapley additive explanations

SHAP is een populaire explanationmethode ontwikkeld door Lundberg et al. [42]. Het is een methode die lokale, modelagnostische post-hoc explanations genereert. De methode ziet iedere inputfactor van een algoritme als een speler, en de voorspelling van het algoritme als de te winnen prijs. Shapley-waarden [1] bepalen dan hoe we de prijs (voorspelling) moeten verdelen over de verschillende spelers (inputfactoren). Dit komt dus overeen met welke bijdrage iedere inputfactor heeft gehad op de voorspelling, en is dus een maat voor het belang van iedere inputfactor (*feature importance*). Hoewel SHAP hevig steunt op wiskundige concepten wordt de uiteindelijke berekening, voor meer efficiëntie, gedaan door benadering. Lundberg et al. [42] stellen meerdere implementaties van SHAP ter beschikking, afhankelijk van de toepassing en het gebruikte model. KernelSHAP is de oorspronkelijke, modelagnostische implementatie. TreeSHAP, DeepSHAP, GradientSHAP en LinearSHAP zijn modelspecifieke implementaties met optimalisaties voor respectievelijk beslisingsbomen, diepe neurale netwerken, gradiëntgebaseerde modellen en lineaire modellen.

Wiskundige achtergrond

In wat volgt wordt er dieper ingegaan op de wiskundige principes die het KernelSHAP algoritme gebruikt, gebaseerd op Molnar [48].
Het doel van KernelShap is om voor een bepaalde voorspelling x, de bijdrage van iedere inputfactor aan deze voorspellling te schatten. Dit gebeurt in vier stappen:

1. Het willekeurig bepalen van coalities $z'_k \in \{0,1\}^M,\ k \in \{1,\ldots,K\}$, met M het aantal inputfactoren van het model en K het gewenste aantal coalities.

   Neem als voorbeeld de coalitievector $z'_k = (1,0,1,1)$. Hier gaat het dus

om een coalitie van de eerste, derde en vierde inputfactor in een model met in totaal vier inputfactoren.

2. Het model een voorspelling laten maken voor elke z'_k, door z'_k om te zetten naar de oorspronkelijke input-vectorruimte door transformatie $h_x \colon \{0,1\}^M \to \mathbb{R}^p$ toe te passen, en vervolgens het model $\hat{f}$ er op toe te passen: $\hat{f}(h_x(z'_k))$.
   De transformatie door h_x gebeurt als volgt: als de waarde in de coalitievector op positie $z'_{ki} = 1$, is de waarde op die positie de waarde van de overeenkomstige inputfactor voor de uit te leggen voorspelling x. Als de waarde in de coalitievector op positie $z'_{ki} = 0$, is de waarde op die positie de waarde van de overeenkomstige inputfactor voor een willekeurige andere voorspelling van het model.

   Neem als voorbeeld terug de coalitievector $z'_k = (1, 0, 1, 1)$, met de inputfactoren voor de uit te leggen voorspelling $x = (20, 26, 13, 2)$. De overeenkomstige getransformeerde inputvector zou dan $h_x(x) = (20, a, 13, 2)$ kunnen zijn met a een willekeurige waarde voor de tweede inputfactor die voorkomt in een voorspelling van het model. $h_x(x)$ wordt dan gebruikt als input voor het model en de bekomen voorspelling $\hat{f}(h_x(x))$ wordt in een later stap gebruikt.

3. Het gewicht voor elke coalitie z'_k berekenen met de SHAP-kernel. De SHAP kernel π_x voor z' wordt gegeven door:

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|} \mid z' \mid (M- \mid z' \mid)}$$

   met M het aantal inputfactoren van het model en $\mid z' \mid$ het aantal aanwezige inputfactoren volgens coalitievector z'_k. Deze formule toegepast op het voorbeeld $z'_k = (1, 0, 1, 1)$ geeft: $\pi_x(z'_k) = \frac{(4-1)}{4 \cdot 3 \cdot (1)} = \frac{3}{12}$

4. Als laatste stap wordt een lineair regressiemodel g gefit:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z' j$$

   Dit model wordt getraind met de volgende lossfunctie:

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z')$$

   De coëfficiënt ϕ_j van het regressiemodel komt dan overeen met de Shapley waarde voor inputfactor j.

SHAP-explanations maken op verschillende manieren gebruik van de berekende Shapley waarden. In deze masterproef wordt er gebruik gemaakt van de *feature importance* explanation. Deze explanation geeft aan welke inputfactoren het belangrijkste waren voor een bepaalde voorspelling van het model. Ze geven ook aan in welke richting ze hebben bijgedragen aan de voorspelling.

Problemen met SHAP

Hoewel SHAP een van de meest gebruikte explanationmethodes is, zijn er ook enkele problemen aan verbonden. Dat SHAP niet altijd robuust is, en dus voor dezelfde invoer een verschillende explanation kan geven werd reeds aangetoond door Alvarez-Melis et al. [3]. Ook het achterliggend gebruik van Shapley-waarden wordt bekritiseerd door onderzoek van Sundararajan et al. [66]. Hierin wordt gesteld dat het gebruik van Shapley waarden zoals gegeven in het originele onderzoek van Shapley [1], niet altijd correct toegepast worden in implementaties waarbij een deel van de kracht ervan verloren gaat. SHAP gaat ook uit van onafhankelijkheid van de inputfactoren en maakt aannames over de datadistributie waaraan niet altijd voldaan is. Ook het feit dat SHAP-explanations additief zijn, terwijl de meeste complexe modellen niet-additief zijn zorgt ervoor dat SHAP vaak niet alle nuances en interacties in een model kan vatten [25]. De betrouwbaarheid van SHAP-explanations is dus niet altijd even groot en afhankelijk van de implementatie en data.

Naast deze theoretische bezwaren zijn er ook nog problemen met de interpretatie van SHAP-explanations door gebruikers. Ongebruikte features in het model kunnen toch een score toegewezen krijgen door SHAP, wat ingaat tegen de intuïtie van gebruikers en voor verwarring zorgt [66]. Slack et al. [65] hebben aangetoond dat het makkelijk is om SHAP te misleiden. Ze zijn er in geslaagd om SHAP-explanations te genereren voor een model dat bevooroordeeld was, zonder dat deze explanations blijk gaven van deze vooroordelen. Dit alles leidt ertoe dat het correct interpreteren van SHAP-explanations niet altijd even makkelijk is.

2.2.5 Counterfactual explanations

Neem terug de kredietscore gebruikt door Amerikaanse banken zoals reeds besproken in paragraaf 2.1.2. Indien een bank geen lening wil geven aan een bepaalde persoon, op basis van de beslissing van een AI-systeem, heeft deze persoon minder baat bij een explanation zoals SHAP die kan bieden. Weten welke inputfactoren de grootste invloed hebben maakt weinig verschil. Waar de persoon wel in geïnteresseerd is, is hoe ervoor gezorgd kan worden dat de beslissing van het algoritme wordt omgezet van een negatief advies naar een positief advies. Dit levert bruikbare inzichten op die de persoon kunnen helpen om alsnog een lening te bekomen.

CounterFactual explanations [68, 49] (*CF*) zijn explanations die de persoon kunnen helpen dit soort inzichten te bekomen. Ze geven een antwoord op de volgende vraag: *Welke inputfactoren moeten op welke manier veranderen om een bepaalde voorspelling te bekomen?*. Meer specifiek beschrijft een counterfactual explanation voor een voorspelling de kleinste verandering aan de inputfactoren die ervoor zou zorgen dat de voorspelling anders zou zijn. In het voorbeeld van de kredietscore kan een mogelijke explanation bijvoorbeeld zijn dat indien de persoon €5.000 per jaar meer zou verdienen en 1 kredietkaart zou hebben in plaats van twee, het systeem een positief advies zou geven. Het denken in contrafeiten (*counterfactuals*) is iets

dat mensen van nature doen, en uitleg gegeven op deze manier is vaak makkelijk te interpreteren. Psychologisch onderzoek door Byrne et al. [11, 12] heeft uitgewezen dat het gebruik van contrafeiten er voor zorgt dat mensen causaal gaan redeneren. Dat zorgt er dan weer voor dat ze verbanden proberen maken tussen de invoer en uitvoer van het algoritme, wat de (waargenomen) transparantie van het algoritme verhoogt. Fernández-Loria et al. [19] hebben ondervonden dat mensen een voorkeur hebben voor counterfactual explanations wanneer ze vergeleken worden met methodes die werken met *feature importance* zoals SHAP.

Er bestaan meerdere implementaties van counterfactual explanations, in zowel modelspecifieke als modelagnostische vorm.

Wiskundige achtergrond

In wat volgt wordt er dieper ingegaan op de wiskundige basis van counterfactual explanations, gebaseerd op Molnar [48]. Dit wordt gedaan door de methode van Wachter et al. [70] te bespreken waarop een aantal varianten van counterfactual explanations gebaseerd zijn. Dit biedt een goed beeld van de achterliggende concepten zonder in detail te bekijken welke optimalisaties en verbeteringen zijn toegevoegd door specifieke implementaties.

De methode van Wachter et al. [70] stelt de volgende lossfuntie voor die geminimaliseerd moet worden om een counterfactual explanation te bekomen:

$$L(x, x', y', \lambda) = \underbrace{\lambda \cdot (\hat{f}(x') - y')^2}_{(1)} + \underbrace{d(x, x')}_{(2)} \tag{2.1}$$

Term (1) van lossfunctie 2.1 stelt de kwadratische afstand voor tussen de voorspelling van het model $\hat{f}$ voor counterfactual $x' = \hat{f}(x')$ en de gevraagde voorspelling y' die door de gebruiker gegeven wordt.
Term (2) van lossfunctie 2.1 stelt de afstand d voor tussen de uit te leggen instantie x en de counterfactual x'. d is gedefinieerd als de Manhattan-afstand gewogen met de inverse van de Median Absolute Deviation (MAD):

$$d(x, x') = \sum_{j=1}^{p} \frac{| \, x_j \; - \; x'_j \, |}{MAD_j}$$

met p het aantal inputfactoren en de Median Absolute Deviation (MAD) gedefineerd als volgt:

$$MAD_j = \text{mediaan}_{i \in \{1,\ldots,n\}}(| \, x_{i,j} \; - \; \text{mediaan}_{l \in \{1,\ldots,n\}}(x_{l,j}) \, |)$$

Parameter λ in lossfunctie 2.1 stelt het gewicht voor dat gegeven wordt aan de afstand van de predictie (term (1)) ten opzichte van de afstand van de inputfactoren (term (2)). Lossfunctie 2.1 wordt opgelost voor een gegeven gewicht λ, de uit te leggen

instantie x en de gevraagde voorspelling y' en geeft dan een counterfactual x' terug. Een grote waarde voor het gewicht λ zorgt er voor dat er counterfactuals gegenereerd worden die zeer dicht liggen bij de gevraagde voorspelling y'. Een kleinere waarde voor λ zorgt voor gegenereerde counterfactuals die inputfactoren hebben die zeer dicht liggen bij de inputfactoren van x. Indien λ zeer groot gekozen wordt zal de methode altijd de counterfactual x' genereren die zorgt voor een voorspelling die het dichtst bij y' ligt, onafhankelijk van hoe ver x' ligt van x. De keuze van λ kan dus gezien worden als de tradeoff tussen de plausibiliteit van de counterfactual (inputfactoren die dicht bij de huidige inputfactoren liggen) en de accuraatheid van de counterfactual (inputfactoren die zorgen voor een voorspelling die dicht bij de gevraagde voorspelling ligt). Het kiezen van deze parameter is in de meeste implementaties van counterfactual explanations iets dat niet zelf gedaan moet worden, maar geoptimaliseerd wordt door het systeem.

Lossfunctie 2.1 moet dan vervolgens geminimaliseerd worden. Hiervoor kan gebruik worden gemaakt van eender welk optimalisatiealgoritme zoals Gradient Descent of Newton's optimalisatiemethode.

Samengevat werkt de methode van Wachter et al. [70] volgens de volgende stappen:

1. Kies de uit te leggen instantie x, de gewenste voorspelling y' en gewicht λ.
2. Minimaliseer lossfunctie 2.1 en geef de gevonden x' terug als counterfactual explanation.
3. (Optioneel) Indien er meerder counterfactuals gewenst zijn kan stap (2) herhaald worden met steeds grotere waarden voor het gewicht λ. Hiervoor kan een stopconditie gegeven door $\mid \hat{f}(x') - y' \mid > \epsilon$ met ϵ de tolerantie van de afstand tussen de voorspelling voor counterfactual x' en de gewenste voorspelling y' gebruikt worden.

Problemen met counterfactual explanations

Een van de grootste problemen met counterfactual explanations is het zogenaamde *Rashomon Effect* [43]. In de meeste gevallen zijn er meerdere counterfactual explanations mogelijk voor één gevraagde voorspelling. Elk van deze counterfactual explanations schetst een verschillend beeld van de mogelijke veranderingen die moeten gebeuren om tot de gevraagde uitkomst te komen. Dit kan tot contradicties leiden waarin de ene counterfactual aangeeft dat inputfactor A moet veranderen en de andere inputfactoren gelijk moet blijven, terwijl een andere counterfactual aangeeft dat de andere inputfactoren gelijk moet blijven en inputfactor B moet veranderen. Het is dus belangrijk om hierin expliciete keuzes te maken tussen het tonen van alle gegenereerde counterfactual explanation voor een bepaalde gevraagde uitkomst, of het selecteren van de 'beste' counterfactual explanation op basis van een bepaald criterium. Een ander probleem bij counterfactual explanations is dat het niet altijd gegarandeerd kan zijn dat een counterfactual plausibel is [33]. Doordat er

gezocht wordt naar een combinatie van inputfactoren die zorgt voor een bepaalde uitkomst kan het zijn dat de gevonden combinatie theoretisch correct is, en dus voor de gevraagde voorspelling zorgt, maar dat de waarden van de inputfactoren niet van toepassing zijn op de situatie. Neem terug het voorbeeld van de kredietscore. Het kan zijn dat er een counterfactual explanation gegenereerd wordt die zegt dat de persoon '-1' kredietkaart moet hebben zodat hij een positief advies krijgt. In realiteit is het natuurlijk onmogelijk om een negatief aantal kredietkaarten te hebben, maar het model geeft voor deze inputfactoren toch de gevraagde voorspelling. Ook de stabiliteit van counterfactual explanations is niet altijd gegarandeerd [38]. Met stabiliteit wordt bedoeld dat twee gelijkaardige inputs tot gelijkaardige explanations moet leiden. In het geval van de kredietscore zou het kunnen dat persoon A €20.000 per jaar verdient en 1 kredietkaart heeft en persoon B €22.000 euro per jaar verdient en 1 kredietkaart heeft. Ze krijgen beiden een negatief advies van het AI-systeem. Indien er nu counterfactual explanations gegenereerd worden voor beide gevallen kan het voorkomen dat de counterfactual voor persoon A stelt dat hij €25.000 euro per jaar moet verdienen en 1 kredietkaart moet hebben, terwijl de counterfactual voor persoon B stelt dat zij €22.000 euro per jaar moet verdienen maar geen kredietkaarten mag hebben. Beide oorspronkelijke inputs liggen dicht bij elkaar, maar de gegenereerde counterfactual explanations zijn zeer verschillend. Dit kan leiden tot wantrouwen van gebruikers.

2.3 Energiedomein

Het toepassingsgebied van deze masterproef is het energiedomein. Binnen dit domein werd er reeds veel onderzoek verricht vanuit verschillende standpunten. Een vaak onderzocht probleem is het voorspellen van de energieprijs. Om accuraat de energieprijs te voorspellen moet er rekening gehouden worden met vele (externe) factoren. De selectie van de inputfactoren voor het model bepaalt voor een groot deel hoe accuraat het model zal zijn. Hoewel er verschillende technieken bestaan binnen de state-of-the-art om inputfactoren te selecteren is er in de literatuur slechts een beperkte hoeveelheid onderzoek te vinden dat zich focust op de selectie van inputfactoren voor energieprijsvoorspellingen. Dat er net zo veel mogelijke inputfactoren zijn leidt tot zeer complexe en foutgevoelige voorspellingsalgoritmen. Naast de keuze van de inputfactoren heeft ook de keuze van het model een grote invloed op de kwaliteit van de voorspellingen. Amjady et al. [4] stellen dat voorspellingstechnieken voor time series goede resultaten leveren om het stabiele seizoensgedrag van de energieprijs te voorspellen, maar moeilijkheden hebben om het niet-lineaire gedrag van de abrupte veranderingen in energieprijs op te vangen. Traditionele neurale netwerken daarentegen zijn beter in het voorspellen van het niet-lineaire gedrag maar vertonen problemen door de veelheid aan complexe niet-lineaire verbanden. Om deze problemen op te lossen wordt er in de state-of-the-art veelal gebruik gemaakt complexe neurale netwerken. Chaudhury et al. [15] vergelijken verschillende AI-algoritmen die proberen de energieprijs op de open markt te voorspellen zoals support vector machines, kNearest neighbor en long short-term memory *(LSTM)* neurale netwerken.

Ze concluderen dat het gebruik van een LSTM de beste voorspelling oplevert. Lu et al. [41] geven een overzicht van de evolutie van energieprijsvoorspellingsalgoritmen en geven een classificatie op basis van het gebruikte voorspellingsmodel. Hieruit blijkt dat het onderzoeksgebied evolueert richting hybride modellen die verschillende technieken combineren (bijvoorbeeld een timeseriesmodel om de seizoensgebonden prijs te voorspellen, gecombineerd met een complex neuraal netwerk om de pieken ten opzichte van die seizoensgebonden prijs te voorspellen). Ze stellen ook dat de huidige state-of-the-art modellen voor energieprijsvoorspelling een gemiddelde absolute procentuele fout (*MAPE*) hebben tussen 0.0003 en 0.2. Dit toont aan dat er binnen de state-of-the-art modellen nog veel variatie zit in accuraatheid van voorspellingen én dat dit onderzoeksgebied nog veel toekomstig potentieel en mogelijkheid tot verbetering heeft.

2.4 Gerelateerd werk

Het gebruik van SHAP voor domeinexperts en de invloed op de mens-AI taakperformantie werd reeds geëvalueerd door Weerts et al. [72]. Hier werd geen significant verschil gevonden in de taakperformantie tussen gebruikers die SHAP-explanations te zien kregen en gebruikers die geen SHAP-explanations te zien kregen. Gebruikers gaven wel aan dat het gebruik van SHAP-explanations een invloed had op het beslissingsproces. Weerts et al. [72] geven de suggestie dat SHAP-explanations mogelijk gecombineerd moeten worden met andere explanationmethodes om betere resultaten te bekomen. Gosiewska et al. [25] benoemen enkele problemen van SHAP-explanations, waaronder de veronderstelling van onafhankelijkheid van de inputfactoren. Ze concluderen dat SHAP-explanations inconsistent zijn en uiteindelijk in vele gevallen niet betrouwbaar zijn. Kaur et al. [32] komen tot de conclusie dat het gebruik van SHAP-explanations voor AI-experts leidt tot 'blind vertrouwen' in de explanations en daarbij ook in het systeem. Ze bemerken ook dat SHAP-explanations vaak misbruikt worden door AI-experts. Alvarez-Melis et al. [3], Sundarajan et al. [66] en Slack et al. [65] gaan dieper in op andere problemen van SHAP zoals al vermeld in paragraaf 2.2.4. Een overzicht van de verschillende implementaties van counterfactual explanations en hun problemen en uitdagingen wordt gegeven in Keane et al. [33]. Warren et al. [71] laten gebruikers een model nabootsen door Model Output Prediction (gebruikers de voorspelling van het model laten voorspellen), na hen eerst kennis te laten maken met het model door gebruik van counterfactual explanations. Ze stellen dat counterfactual explanations zorgen voor een licht verhoogde accuraatheid van de voorspellingen van gebruikers. Ze onderzoeken ook of het gebruik van continue en categorische inputfactoren een invloed heeft op de accuraatheid van de voorspellingen van de gebruikers en komen tot de conclusie dat het gebruik van categorische inputfactoren de voorkeur heeft.

2.5 Besluit

Het XAI-domein is een zeer breed en maatschappelijk relevant onderzoeksdomein. Het huidige onderzoek in verband met XAI richt zich echter veelal op Data-Experts en AI-experts. Er is in de realiteit niet enkel nood aan explanations voor deze twee doelgroepen, maar ook voor effectieve eindgebruikers zonder eerdere ervaring. In onze huidige maatschappij wordt er steeds meer gebruik gemaakt van AI-systemen. Deze systemen kunnen mensen helpen beslissingen te maken. Indien er de hoop is AI-systemen universeel te kunnen gebruiken en voor iedereen beschikbaar te stellen moet er voor gezorgd worden dat AI-novices niet uit het oog verloren worden.

Hoofdstuk 3

Methode

In deze masterproef ligt de focus op het vergelijken van het vertrouwen in en het begrip van AI-novices in black-box AI-algoritmen, gerealiseerd door gebruik te maken van lokale, modelagnostische post-hoc explanationmethodes. Meer specifiek worden er twee state-of-the-art explanationmethodes bestudeerd: Shapley additive explanations [42] (*SHAP*) en counterfactual explanations [68] (*CF*). Deze explanationmethodes zullen worden toegepast op een neuraal netwerk dat prijsvoorspellingen doet van de energieprijs in België. Het evalueren van de explanations zal gebeuren aan de hand van een gebruikersstudie, goedgekeurd door de Sociaal-Maatschappelijke Ethische Commissie van de KU Leuven (dossiernummer G-2022-497). De gebruikersstudie zal bestaan uit drie verschillende groepen. Gebruikers worden willekeurig toegewezen aan één specifieke groep. Tabel 3.1 geeft weer welke informatie en explanations de gebruikers uit iedere groep te zien krijgt.

Groep	Informatie/Explanations
A	Inputfactoren
B	Inputfactoren & SHAP
C	Inputfactoren, SHAP & CF

Tabel 3.1: Studiegroepen en de beschikbare informatie in de gebruikersstudie.

Er werd gekozen voor het energiedomein als toepassingsgebied vanwege de maatschappelijke relevantie van dit domein, alsook vanwege de persoonlijke interesse van de auteur. De relevantie van XAI-onderzoek in de context van novice gebruikers werd reeds toegelicht in paragraaf 2.5. Aangezien SHAP een van de meest gebruikte explanationmethodes is voor data-experts en AI-experts zal deze masterproef evalueren of SHAP-explanations ook nuttig kunnen zijn voor AI-novices. Meer bepaald zal er geëvalueerd worden of SHAP-explanations zorgen voor meer vertrouwen in en algoritmische transparantie van black-box AI-algoritmen bij AI-novices. Ook zal er worden bekeken of counterfactuals een toevoeging kunnen vormen voor SHAP om diezelfde doelen te bereiken. Hiervoor worden de volgende onderzoeksvragen *(RQ)* voorgesteld:

RQ1: Hoe beïnvloedt het gebruik van Shapley additive explanations het begrip van en het vertrouwen in een black-boxmodel bij AI-novices?

RQ2: Leiden counterfactual explanations als toevoeging voor Shapley additive explanations tot een groter vertrouwen in en een beter begrip van een black-boxmodel voor AI-novices?

In paragraaf 3.1 wordt er dieper ingegaan op de technische implementatie van het voorspellingsmodel en de explanationmethodes. De evaluatiemethode van de gebruikersstudie zal worden toegelicht in paragraaf 3.2.

3.1 Implementatie van het voorspellingsmodel en de explanations

3.1.1 Voorspellingsmodel

Om de gewenste explanationmethodes te kunnen evalueren moeten ze toegepast worden op een black-boxmodel. In deze masterproef is er gekozen om de energieprijs in België te voorspellen met behulp van een neuraal netwerk.

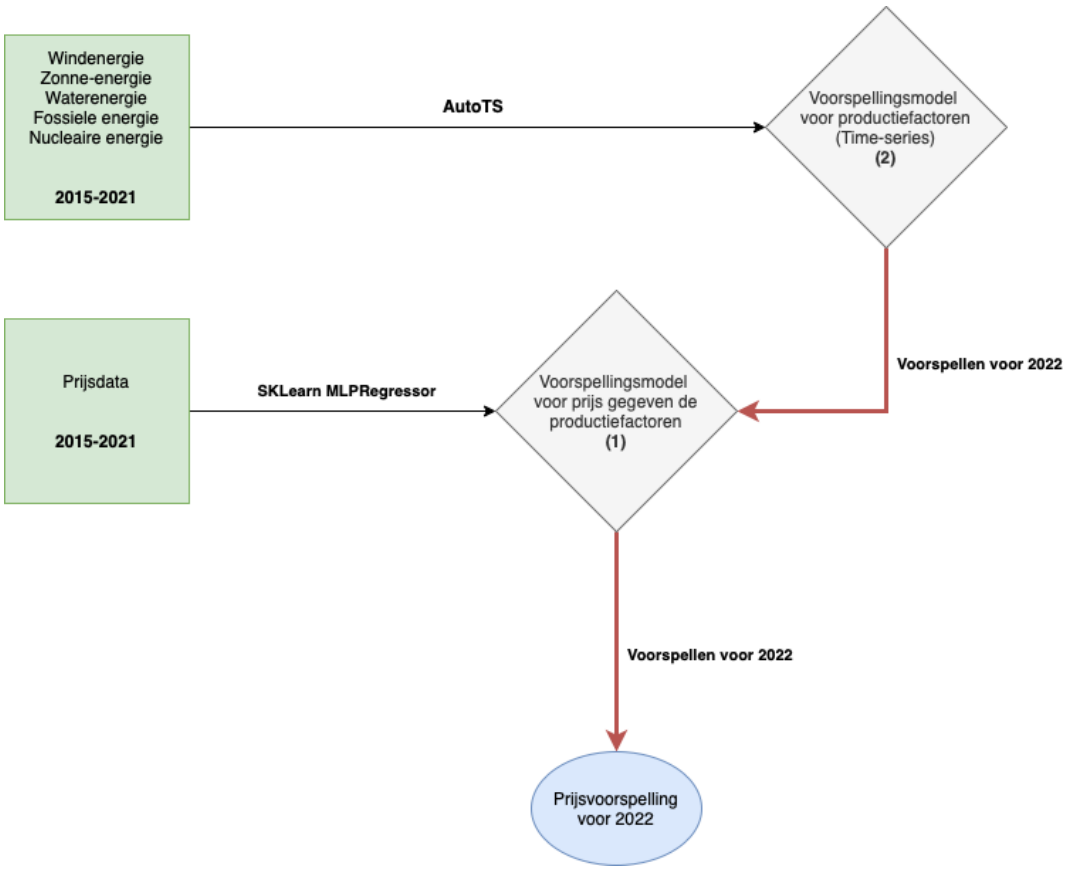
Trainingsdata

Het model werd getraind met data van het 'ENTSO-E Transparency Platform'. Dit is een open-dataplatform van de Europese Unie, specifiek gericht op data van de energiemarkt [1]. Er werd gekozen om de energieprijs in België te voorspellen op basis van de hoeveelheid energieproductie van verschillende types: windenergie, zonne-energie, waterenergie, energie uit fossiele brandstoffen en nucleaire energie. Om voorspellingen te doen over de energieprijs in de toekomst, waarbij de productiefactoren als invoer voor het neuraal netwerk gebruikt kunnen worden werd er besloten om de voorspelling met behulp van twee modellen te doen zoals samengevat in fig. 3.1:

1. Een neuraal netwerk getraind met als inputfeatures de productiefactoren van 2015 tot en met 2021 en al doelvariabele de dagprijs van energie, beiden afkomstig uit het 'ENTSO-E Transparency Platform'. Met dit model kan de dagprijs van energie in België bepaald worden op basis van de hoeveelheid energieproductie van ieder type. Voor de implementatie van dit neuraal netwerk werd gekozen voor de *MLPRegressor* uit het populaire Python machine learning framework Scikit-Learn [2].

[1] https://transparency.entsoe.eu.
[2] https://scikit-learn.org/.

2. Het voorspellen van de dagwaarde van de productie van elk type voor het jaar 2022 op basis van de data uit het 'ENTSO-E Transparency Platform' van 2015 tot en met 2021 gebruik makend van AutoTS [3], een Python-library voor timeseries voorspellingen. Deze dagwaarden zijn nodig voor het jaar 2022 omdat model (1) voorspelt op basis van deze waarden. Aangezien het gaat om dagwaarden van toekomstige dagen is deze data niet beschikbaar, en moet deze voorspeld worden.



Figuur 3.1: Samenvatting van de gebruikte modellen voor voorspellingsmodel.

Wiskundige achtergrond

Zoals al eerder aangehaald wordt er voor het uiteindelijke prijsvoorspellingsmodel gebruik gemaakt van de *MLPRegressor* uit Scikit-Learn. *MLPRegressor* staat voor multilayer perceptron regressor, en is een subtype van neurale netwerken dat gebruikt wordt voor regressiedoeleinden. In wat volgt wordt er een algemene uitleg gegeven over de werking van een multilayer perceptron voor binaire data, gebaseerd op het werk van Shalev-Shwartz et al. [62] en Nielsen [51].

[3] https://github.com/winedarksea/AutoTS.

Een perceptron is het kleinste element in een neuraal netwerk, gebaseerd op de neuronen in de hersenen van de mens. Het doel van een perceptron is het omzetten van meerdere binaire inputs naar een output. Zoals te zien in fig. 3.2 worden de verschillende inputs $x_1, \ldots, x_n$ door de perceptron vermenigvuldigd met de bijbehorende gewichten $w_1, \ldots, w_n$. Vervolgens worden ze gesommeerd. Deze som $\sum_{i=1}^{n} w_i x_i$ wordt vervolgens door een activatiefunctie g gehaald. Deze activatiefunctie bepaalt voor binaire invoer of de perceptron 'actief' is $(g(\sum_{i=1}^{n} w_i x_i) = 1)$ of 'inactief' is $(g(\sum_{i=1}^{n} w_i x_i) = 0)$. In het meest simpele geval is deze activatiefunctie een stapfunctie die vanaf een bepaalde drempelwaarde σ activeert. Meer exact:

$$g(\sum_{i=1}^{n} w_i x_i) = \begin{cases} 0 & \text{als } \sum_{i=1}^{n} w_i x_i \leq \sigma \\ 1 & \text{als } \sum_{i=1}^{n} w_i x_i > \sigma \end{cases}$$

Een multilayer perceptron is niet meer dan het woord zelf zegt: meerdere lagen aan perceptrons zoals te zien in fig. 3.3. Iedere perceptron in een multilayer perceptron doet een sommatie van zijn gewogen inputs en haalt deze vervolgens door een activatiefunctie. Het trainen van een multilayer perceptron bestaat uit het vinden van de meest optimale gewichten $w_{i,j}$ voor iedere input i en laag j zodat de perceptrons voor de correcte combinaties van invoer $(x_1, \ldots, x_n)$ activeren en zorgen voor de gewenste uitvoer. Dit wordt gedaan aan de hand van backpropagation. Backpropagation is het mechanisme dat iteratief de gewichten in het netwerk aanpast tot de gevonden gewichten zorgen voor de optimale output. In backpropagation wordt dit gedaan aan de hand van een lossfunctie. In het geval van de MLPRegressor van Scikit-Learn gaat het om de kwadratische fout tussen de gewenste uitvoer y en de effectieve uitvoer $\hat{y}$:

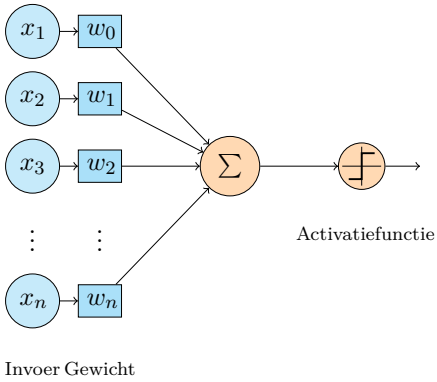
$$L(y, \hat{y}) = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

In meer complexe neurale netwerken kan er ook vertrokken worden van continue numerieke data, en is er een veelheid aan activatiefuncties beschikbaar, elk met hun eigen voor- en nadelen. Het volledig beschrijven van de wiskundige basis van neurale netwerken in al zijn vormen en de exacte werking van het backpropagation-algoritme valt buiten het bestek van deze masterproef.

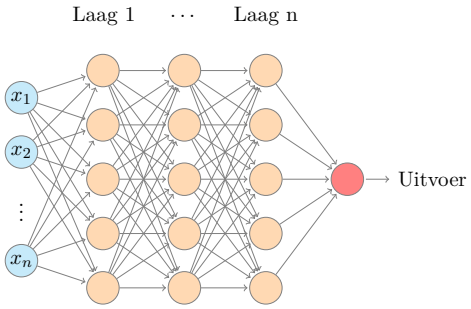
3.1.2 Shapley additive explanations

In de implementatie voor de gebruikersstudie van deze masterproef wordt er gebruik gemaakt van de KernelSHAP implementatie van Lundberg et al. [42]. Deze implementatie is te downloaden en te gebruiken als Python package [4]. Het enige dat deze implementatie van SHAP nodig heeft om explanations te kunnen genereren is het prijsvoorspellingsmodel uit paragraaf 3.1.1 en de inputfactoren van de voorspelling waarvan een uitleg gegenereerd moet worden. Voor de gebruikersstudie werden

[4] https://github.com/slundberg/shap.



Figuur 3.2: Grafische weergave van een perceptron



Figuur 3.3: Grafische weergave van een multilayer perceptron

SHAP-explanations gegenereerd voor iedere voorspelling van het model. Deze werden gebruikt in de webapplicatie waarmee de gebruikersstudie uitgevoerd werd. Meer details over deze webapplicatie worden gegeven in hoofdstuk 4.

3.1.3 Counterfactual explanations

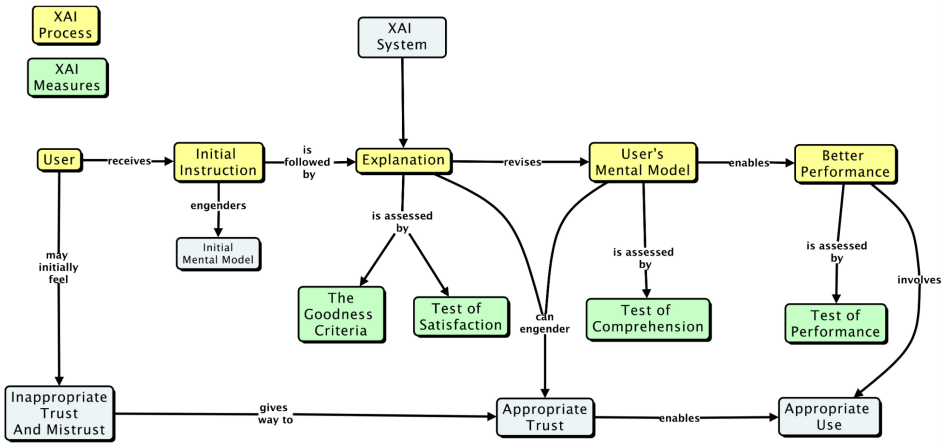
Voor de implementatie van de counterfactual explanations werd gekozen voor DiCE [5]. DiCE (Diverse Counterfactual Explanations) is een methode voor het genereren van counterfactual explanations ontwikkeld door onderzoekers van Microsoft [49]. Ook deze methode is net zoals SHAP beschikbaar als Python package. Hoewel DiCE gezien wordt als modelagnostische explanationmethode bevat het DiCE Python package ook modelspecifieke implementaties die sterk geoptimaliseerd zijn voor efficiëntie. Om explanations te genereren voor de gebruikersstudie werd er gekozen voor de modelagnostische implementatie aangezien de modelspecifieke implementaties niet geschikt waren voor het gekozen model. Om de modelagnostische implementatie te kunnen gebruiken moet er gekozen worden tussen drie varianten: random, genetic of kd-tree. Deze varianten hebben invloed op de manier waarop er nabije instanties gekozen worden om te evalueren als potentiële counterfactual explanation. Na experimentatie met de drie verschillende opties werd er gekozen voor de genetic variant. Deze variant zorgde voor de beste counterfactual explanations met de kortste berekeningstijd. Voor de gebruikersstudie werden er counterfactual explanations gegenereerd voor iedere voorspelling van het model. Hoofdstuk 4 gaat dieper in op hoe deze explanations verwerkt werden in de gebruikersstudie.

3.2 Evaluatiemetrieken

Om een antwoord te kunnen bieden op de onderzoeksvragen moeten de verschillende explanationmethodes geëvalueerd worden aan de hand van de verschillende evalua-

[5] https://github.com/interpretml/DiCE.

tiecriteria zoals voorgesteld door Mohseni et al. [47] en Hoffman et al. [28]. In deze masterproef worden er twee evaluatiecriteria gebruikt: mentaal model en vertrouwen. Bovenop de verzamelde data rond deze twee evaluatiecriteria wordt er ook nog extra data verzameld om de onderzoeksvragen mee te kunnen beantwoorden. Hoffman et al. [28] stellen een conceptueel model voor van het explanationproces in fig. 3.4. Hetgene gemeten zal worden in de gebruikersstudie in deze masterproef komt overeen met 'User's Mental Model', 'Inappropriate Trust And Mistrust' en 'Appropriate Trust'. Hiermee behandelen we een groot deel van het explanationproces. Toch zijn er onderdelen van het proces die in de gebruikersstudie niet aan bod zullen komen zoals de voldoening die een explanation geeft aan de gebruiker en het initiële mentaal model.



Figuur 3.4: Conceptueel model van het explanationproces, uit Hoffman et al. [28].

3.2.1 Mentaal model

Wanneer het mentaal model van de gebruiker geëvalueerd wordt, is het doel om te weten te komen in hoeverre de gebruiker begrijpt hoe het systeem werkt en hoe het tot een voorspelling komt. Het kan ook gezien worden als proxy voor de algoritmische transparantie. Wanneer de gebruiker een accuraat mentaal model heeft van het systeem, wil dat zeggen dat het systeem een zekere mate van transparantie heeft en dus de gebruiker toelaat in te zien hoe het tot een beslissing of voorspelling komt. Mohseni et al.[47] stellen verschillende evaluatiemethodes voor om het mentaal model van een gebruiker te meten. In deze masterproef is er gekozen om gebruik te maken van Model Output Prediction. Hierbij krijgt de gebruiker een mogelijke invoer voor het systeem te zien, en is het de bedoeling dat de gebruiker aangeeft welke voorspelling het systeem voor deze gegeven invoer zal doen, samen met de redenering die de gebruiker heeft gevolgd om tot die conclusie te komen. De gebruikte vraagstelling hiervoor is te vinden in fig. A.3. De gegeven voorspellingen van de gebruikers worden omgezet naar numerieke waarden door de kwadratische

gemiddelde afwijking *(RMSE)* ten opzichte van de voorspelling van het systeem te berekenen. De kwadratische afwijking kan een waarde hebben gaande van 0, voor een gebruiker die alle voorspellingen correct heeft, tot 53, voor een gebruiker die voor alle voorspellingen zo ver mogelijk van de voorspelling van het systeem zit. Hoe kleiner de gemiddelde kwadratische afwijking, hoe beter de voorspelling van de gebruiker, en dus ook het mentaal model van de gebruiker. Indien een gebruiker geen enkele voorspelling geeft door optie *'I don't know'* aan te duiden bij alle vragen wordt de score van de gebruiker verder niet gebruikt.

Ander mogelijkheden om het mentaal model mee te evalueren zijn: interviews met de gebruiker en Likertschaal vragenlijsten. Er werd niet gekozen voor een interview als evaluatiemethode aangezien dit het aantal deelnemers van de studie zou beperken. Een Likertschaal vragenlijst over mentaal model werd niet gebruikt omdat dit in combinatie met de andere vragenlijsten van de gebruikersstudie zou resulteren in een te tijdsintensieve studie voor de deelnemers.

3.2.2 Vertrouwen

Het vertrouwen van gebruikers in een model is een complex gegeven. Het is een moeilijk meetbaar concept en hangt af van vele factoren. Er zijn onderzoekers die van mening zijn dat vertrouwen steeds voortkomt uit andere doelstellingen van XAI en geen doel op zich moet zijn. Yin et al. [73] hebben bijvoorbeeld ondervonden dat de waargenomen accuraatheid van een systeem een grote invloed heeft op het vertrouwen van een gebruiker. Binnen het concept van vertrouwen zijn er nog verschillende soorten te onderscheiden. Er is wantrouwen, ongepast wantrouwen, ongepast vertrouwen, en gepast vertrouwen.

Met wantrouwen wordt de initiële wantrouwigheid van gebruikers bedoeld. Sommige mensen zijn erg op hun hoede voor AI. Er zijn zelfs mensen die angstig worden van het idee van het gebruik van AI. Hiermee moet rekening gehouden worden bij het evalueren van explanationmethodes. Gebruikers die van nature wantrouwig staan ten opzichte van het gebruik van AI zullen een ander niveau van vertrouwen hebben in een systeem dan gebruikers die hier meer positief tegenover staan. Wanneer dit wantrouwen buitenproportioneel groot is wordt dit gezien als ongepast wantrouwen. Om dit initiële wantrouwen te meten wordt er in deze masterproef gebruik gemaakt van vragen gebaseerd op de *unified theory of acceptance and use of technology* [67]. De vragenlijst die hiervoor gebruikt werd is gebaseerd op Andrews et al. [5] en is te vinden in fig. A.4. De antwoorden op deze vragen worden omgezet naar numerieke waarden door ze scores te gegeven volgens de Likertschaal gaande van een tot vijf. Met behulp van deze scores wordt dan het initiële vertrouwen van de gebruiker berekend door het gemiddelde van de score te nemen over alle vragen. Een initieel vertrouwen groter dan drie betekent dat de gebruiker van nature reeds vertrouwen heeft in AI, terwijl een initieel vertrouwen kleiner dan drie betekent dat de gebruiker van nature wantrouwig is ten opzichte van AI. Een volledig neutrale gebruiker heeft een initieel vertrouwen gelijk aan drie.

Ongepast vertrouwen is het vertouwen dat gebruikers onterecht hebben in een systeem. Dit kan bijvoorbeeld voorkomen wanneer de explanations die gegeven worden zeer inaccuraat zijn. Gebruikers kunnen dan vertrouwen krijgen in het systeem op basis van de gegeven explanations, terwijl de explanations niet het juiste beeld geven van het systeem. Hierdoor kan het zijn dat het vertrouwen ongepast is wanneer het systeem in werkelijkheid anders werkt dan de explanations doen uitschijnen. Wanneer gebruikers het systeem vertrouwen, en dat vertrouwen terecht is omdat het systeem goed werkt wordt er gesproken van gepast vertrouwen. Dit is het ultieme doel van vertrouwen gecreëerd door explanations. Het is niet de bedoeling dat de explanations zorgen voor een blindelings vertrouwen in het systeem, wel dat explanations op de juiste momenten zorgen voor vertrouwen in de delen van het systeem die effectief betrouwbaar zijn. Om het vertrouwen in de gebruikte explanations in de gebruikersstudie te meten wordt er gebruik gemaakt van vragen uit Hoffman et al. [28]. Deze vragenlijst is te vinden in fig. A.5. De antwoorden op deze vragen worden ook omgezet naar numerieke data door gebruik te maken van de Likertschaal gaande van een tot vijf. Hiervan wordt voor iedere gebruiker het gemiddelde over alle negen vragen berekend.

Het meten van zowel het initiële vertrouwen als het vertrouwen in het systeem na het gebruik laat toe het vertrouwen opgebouwd door gebruik te maken van het systeem te isoleren zonder het vertekend beeld dat mogelijk optreedt indien de gebruiker van nature reeds zeer wantrouwig of extreem positief is ten opzichte van Artificiële Intelligentie.

3.2.3 Overige data

Naast de vragen die het mentaal model en vertrouwen peilen worden er nog meer vragen gesteld aan en data verzameld van de gebruikers. Eerst wordt er algemene informatie verzameld van de gebruiker: de leeftijd, het opleidingsniveau en het land waar de gebruiker op dit moment woont. Ook het kennisniveau van de gebruiker wordt gemeten met vragen gebaseerd op Ooge et al. [53], te vinden in fig. A.2. De bedoeling van deze vragen is te kunnen inschatten of de gebruiker tot de doelgroep van AI-novices behoort én in te schatten of de gebruiker reeds kennis heeft over het energiedomein. Deze vragen worden omgezet naar numerieke waarden aan de hand van de codering in tabel 3.2. Indien een gebruiker meerdere antwoorden heeft aangeduid wordt hiervan de som genomen. Vervolgens wordt hier het gemiddelde van berekend voor vraag een tot en met drie (AI-kennisniveau) en vraag drie tot en met zes (energiedomein-kennisniveau). Gebruikers die een gemiddelde score hebben groter dan drie op minimaal één van de domeinen werden gelabeld als expert en worden verder buiten beschouwing gelaten aangezien ze geen deel uitmaken van de doelgroep van deze gebruikersstudie.

Ook wordt er gevraagd aan de gebruiker of hij de gegeven explanations nuttig vond en waarom, en of zij het nuttig zou vinden om een ander soort explanation te krijgen en waarom. Deze vragen zijn te vinden in fig. A.6 tot en met fig. A.10.

Antwoord	Waarde
'I don't know what it is'	0
'I know the word'	1
'I often use it'	2
'I can explain it'	3

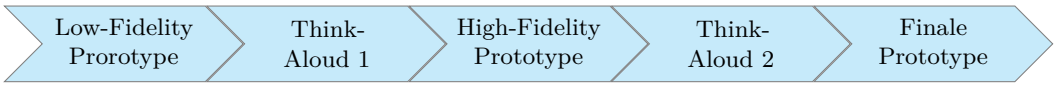
Tabel 3.2: Codering van het kennisniveau, gebaseerd op [53].

Aan het einde van de vragenlijst wordt er ook nog een open vraag gesteld aan de gebruiker waarin hij eventueel nog extra opmerkingen of suggesties kwijt kan. Deze vraag is te vinden in fig. A.11. Tijdens het experiment wordt ook de tijd die de gebruiker spendeert in ieder scherm van de explanationinterface bijgehouden.

Hoofdstuk 4

Ontwikkelingsproces

In dit hoofdstuk wordt er dieper ingegaan op het ontwikkelingsproces dat gevolgd werd bij het ontwerpen van de prototypes voor de gebruikersstudie. Er werd gekozen voor een iteratief proces waarbij de eindgebruiker centraal staat en betrokken wordt bij de tussentijdse evaluaties van het ontwerp. Figuur 4.1 geeft weer welke stappen er ondernomen werden.



Figuur 4.1: Het ontwikkelingsproces van het finale prototype bestaande uit twee eerdere prototypes en twee think-alouds.

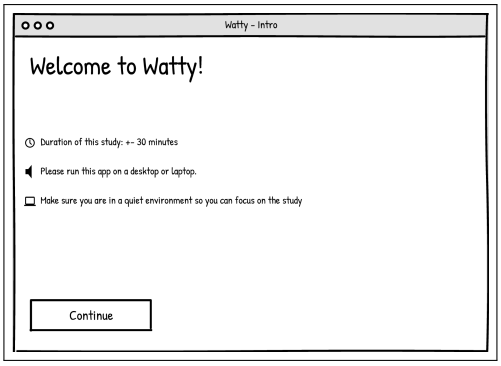
4.1 Low-fidelity prototype

De eerste versie van het prototype werd ontwikkeld in Figma [1]. Figma is een tool specifiek ontworpen om op een snelle manier digitale prototypes te ontwikkelen. Het laat toe makkelijk interactieve versies te maken van een prototype zodat ze afgetoetst kunnen worden bij de gebruiker. Het prototype bestaat uit drie verschillende delen: uitleg van de gebruikersstudie, de explanationinterface van de gebruikersstudie en de vragenlijsten over de gebruikersstudie.

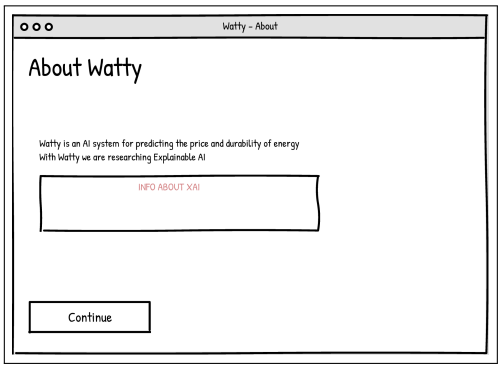
De gebruikersstudie start met de landingspagina zoals te zien in fig. 4.2. Deze pagina geeft enkele richtlijnen aan de gebruiker zoals op welke apparaten de applicatie kan gebruikt worden en hoe lang de gebruikersstudie ongeveer zal duren. Vervolgens krijgt de gebruiker de pagina in fig. 4.3 te zien waarbij er een algemene uitleg wordt gegeven rond het doel van de gebruikersstudie. Hierna dient de gebruiker een geïnformeerde toestemming te geven zoals bepaald door de richtlijnen van de Sociaal-Maatschappelijk Ethische Commissie van de KU Leuven. Deze pagina is te

[1] https://figma.com

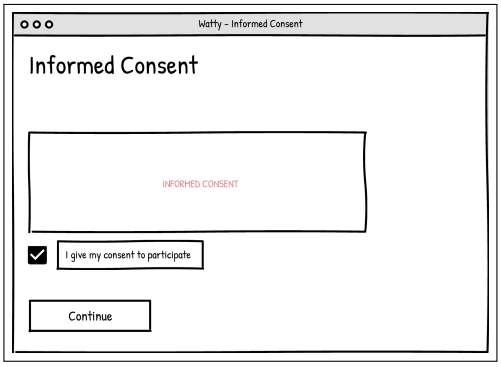
zien in fig. 4.4. Als laatste pagina voor de effectieve gebruikersstudie start krijgt de gebruiker een overzicht van het verdere verloop van de gebruikersstudie te zien zoals in fig. 4.5.



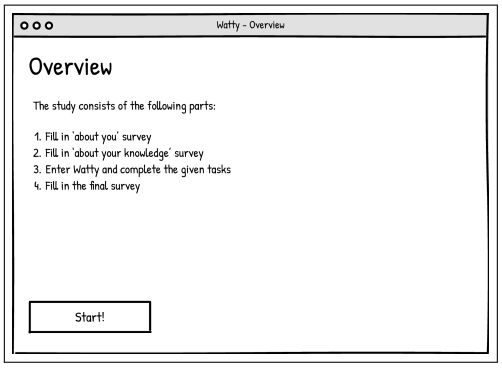
Figuur 4.2: Low-fidelity prototype van de landingspagina.



Figuur 4.3: Low-fidelity prototype van de informatiepagina.



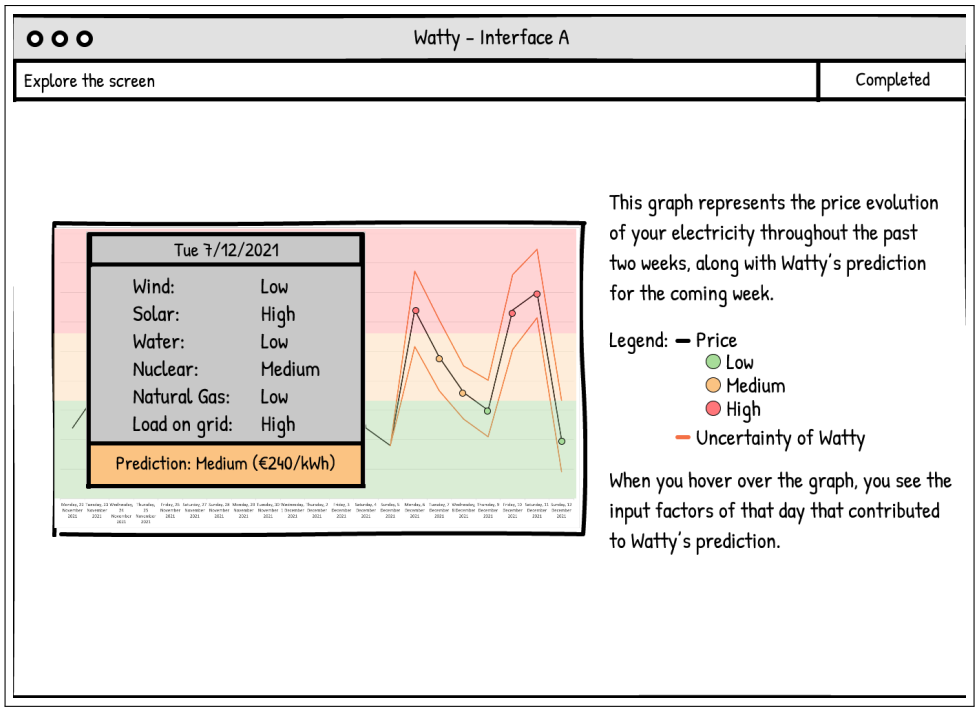
Figuur 4.4: Low-fidelity prototype van de geïnformeerde toestemmingspagina.



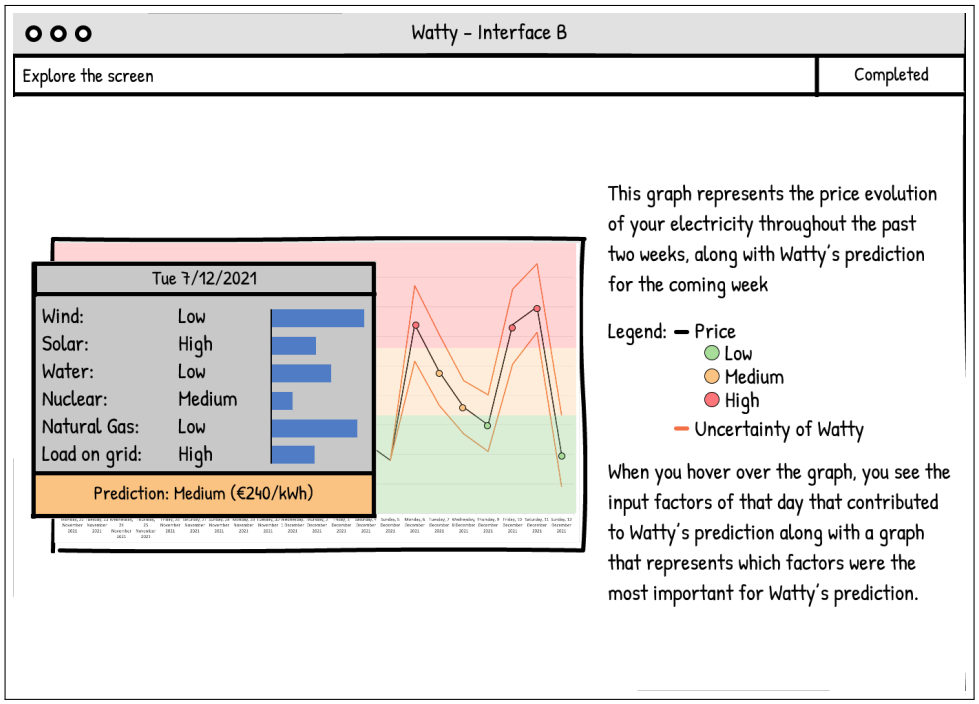
Figuur 4.5: Low-fidelity prototype van de overzichtspagina.

Het belangrijkste deel van het prototype is de explanationinterface van de gebruikersstudie. Hierin krijgen de gebruikers een grafiek te zien met daarin de energieprijs uit het verleden, samen met de voorspelling van het systeem voor de toekomstige energieprijs. De gebruiker krijgt op het scherm ook uitleg over de informatie die hij te zien krijgt. Wanneer de gebruiker op een datapunt klikt opent er een popupvenster met daarin extra informatie, afhankelijk van de groep waar hij/zij willekeurig aan wordt toegewezen door de applicatie bij het starten van de gebruikersstudie.

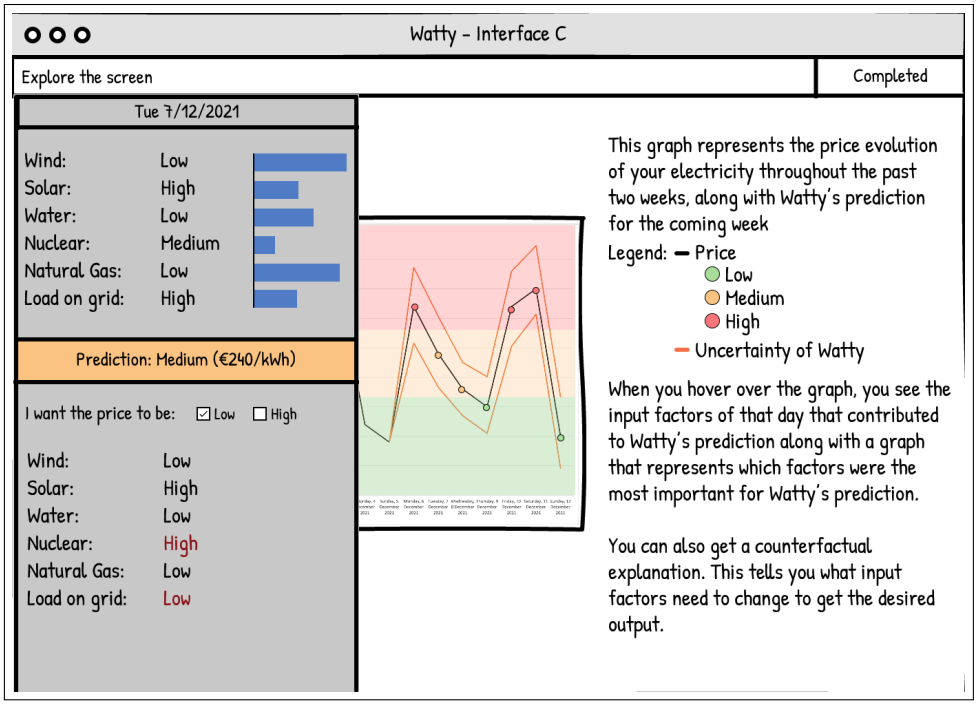
Figuur 4.6 toont het popupscherm dat gebruikers in groep A te zien krijgen. Bij het klikken op een datapunt in de prijsgrafiek krijgt de gebruiker enkel de waarde van de inputfactoren op de geselecteerde dag in tabelvorm te zien. In groep A krijgt



Figuur 4.6: Low-fidelity prototype voor groep A.



Figuur 4.7: Low-fidelity prototype voor groep B.



Figuur 4.8: Low-fidelity prototype voor groep C.

de gebruiker dus geen effectieve explanations te zien voor de voorspelling van de geselecteerde dag. De gebruikers in groep B krijgen zoals te zien in het prototype in fig. 4.7 meer informatie. Bovenop de inputfactoren van de geselecteerde dag krijgen ze ook een SHAP feature importance explanation in de vorm van een horizontaal staafdiagram te zien. De gebruikers in groep C krijgen het meeste informatie te zien wanneer ze een datapunt selecteren. Naast de inputfactoren en SHAP-explanation wordt hier ook nog de mogelijkheid gegeven aan de gebruiker om op een interactieve manier counterfactual explanations op te vragen door de gewenste voorspelling van het systeem aan te duiden, zoals te zien in het prototype in 4.8. De gebruikers krijgen in alle groepen dezelfde taak om het scherm te exploreren tot ze ervan overtuigd zijn dat alle informatie bekeken is.

4.2 Think-aloud studie 1

Na het ontwerp van het low-fidelity prototype werd er een think-aloud studie gedaan met 5 deelnemers. Vier van de deelnemers behoorden tot de groep van AI-novices en één deelnemer is een AI-expert. De AI-expert werd toegevoegd om meer gedetailleerde technische feedback te verzamelen. De leeftijd van de deelnemers varieert tussen 22 en 56 jaar, met een opleidingsniveau gaande van secundair onderwijs tot doctoraatstudent. De think-aloud werd gedaan via een videogesprek en de deelnemers werden gevraagd hun scherm te delen terwijl ze een interactieve versie van het prototype van de volledige gebruikersstudie uitvoerden, van de informatieve pagina's

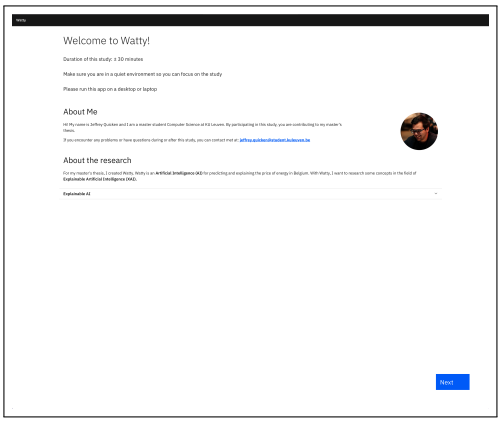
aan het begin tot en met de vragenlijsten aan het einde. De lijst met verzamelde feedback is terug te vinden in appendix B.

De meest frequent voorkomende opmerking was dat de betekenis van de input-factoren zoals gegeven in de popup onduidelijk was. Ook het verschil tussen de inputfactoren en de SHAP-explanation was onduidelijk. Een mogelijke oplossing hiervoor is het aanpassen van de verwoording van de inputfactoren en het geven van een duidelijkere tekstuele uitleg over de betekenis van zowel de inputfactoren als SHAP. Een ander frequent probleem was dat de gebruikers niet voldoende tijd spenderen in de explanationinterface waardoor ze niet voorbereid waren op de vragen die achteraf volgden. Dit komt door de taakomschrijving die nogal breed is ('exploreer het scherm'). Een mogelijke oplossing hiervoor is om de gebruikers een meer specifieke taak te laten vervolledigen waardoor ze de interface langer moeten exploreren.

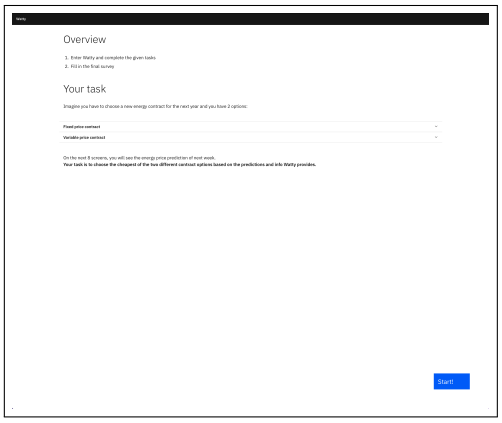
4.3 High-fidelity prototype

Na verschillende iteraties van het low-fidelity prototype werd er overgegaan tot de ontwikkeling van het high-fidelity prototype. Om dit te ontwikkelen werd er gekozen voor SvelteKit [2], een framework om webapplicaties mee te bouwen gebaseerd op het recent populaire JavaScript-framework Svelte [3]. In het high-fidelity prototype werden er verschillende aanpassingen gedaan op basis van de feedback die gegeven werd door de deelnemers aan de eerste think-aloud studie. De landingspagina en informatiepagina werden gecombineerd tot één overzichtspagina zoals te zien in fig. 4.9. De belangrijkste toevoeging in het high-fidelity prototype is dat de taak die aan gebruikers gegeven werd is aangepast, en getoond wordt op een aparte pagina zoals te zien in fig. 4.10. In plaats van enkel het scherm te moeten exploreren wordt er nu een beslissingstaak geformuleerd voor de gebruikers: *'Kies de goedkoopste optie tussen een vast energiecontract en een variabel energiecontract'*. De prijs van het vast contract wordt gegeven, en de prijs van het variabel contract dient afgeleid te worden uit de prijsvoorspelling gegeven door het systeem. De gebruiker maakt dan zelf de keuze voor een contract en geeft een verklaring voor deze beslissing. Door deze taak te geven spenderen de gebruikers meer tijd in de explanationinterface en interageren ze meer met de gegeven explanations. Een ander verschil is dat de explanations en informatie nu niet meer getoond worden in een popup, maar naast de lijngrafiek van de prijsvoorspelling. Het tonen als popup zorgt ervoor dat een deel van de lijngrafiek niet meer zichtbaar is, wat het mogelijk moeilijk maakt voor de gebruiker om alle gegeven data in rekening te nemen bij de beslissing. Figuur 4.11, fig. 4.12 en fig. 4.13 tonen de explanation interfaces die gebruikers in respectievelijk groep A, groep B en groep C te zien krijgen.

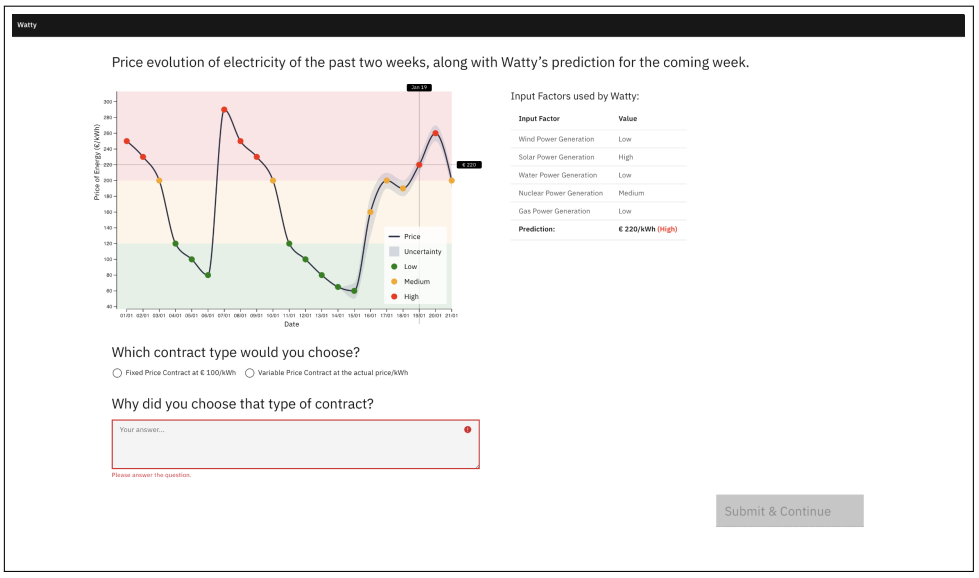
[2] https://kit.svelte.dev
[3] https://svelte.dev



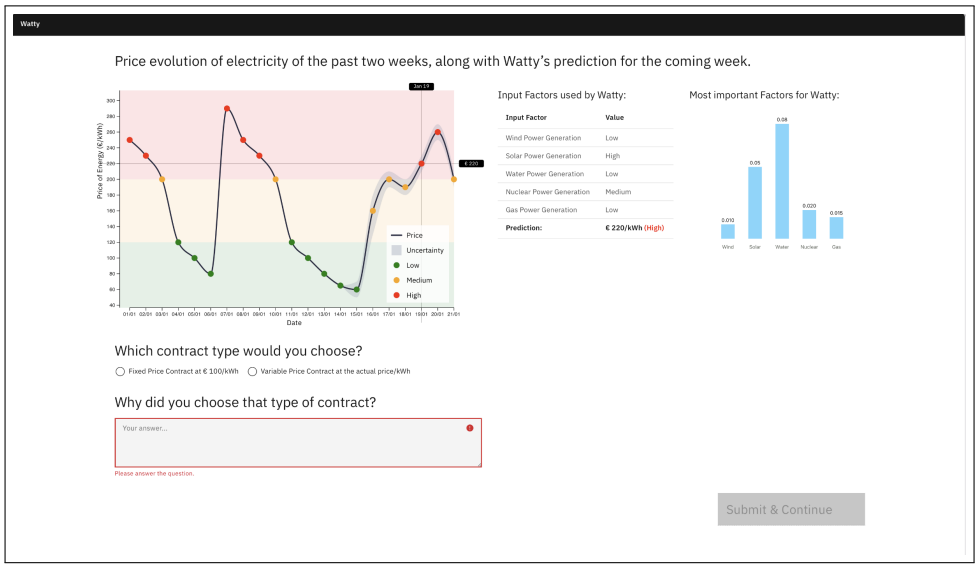
Figuur 4.9: High-fidelity prototype van de landingspagina.



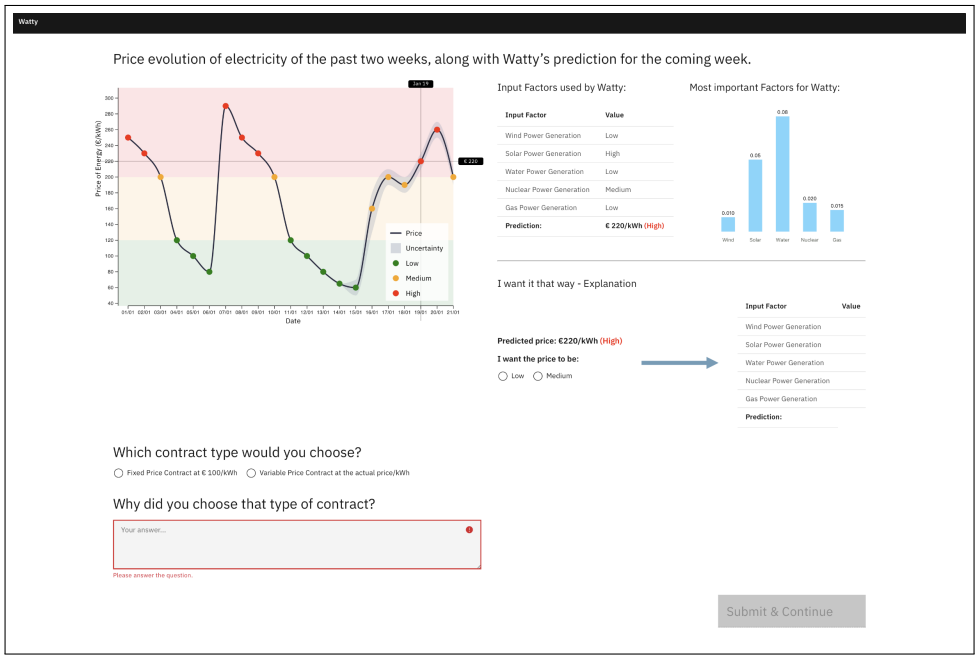
Figuur 4.10: High-fidelity prototype van de taakpagina.



Figuur 4.11: High-fidelity prototype voor groep A.



Figuur 4.12: High-fidelity prototype voor groep B.



Figuur 4.13: High-fidelity prototype voor groep C.

4.4 Think-aloud studie 2

Ook met het high-fidelity prototype werd een think-aloud uitgevoerd met zes deelnemers. Vier van de deelnemers zijn AI-novices en twee deelnemers zijn AI-Experts. De leeftijd varieert tussen 21 en 25 jaar, met een opleidingsniveau gaande van secundair onderwijs tot universitaire master. Ook deze think-aloud werd gedaan via een videogesprek waarbij de deelnemers gevraagd werd hun scherm te delen terwijl ze door de webapplicatie gingen. In deze think-aloud werd er gefocust op de explanationinterface van groep C, aangezien deze de meest complexe is en ook alle elementen bevat die aanwezig zijn in de explanationinterface van groep A en groep B. Ook werd er feedback gevraagd over de vragenlijsten aan het einde van de gebruikersstudie. De nadruk van deze think-aloud lag voornamelijk bij de duidelijkheid van de interface, begrijpbaarheid van de explanations, toelichtingen en taak, en op de begrijpbaarheid van de gestelde vragen. De verzamelde feedback is te zien in appendix C.

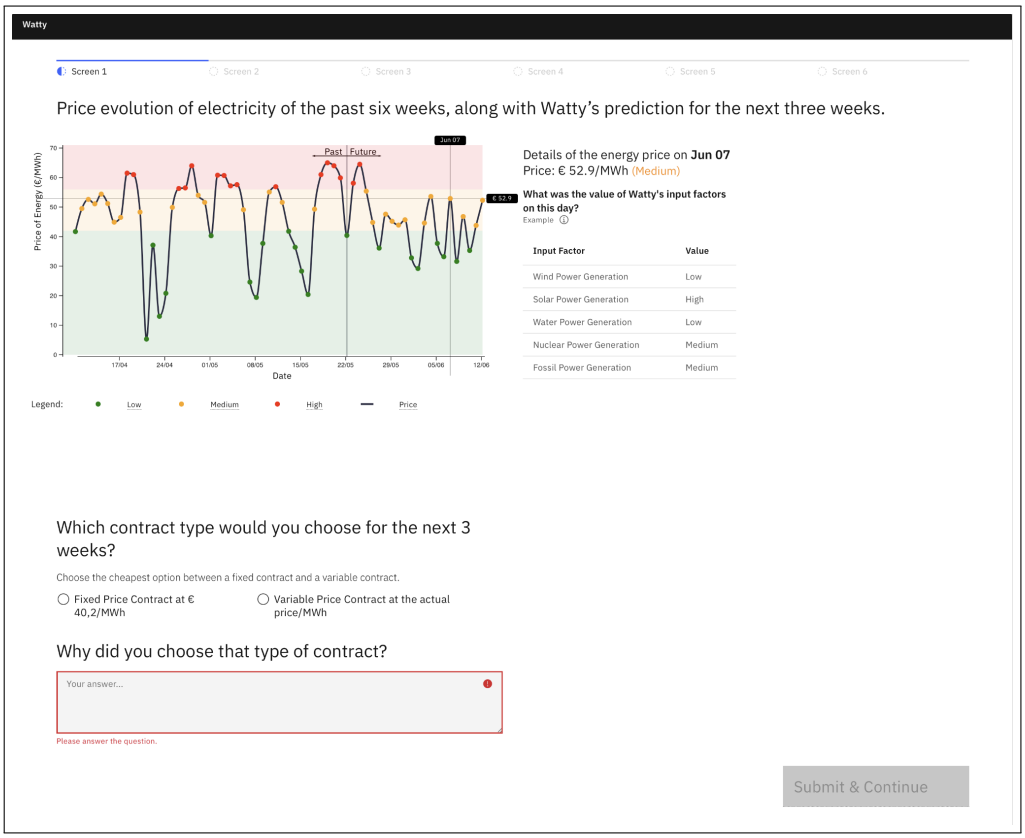
Een vaak voorkomende opmerking bij de deelnemers van deze think-aloud studie was dat de periode van het te kiezen contract (één jaar) niet afgestemd was op de periode van de voorspelling die gegeven werd (één week). Ook het feit dat er slechts voor twee weken data uit het verleden getoond werd vormde een probleem. Een oplossing hiervoor is om meer data te tonen en de periode van voorspelling gelijk te stellen met de periode van het te kiezen contract. Een ander vaak gehoord probleem tijdens de think-aloud was dat het voor de deelnemers niet altijd duidelijk was waar de verleden data overging in voorspellingen. Het toevoegen van een duidelijke scheidingslijn en labels kan hiervoor een oplossing zijn. Het was voor de deelnemers ook nog steeds niet altijd duidelijk wat de explanations net betekende en wat het verschil was tussen de verschillende explanations. Het toevoegen van informatiepopups kan hier een oplossing voor zijn.

Er kwamen ook enkele problemen naar boven die inherent zijn aan de gegeven explanations, bijvoorbeeld het feit dat de gegeven data en explanation niet altijd overeenkwam met de verwachtingen. Dit zorgde voor verwarring bij de deelnemers. Dit komt door de limitaties van het model en de explanationmethodes. Deze worden verder besproken in hoofdstuk 6.

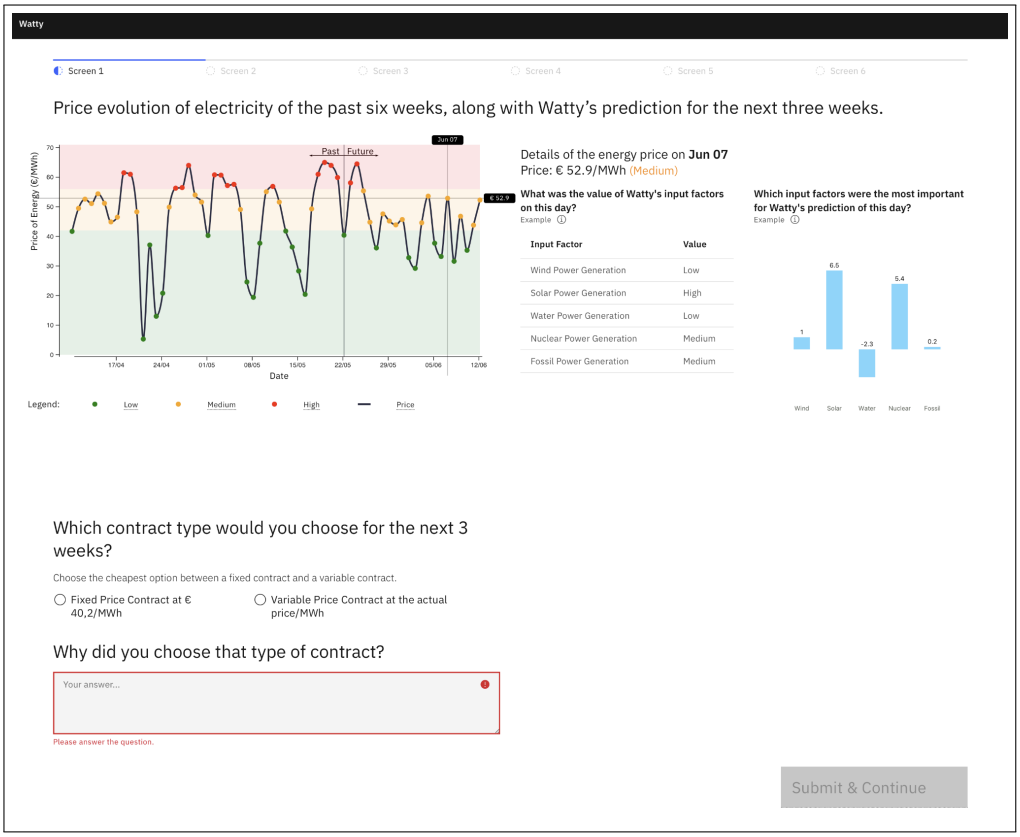
4.5 Finale prototype

In het finale prototype werd er waar mogelijk rekening gehouden met de gegeven feedback van de deelnemers aan de think-aloud studies. Ook werden er in de ontwikkelingsperiode tussen de think-aloud studies nog verschillende iteraties gedaan die informeel afgetoetst werden bij gebruikers in de omgeving van de auteur. Het eindresultaat is de applicatie waar de gebruikersstudie in uitgevoerd zal worden. In dit finale prototype werden er heel wat aanpassingen doorgevoerd. Er wordt nu zes weken aan verleden data getoond en drie weken aan voorspelde data. Ook de duur van het te kiezen energiecontract is afgestemd op de hoeveelheid aan voorspelde data die getoond

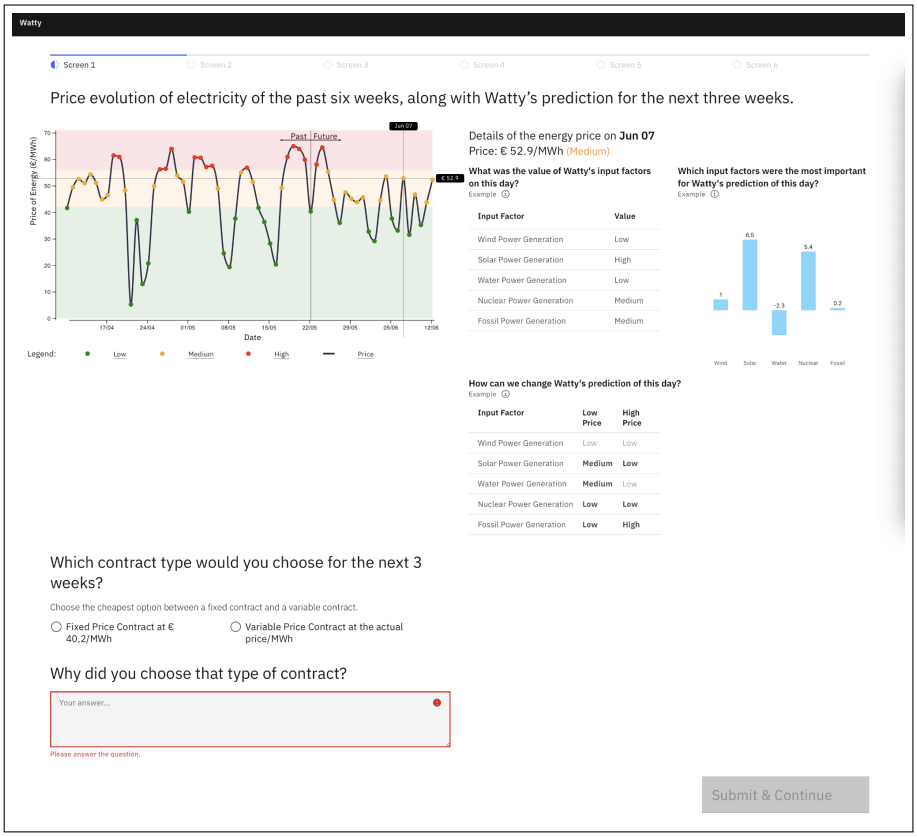
wordt aan de gebruiker. Er is een duidelijke en gelabelde scheidingslijn toegevoegd aan de lijngrafiek zodat gebruikers makkelijker kunnen zien waar de verleden data stopt en de voorspellingen beginnen. Om meer duidelijkheid te scheppen in de gegeven explanations werden er informatiepopups toegevoegd die de explanations kaderen en ook steeds een voorbeeld bevatten van hoe de explanation geïnterpreteerd moet worden. Ook werd de interactiviteit bij de counterfactual explanations verwijderd. Beide alternatieve voorspellingsklassen worden steeds getoond zodat de gebruiker niet meer moet kiezen aangezien dat voor verwarring zorgde. Er is ook een voortgangsbalk toegevoegd aan de applicatie zodat gebruikers altijd weten waar ze zich bevinden in de gebruikersstudie. Ook werden er enkele visuele veranderingen doorgevoerd zoals het groter en responsief maken van de lijngrafiek. Figuur 4.14, fig. 4.15 en fig. 4.16 tonen de explanation interfaces die deelnemers aan de gebruikersstudie in respectievelijk groep A, groep B en groep C te zien krijgen.



Figuur 4.14: Finale prototype voor groep A.



Figuur 4.15: Finale prototype voor groep B.

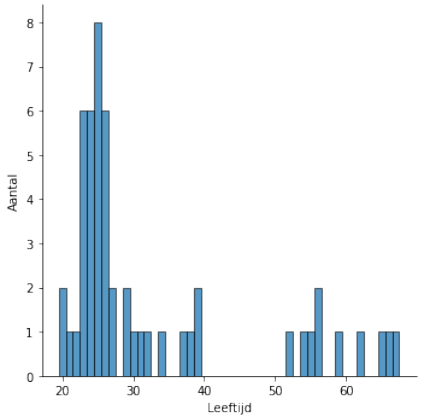


Figuur 4.16: Finale prototype voor groep C.

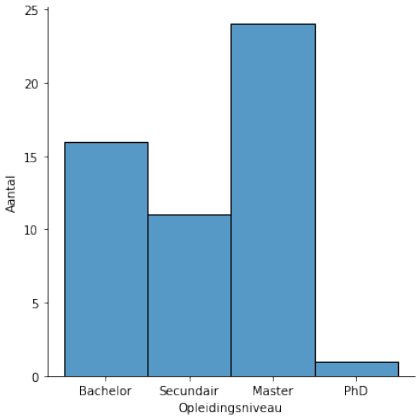
Hoofdstuk 5

Resultaten

Er namen in totaal 67 gebruikers deel aan de gebruikersstudie. 15 gebruikers werden weggefilterd omdat ze de studie niet volledig doorliepen of de studie voltooiden in minder dan 5 minuten. Figuur 5.1 en fig. 5.2 geven de leeftijdsverdeling en het opleidingsniveau van de gebruikers weer. 18 van de gebruikers werden toegewezen aan groep A, 15 gebruikers aan groep B en 19 gebruikers aan groep C.



Figuur 5.1: Leeftijdsverdeling van de gebruikers.



Figuur 5.2: Opleidingsniveau van de gebruikers.

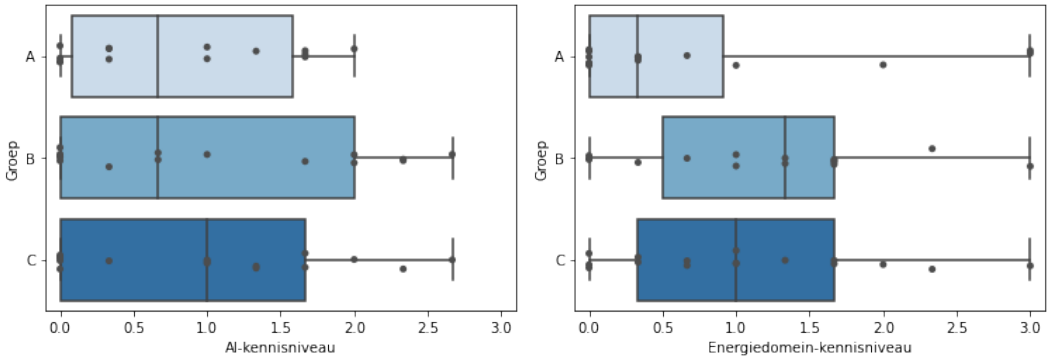
5.1 Kwantitatieve resultaten

5.1.1 Kennisniveau

Zes van de 52 gebruikers bleken AI-expert te zijn en twee van de gebruikers bleken ervaren te zijn in het energiedomein. Er was één gebruiker die expert was in zowel het AI- als het energiedomein. Na deze filtering bleven er nog 14 gebruikers over in groep A, 15 gebruikers in groep B en 16 gebruikers in groep C, voor een totaal van

45 novice gebruikers die geen kennis hebben van het AI en/of energiedomein.

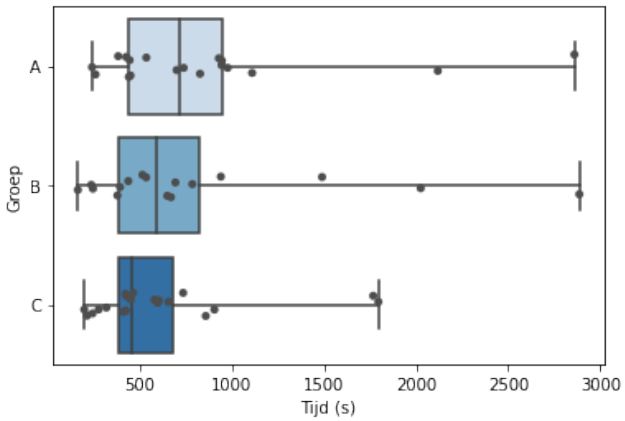
Vervolgens werd het gemiddelde kennisniveau per groep voor het AI-domein en energiedomein vergeleken. Figuur 5.3 toont deze vergelijking. Om te bepalen of er een significant verschil is in het kennisniveau tussen de verschillende groepen werd er een one-way ANOVA-test gedaan. Voor een significantie van $\alpha = 5\%$ is het verschil in AI-kennisniveau tussen de verschillende groepen niet significant (p-waarde = 0.742). Ook het verschil in energiedomein-kennisniveau is niet significant (p-waarde = 0.489). Er wordt dus aangenomen dat de groepen gelijk verdeeld zijn op gebied van kennisniveau.



Figuur 5.3: Boxplots van het kennisniveau van het AI- en energiedomein.

5.1.2 Tijdsbesteding

Tijdens het experiment werd er bijgehouden hoeveel tijd de gebruikers in totaal spendeerden in de explanationinterface van de gebruikersstudie.Figuur 5.4 toont de gespendeerde tijd in de verschillende groepen.



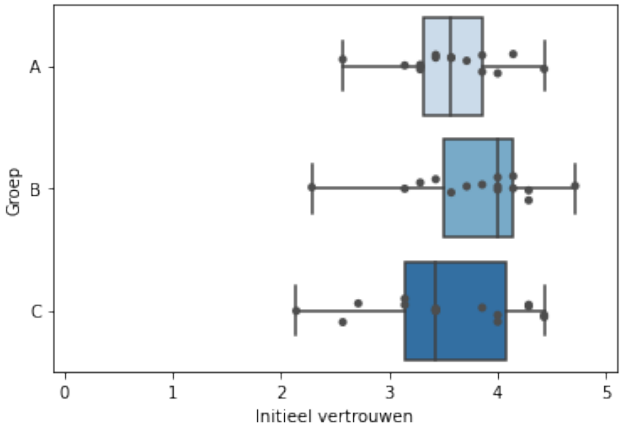
Figuur 5.4: Boxplots van de gespendeerde tijd in de explanationinterface.

Een one-way ANOVA-test uitgevoerd op de verzamelde tijdsdata met een significantie van $\alpha = 5\%$ toont aan dat het verschil in gespendeerde tijd tussen de verschillende groepen niet significant is (p-waarde = 0.456). Wel valt er op te merken dat hoewel de gebruikers gericht een taak moesten vervolledigen, ze gemiddeld slechts 88 seconden spendeerden per scherm in de explanationinterface. Er valt ook op te merken dat, hoewel het verschil in tijdsbesteding tussen de groepen niet statistisch significant is, de gebruikers van groep C gemiddeld 30 seconden minder spenderen per scherm dan gebruikers van groep A of groep B.

5.1.3 Vertrouwen

Figuur 5.5 toont het initiële vertrouwen van de gebruikers in de verschillende groepen. Het is duidelijk dat een groot deel van de gebruikers reeds van nature positief vertrouwen heeft in Artificiële Intelligentie.

Om te bepalen of er een significant verschil is in het initieel vertrouwen tussen de verschillende groepen werd er een one-way ANOVA-test uitgevoerd. Voor een significantie van $\alpha = 5\%$ is het verschil in initieel vertrouwen tussen de verschillende groepen niet significant (p-waarde = 0.486). Er wordt dus aangenomen dat de groepen eerlijk verdeeld zijn op gebied van initieel vertrouwen.

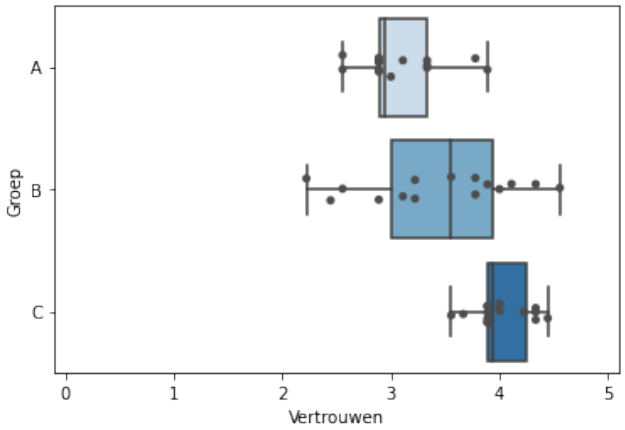


Figuur 5.5: Boxplots van het initieel vertrouwen.

Figuur 5.6 toont het verschil in het gemiddelde vertrouwen tussen de verschillende groepen.

Een one-way ANOVA-test tussen het gemiddelde van het vertrouwen van de gebruikers in de drie groepen toon voor een significantie van $\alpha = 5\%$ aan dat het verschil in vertrouwen tussen de groepen significant is (p-waarde = 2.8e−5). Om te bepalen tussen welke groepen het verschil in vertrouwen significant is werd er gebruik gemaakt van Tukey's range test. Voor een significantie van $\alpha = 5\%$ blijkt dat het verschil in vertrouwen tussen groep A en groep B niet significant is (p-waarde =

0.145) terwijl het verschil in vertrouwen tussen groep A en groep C (p-waarde = 0.001) en groep B en groep C (p-waarde = 0.007) wel significant is.

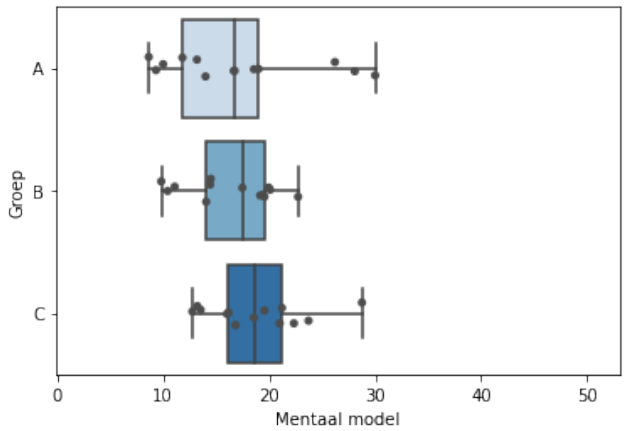


Figuur 5.6: Boxplots van het vertrouwen van gebruikers in het systeem.

5.1.4 Mentaal model

Er waren enkele gebruikers die geen voorspellingen hebben gemaakt bij de model output prediction vraag in de gebruikersstudie (één gebruiker in groep A, twee gebruikers in groep B en drie gebruikers in groep C). Deze gebruikers werden niet in beschouwing genomen voor de vergelijking van het mentaal model. Figuur 5.7 toon het verschil van de RMSE score tussen de verschillende groepen.

Ook hier werd er weer gebruik gemaakt van een one-way ANOVA-test om te bepalen of er een statistisch significant verschil is in het mentaal model van de gebruikers tussen de verschillende groepen. Voor een significantie van $\alpha = 5\%$ is het verschil in mentaal model tussen de groepen niet significant (p-waarde = 0.533).



Figuur 5.7: Boxplots van de RMSE van de model output prediction van de gebruikers.

5.2 Kwalitatieve resultaten

Tabel 5.1 geeft een samenvatting van de verschillende thema's die naar boven komen uit de antwoorden van de gebruikers op de open vragen over het nut en de kwaliteit van de SHAP-explanations. Van de 31 novice gebruikers die in aanraking gekomen zijn met de SHAP-explanation (in groep B en groep C) geven er 18 aan dat ze de SHAP-explanation onduidelijk vinden. Zelfs na het lezen van de infopopup begrijpt dit deel van de gebruikers niet wat de waarden net betekenen. Van de overige 13 gebruikers geven er 2 aan dat de SHAP-explanation een goede aanvulling is op de inputfactoren, maar nog steeds niet voldoende informatie geeft. Eén gebruiker geeft specifiek aan dat de SHAP-explanation zorgt voor meer vertrouwen in het systeem.

Tabel 5.2 geeft een samenvatting van de verschillende thema's die naar boven komen uit de antwoorden van de gebruikers op de open vragen over het nut en de kwaliteit van de CF-explanations. Van de 16 gebruikers aan wie de counterfactual explanations getoond werden geven er twee aan dat ze de explanation onduidelijk vinden en niet goed begrijpen welke informatie de explanation meegeeft. Een ander thema dat hier naar bovenkomt is de impact die de gebruiker heeft op het beïnvloeden van de voorspelling. Er zijn gebruikers die opmerken dat er explanations worden gegeven waar de gebruiker zelf geen invloed op heeft.

Thema	Freq.	Voorbeeldcitaat
Begrip (Is de explanation duidelijk voor de gebruiker)		
Onduidelijke explanation	18	'Helemaal niet duidelijk, dit heb ik dus ook niet gebruikt om mijn keuzes op te baseren'
Te technische explanation	2	'Ik ben een technisch persoon, dus ik begrijp de grafiek en ik vind het interessant om te zien wat er achter de schermen gebeurt. Maar minder technische mensen zullen in de war raken denk ik.'
Opgedane inzichten	3	'Nuttig, je ziet welke invloed de energiemix heeft op de prijs.'
Stijl (Is de explanation visueel duidelijk en aangenaam)		
Visuele aantrekkingskracht	2	'Nuttig, door middel van de grafieken krijg je ook meteen een visueel zicht. Helpt om bovenstaande input factors nog beter te begrijpen.'
Informatieweergave	1	'Dit was nuttig, maar er zou waarschijnlijk een as bij kunnen die aangeeft dat naar beneden betekent dat de kosten lager zijn en naar boven dat ze hoger zijn. De meeste mensen lezen waarschijnlijk de info niet dus zouden het wel niet snappen denk ik'
Volledigheid (Geeft de explanation alle informatie weer)		
Meer details nodig	2	'Geeft een duidelijker beeld dan inputfactoren maar nog steeds onvoldoende duidelijkheid naar verbruik/ onderling percentage.'
Vertrouwen (Zorgt de explanation voor vertrouwen in het systeem)		
Geeft vertrouwen in systeem	1	'Nuttig, als ik het gewicht zie van de inputfactoren op watty's voorspelling heb ik het gevoel dat ik het meer kan vertrouwen.'
Verwachtingen (Komt de explanation overeen met de verwachtingen van de gebruiker)		
Eigen verwachtingen	4	'Ik dacht dat de factoren de prijs anders zouden beïnvloeden. Ik dacht dat de prijs omhoog zou gaan als nucleaire en fossiele energieopwekking omhoog gaan.'

Tabel 5.1: Thematische analyse van de antwoorden van de gebruikers op de open vragen over SHAP-explanations.

Thema	Freq.	Voorbeeldcitaat
Begrip (Is de explanation duidelijk voor de gebruiker)		
Onduidelijke explanation	3	'Deze informatie is te gecompliceerd voor snelle interpretatie.'
Opgedane inzichten	6	'Nuttig, je kan leren wat er moet veranderen om de prijs beter (of slechter) te maken.'
Stijl (Is de explanation visueel duidelijk en aangenaam)		
Opsplitsing in categorieke waarden niet duidelijk	1	'Nee, het is verwarrend om 'Low', 'Medium' en 'High' te gebruiken voor de prijs en de kans dat die zou veranderen'
Vetgedrukte factoren onduidelijk	2	'Nee, ik begrijp niet zo goed de meerwaarde van de vetgedrukte woorden.'
Verwachtingen (Komt de explanation overeen met de verwachtingen van de gebruiker)		
Impact	2	'Niet nuttig, we kunnen de input factors immers zelf niet aanpassen.'

Tabel 5.2: Thematische analyse van de antwoorden van de gebruikers op de open vragen over CF-explanations.

Hoofdstuk 6

Discussie

In dit hoofdstuk worden de resultaten van de analyse in hoofdstuk 5 besproken en gebruikt om een antwoord te formuleren op de eerder vooropgestelde onderzoeksvragen. Ook worden de limitaties van dit onderzoek besproken en wordt er gekeken naar mogelijk toekomstig onderzoek.

6.1 Antwoord op de onderzoeksvragen

> RQ1: Hoe beïnvloedt het gebruik van Shapley additive explanations het begrip van en het vertrouwen in een black-boxmodel bij AI-novices?

Het gebruik van SHAP-explanations heeft geen invloed op het vertrouwen in een black-boxmodel bij AI-novices. Uit de geanalyseerde data blijkt dat er geen significant verschil is tussen gebruikers die enkel inputfactoren getoond krijgen (groep A) en gebruikers die als aanvulling SHAP-explanations getoond krijgen (groep B). Het vertrouwen in het systeem bij zowel groep A als groep B is relatief klein. Er werd ook geen significant verschil gevonden in het begrip van het model tussen gebruikers van groep A en groep B.

Dat SHAP-explanations in deze gebruikersstudie niet tot meer vertrouwen leiden bij AI-novices kan meerdere mogelijke verklaringen hebben. Uit de thematische analyse van de gebruikersstudie komt naar boven dat 18 van de 31 gebruikers die SHAP-explanations hebben gekregen aangeven dat ze de explanations niet duidelijk vinden. Ook al is er een infopopup die uitlegt wat de SHAP-explanation net voorstelt, zijn er gebruikers die niet begrijpen wat de explanation net wil zeggen. Of gebruikers deze infopopup gelezen hebben werd niet bijgehouden. Deze informatie had mogelijk meer inzicht kunnen geven. Er zijn ook enkele gebruikers die aanhalen dat de explanation te technisch is, en kan zorgen voor problemen bij mensen die een beperkte ervaring hebben met het lezen en interpreteren van grafieken. Dit is een mogelijk probleem voor de standaard implementatie van SHAP door Lundberg et al [42]. Hoewel er in deze gebruikersstudie slechts voor één van de mogelijke visualisaties is gekozen,

bestaan de andere visualisaties uit complexere grafieken, die potentieel voor nog meer verwarring kunnen zorgen. Deze complexiteit zorgt ervoor dat gebruikers mogelijk niet begrijpen wat het systeem doet en hoe het tot een voorspelling komt, wat er toe leidt dat gebruikers liever vertrouwen op hun eigen intuïtie. Het kan dus zijn dat een alternatieve, minder visueel complexe voorstellingswijze van de SHAP-explanation wenselijk is wanneer de explanation geïnterpreteerd moet worden door AI-novices.

Een andere mogelijke verklaring van het gebrek aan vertrouwen in SHAP-explanations kan zijn dat ze niet altijd overeenkomen met de verwachtingen van gebruikers. Dit wordt ook zo aangegeven door een aantal gebruikers in de gebruikersstudie. Eén van de gebruikers geeft als opmerking: *'Ik dacht dat de factoren de prijs anders zouden beïnvloeden. Ik dacht dat de prijs omhoog zou gaan als nucleaire en fossiele energieopwekking omhoog gaan.'*. Door de energiecrisis die op dit moment gaande is, komt er veel in het nieuws over de energieprijzen. Mensen hebben vaak al een initieel mentaal model van hoe bepaalde energiesoorten de energieprijs beïnvloeden, ook al strookt dit niet per se met de werkelijkheid. Wanneer de SHAP-explanations ingaan tegen een verband dat de gebruiker ziet in zijn eigen mentaal model kan dit zorgen voor een gebrek aan vertrouwen in de explanation. Dit gebrek aan vertrouwen in de explanation kan op zijn beurt zorgen voor wantrouwen in het systeem. Riveiro et al. [59] ondervonden reeds dat wanneer een systeem ingaat tegen de verwachtingen van de gebruikers het zeer moeilijk is om een explanation te geven die de gebruiker ook effectief zal vertrouwen en zal aanvaarden. De mogelijke inconsistentie van SHAP-explanations zoals bestudeerd door Gosiewska et al.[25] speelt eveneens mogelijk een rol. Het kan bijvoorbeeld voorkomen dat gebruikers voor een bepaalde voorspelling een SHAP-explanation krijgen die zegt dat nucleaire energie een kleine positieve invloed heeft en fossiele energie een grote positieve invloed heeft, en voor een dag met een gelijkaardige voorspelling krijgen ze de tegenovergestelde explanation. Deze tegenstelling kan ervoor zorgen dat de gebruiker minder vertrouwen heeft in de explanations en bijgevolg ook in het systeem omdat ze ervan uitgaan dat de explanation accuraat is ten opzichte van het systeem en dus het beslissingsproces van het systeem accuraat uitlegt.

Kaur et al. [32] komen in eerder onderzoek tot de conclusie dat SHAP-explanations bij AI-experts zorgt voor blind vertrouwen in de explanations en het model. Dat dit bij de AI-novices in deze gebruikersstudie niet het geval is kan mogelijk komen door het verschil in perceptie van de explanationmethode. In dit eerder onderzoek wordt expliciet aan de AI-experts vermeld dat er gebruik wordt gemaakt van SHAP-explanations. Het merendeel van de AI-experts zijn op de hoogte van het feit dat SHAP-explanations veelgebruikt en publiek beschikbaar zijn. Dit wordt gezien als een verklaring voor een blind vertrouwen in de explanations en bijgevolg ook in het model. In de gebruikersstudie van deze masterproef werd er niet vermeld welke specifieke soort explanations er getoond worden. De gebruikers zijn AI-novices, dus hebben zelf geen indicatie van welk soort explanations er gebruik wordt gemaakt en kunnen niet opzoeken welke waarde de explanations hebben en of ze te vertrouwen zijn.

Dat het mentaal model van AI-novices die in deze gebruikersstudie SHAP-explanations getoond krijgen niet beter is dan dat van gebruikers die geen explanations getoond krijgen kan mogelijk verklaard worden door een combinatie van de reeds aangehaalde verklaringen voor het gebrek aan vertrouwen. Het onduidelijk zijn van de explanation zorgt ervoor dat de gebruiker geen tot weinig nuttige informatie uit de explanation kan halen. Iets dat je niet begrijpt kan immers weinig tot geen inzicht opleveren. Ook het té technisch zijn van de explanation kan hier een bijdrage leveren en ervoor zorgen dat gebruikers geen inzicht halen uit de gegeven explanations. Het kan zijn dat, door de technische aard van de explanation, de gebruikers een te grote inspanning moeten leveren om de explanation te kunnen begrijpen. Dat dit kan leiden tot een gebrek aan vertrouwen en een vermindering in begrip werd reeds in eerder onderzoek ondervonden door Hudon et al. [30]. Uit de verzamelde gegevens over de tijdsbesteding van de gebruikers blijkt dat de gebruikers in groep B gemiddeld evenveel tijd hebben doorgebracht in de explanationinterface als gebruikers van groep A. Het begrijpen en interpreteren van de gegeven SHAP-explanations kost mogelijk meer tijd dan het begrijpen van enkel de inputfactoren. Indien gebruikers hier de tijd niet voor hebben genomen kan dit mogelijk leiden tot een gebrek aan verworven inzicht en dus een gebrek aan een beter mentaal model.

RQ2: Leiden counterfactual explanations als toevoeging voor Shapley additive explanations tot een groter vertrouwen in en een beter begrip van een black-boxmodel voor AI-novices?

Het toevoegen van counterfactual explanations aan SHAP-explanations heeft een significant positieve invloed op het vertrouwen in een black-boxmodel bij AI-novices. Er werd geen significant verschil gevonden in het begrip van het model tussen gebruikers van groep B en groep C.

Dat gebruikers meer vertrouwen hebben in het systeem wanneer er counterfactual explanations toegevoegd worden kan verschillende mogelijke oorzaken hebben. Hoewel een klein deel van de gebruikers aangeeft dat de explanation complex is, geeft de meerderheid aan dat de counterfactual explanations nuttig en begrijpbaar zijn. Eén van de gebruikers geeft als opmerking: *'...hierdoor kan je beter inschatten welke prijsinvloed elke energiebron heeft op het geheel.'*. Hoewel er geen verbetering is in het mentaal model en de gebruikers in werkelijkheid niet beter kunnen inschatten welke prijsinvloed elke energiebron heeft op het geheel, is er toch meer vertrouwen in het systeem. Eerder onderzoek van onder andere Byrne et al [11] en [19] heeft uitgewezen dat mensen een voorkeur hebben voor counterfactual explanations en dat het gebruik van dit soort explanations de (waargenomen) transparantie van het systeem verhoogt. Mensen leggen vaak zelf dingen uit aan de hand van contrafeiten. Dit kan er dus voor zorgen dat de counterfactual explanations natuurlijker overkomen voor mensen, beter begrijpbaar zijn en dus resulteren in een verhoogd (maar niet altijd terecht)

vertrouwen in het systeem. Hoewel enkele gebruikers specifiek aangeven dat ze de counterfactual explanations duidelijker vinden dan de SHAP-explanations zijn er geen gebruikers die een specifieke voorkeur hebben aangegeven. Waar SHAP-explanations bij AI-experts zorgen voor blind vertrouwen lijkt het erop dat counterfactual explanations mogelijk hetzelfde doen bij AI-novices. Warren et al.[71] verkregen gelijkaardige resultaten bij hun onderzoek. Ze komen tot de conclusie dat er voorzichtig omgegaan moet worden met counterfactual explanations aangezien een deel van de gebruikers dankzij counterfactual explanations 'een goed gevoel' krijgt over het systeem, zonder enig inzicht te hebben in hoe het systeem tot een voorspelling is gekomen of hoe het werkt.

Warren et al. [71] komen in hun onderzoek wel tot de conclusie dat wanneer gebruikers gevraagd wordt om model output prediction te doen, het gebruik van counterfactual explanations in het algemeen leidt tot een licht verhoogde accuraatheid en dus beter mentaal model. Dat hetzelfde resultaat in deze gebruikersstudie niet naar boven komt kan liggen aan het feit dat de tijdsbesteding ook hier, net zoals bij SHAP-explanations, eerder beperkt is. In een explanationinterface waarbij er zowel SHAP-explanations als counterfactual explanations getoond worden lijkt het dat de hoeveelheid tijd die de gebruikers gemiddeld spendeerden niet voldoende is om alle informatie op te nemen en de interface voldoende te exploreren om zo tot inzichten te komen. Het valt ook op dat de gebruikers van groep C gemiddeld minder tijd spenderen per scherm dan de gebruikers van groep A en groep B. Een mogelijke oorzaak hiervan kan zijn dat het combineren van de twee explanations leidt tot een overvloed aan informatie die de gebruikers afschrikt om verder te exploreren. Ook het feit dat gebruikers de counterfactual explanations mogelijk incorrect interpreteren kan een rol spelen bij het gebrek aan een beter mentaal model. Dat niet alle gebruikers counterfactual explanations correct interpreteren blijkt uit de thematische analyse. Een van de gebruikers geeft de volgende opmerking over counterfactual explanations: *'... het laat zien welke inputfactoren een effect hebben op prijsstijgingen'*. Dit geeft aan dat sommige gebruikers de gegeven explanation generaliseren op een foute manier. Indien een counterfactual aangeeft dat de energieprijs laag zal zijn als er veel windenergie en zonne-energie is terwijl de andere inputfactoren laag zijn wil dat enkel zeggen dat deze counterfactual geldt voor deze exacte waarden. Er kan niet veralgemeend worden dat veel wind- en zonne-energie altijd tot een lage energieprijs zal leiden. Om dit soort conclusies te kunnen trekken moeten meerdere counterfactuals geïnspecteerd worden, en zelfs dan kan er niet met zekerheid geconcludeerd worden dat dit verband altijd geldt.

Uit de verzamelde data in deze masterproef is het moeilijk een definitief oordeel te vellen over welke explanation te verkiezen valt voor AI-novices. Hoewel counterfactual explanations belovend lijken, en zorgen voor meer vertrouwen bij gebruikers, werd er geen verschil in begrip vastgesteld. Wel vinden gebruikers counterfactual explanations duidelijker dan de SHAP-explanations. Dit geeft een indicatie van het feit dat gebruikers van nature uitleg gegeven aan de hand van contrafeiten prefereren, en dat counterfactual explanations mogelijk beter aansluiten bij het denkpatroon van de mens.

6.2 Limitaties

Limitaties van het voorspellingsmodel

In deze masterproef is er geprobeerd om het model los te koppelen van de explanations. In dit onderzoek heeft het model een puur utilitaire functie en dient het enkel als middel om de explanations te kunnen genereren en evalueren. Achteraf gezien is dit een moeilijk gegeven. Uit de thematische analyse komt naar boven dat vele gebruikers al specifieke verwachtingen hadden over wat het systeem zou kunnen doen. Wanneer deze verwachtingen niet ingelost worden kan dit de resultaten mogelijk beïnvloeden. In toekomstig werk moet er zorgvuldig omgegaan worden met de keuze van het toepassingsdomein om meer accurate resultaten te bekomen. Hoewel er zorg is gestoken in het zo accuraat mogelijk maken van het gebruikte voorspellingsmodel heeft het toch tekortkomingen:

- Het model is tweelagig. Dit introduceert afwijkingen op twee verschillende plaatsen wat er voor zorgt dat het model minder accuraat is.
- Het model neemt enkel de productiefactoren in rekening. In werkelijkheid spelen er nog veel meer factoren mee zoals het weer, de afnamehoeveelheid, het seizoen en zelf het dagelijkse nieuws. Het ontwerpen van een model dat de energieprijs accuraat voorspelt is complex en vormt een onderzoeksgebied op zichzelf [15, 50, 54, 9].
- Het model gaat uit van de onafhankelijkheid van de verschillende inputfactoren. In werkelijkheid is er wel een afhankelijkheid tussen de inputfactoren. De energiemarkt is steeds gebalanceerd. Dit wil zeggen dat er steeds exact evenveel productie van energie is als dat er afname is. Dit leidt tot afhankelijkheid van de inputfactoren. Wanneer er bijvoorbeeld meer zonne-energie geproduceerd wordt zullen er andere inputfactoren zijn waarvan de injectie op het energienet zal moeten verminderen om deze balans te houden.
- De data waarop het model getraind is bevat missende en onaannemelijke waarden. Deze werden waar mogelijk vervangen door middel van k-Nearest Neighbor imputatie.
- De data waarop het model getraind is komt uit het verleden (2015 tot en met 2021). Door de huidige geopolitieke situatie is de energieprijs in 2022 meer dan verdrievoudigd ten opzichte van de energieprijs van 2021. De data uit het verleden is dus niet meer representatief voor de huidige situatie. Het model houdt hier geen rekening mee.

Limitaties van de gebruikersstudie

De gebruikersstudie kent ook enkele limitaties:

- Hoewel in de gebruikersstudie vermeld wordt dat het systeem niet de bedoeling heeft om de energieprijs perfect te voorspellen blijkt uit een aantal opmerkingen

van gebruikers dat deze boodschap niet altijd is overgekomen. Dit kan mogelijk opgelost worden door de studie beter te kaderen bij de gebruikers.

- Na analyse van de tijdsbesteding blijkt dat de gebruikers mogelijk niet voldoende tijd spenderen in de explanationinterface. Dit kan de resultaten vertekenen aangezien gebruikers voldoende tijd nodig hebben om de explanations te interpreteren en zo ook het gedrag van het model te kunnen linken aan de explanations. Het kan ook zijn dat het niet altijd realistisch is om van gebruikers te verwachten dat ze met een beperkte uitleg zelfstandig de explanations kunnen begrijpen. Hoewel dit een nobel doel is kan het zijn dat gebruikers er baat bij hebben om een korte introductie/trainingssessie te volgen.
- Het energiedomein leek eind 2021 een interessant toepassingsdomein. De huidige geopolitieke situatie heeft ervoor gezorgd dat gebruikers over het algemeen meer begaan zijn met de problematiek rond de energieprijs en vaak een zeer sterk initieel mentaal model hebben waar ze moeilijk van afwijken. Dit komt duidelijk naar voor in antwoorden van gebruikers die aanhalen dat hoewel ze bepaalde explanations nuttig vinden en vertrouwen, ze toch nog meer vertrouwen hechten aan hun eigen intuïtie. Dit kan mogelijk zorgen voor vertekende resultaten.
- Het vertrouwen wordt in deze masterproef behandeld als een statisch gegeven. Er zijn echter ook onderzoeken die uitwijzen dat het interessant kan zijn om het vertrouwen van gebruikers te onderzoeken als een proces in de tijd [28, 29, 46], en dus over een langere termijn van gebruik van het systeem. Holliday et al. [29] ondervonden dat er verschillende patronen waarneembaar zijn in de evolutie van het vertrouwen van gebruikers in een AI-systeem. Deze patronen werden niet gecapteerd in deze masterproef.

6.3 Toekomstig werk

Het kan interessant zijn om dit onderzoek te voeren op alternatieve wijze. In plaats van te werken met een groter aantal deelnemers die meedoen aan een online experiment waaraan zelfstandig wordt deelgenomen lijkt het interessant om met een kleinere groep deelnemers te werken die de experimenten doen onder begeleiding van de onderzoekers. Tijdens de analyse van de resultaten werd er opgemerkt dat sommige gebruikers zeer summiere antwoorden geven op de open vragen, terwijl deze net tot een beter inzicht in de onderzoeksvragen leiden. Indien het experiment uitgevoerd wordt met achteraf een interview met de gebruiker kan er meer gericht vragen gesteld worden en andere inzichten opgedaan worden. Het uitvoeren van een gebruikersstudie met meer uitgebreide open vragen kan hier een andere mogelijke oplossing voor zijn.

Het toepassingsdomein variëren kan mogelijk leiden tot verbeterde resultaten. Zoals al eerder vermeld is op dit moment het energiedomein niet het meest opportune toepassingsdomein. Kiezen voor een toepassing waar gebruikers minder sterke vooroordelen over hebben kan betere resultaten opleveren. Het werken met synthetische

data kan mogelijk een andere oplossing zijn om de vooroordelen van gebruikers te counteren. Er kan dan ingespeeld worden op de verwachtingen van de gebruikers. Er moet dan wel over gewaakt worden dat de gebruikte data realistisch is en er moet voorzichtig omgegaan worden met de verkregen resultaten. Er nagegaan worden of het vertrouwen van de gebruikers in een systeem met synthetische data niet enkel voortkomt uit het voldoen aan de verwachtingen van de gebruikers.

Uit de resultaten blijkt dat counterfactual explanations zorgen voor meer vertrouwen in het systeem, maar niet voor meer begrip van het systeem. Er kan in toekomstig werk dieper in worden gegaan op hoe SHAP-explanations en counterfactual explanations mogelijk aangepast kunnen worden om beter geschikt te zijn voor AI-novices, zodat ze beiden zorgen voor meer gepast vertrouwen én meer begrip. Er kan bijvoorbeeld onderzocht worden of de explanations in natuurlijke taal geven aan gebruikers een beter alternatief is. Er kan gekeken worden welke invloed interactiviteit heeft op het vertrouwen en begrip door de explanations te geven aan de hand van een chatbot waaraan de gebruiker nog extra vragen kan stellen en verduidelijkingen kan vragen. Ook onderzoek naar het verschil in perceptie, vertrouwen en begrip van black-boxmodellen tussen AI-experts en AI-novices kan mogelijk leiden tot explanations die meer op maat zijn van novice gebruikers. Ooge et al. [53] ondervonden reeds dat het verschil in vertrouwen niet steeds verklaard kan worden enkel op basis van het kennisniveau. Het kan dus opportuun zijn om onderzoek te doen naar personaliseerbare explanations waarbij de gebruiker zelf kan aangeven op welke manier en met welke complexiteit de explanations gegeven worden.

Hoofdstuk 7

Besluit

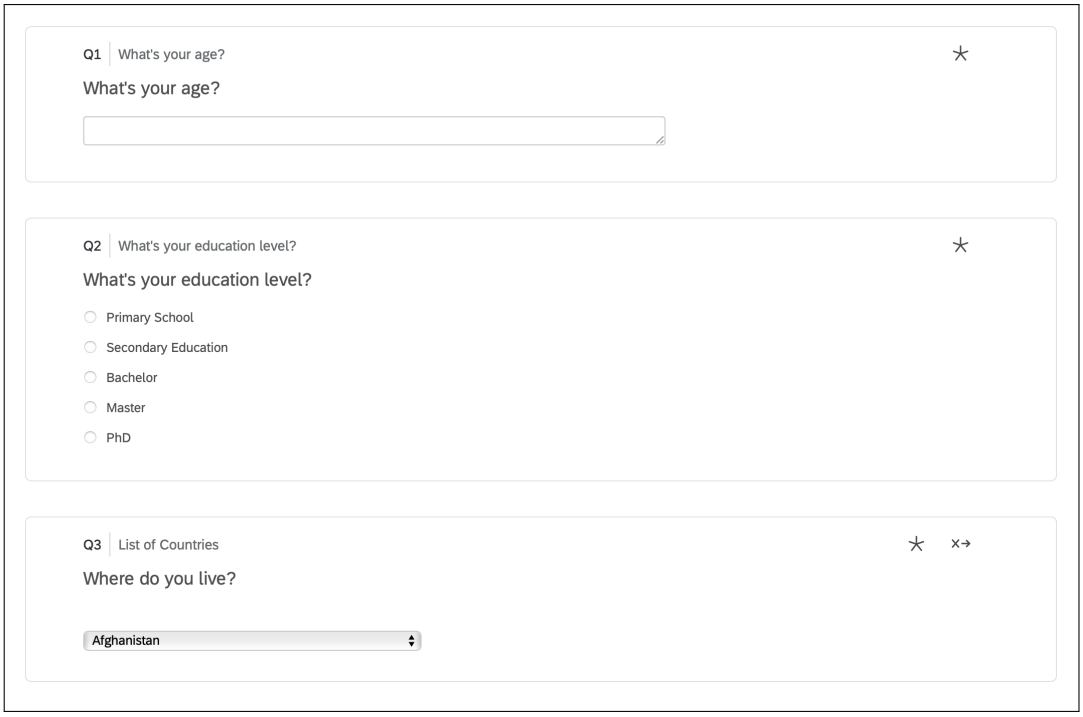
In deze masterproef werd het vertrouwen in en het begrip van AI-novices in black-boxalgoritmen bestudeerd voor twee state-of-the-art methodes voor lokale, post-hoc explanations: Shapley additive explanations en counterfactual explanations. Eerst werd er een overzicht gegeven van de relevante literatuur, waarbij de meest belangrijke concepten en bevindingen werden toegelicht. Er werd aandacht besteedt aan de karakteristieken van een kwalitatieve explanation en aan hoe explanations geëvalueerd kunnen worden. Zowel Shapley additive explanations als counterfactual explanations werden in meer detail besproken en de wiskundige fundamenten van beide methodes werden behandeld. Ook werd er een overzicht gegeven van de relevante literatuur binnen het energiedomein én het gerelateerd werk. Vervolgens werden de onderzoeksvragen voorgesteld, samen met de methodologie die gevolgd werd om een antwoord op de onderzoeksvragen te kunnen bieden. De evaluatiemethode van de gebruikersstudie werd besproken en de effectieve implementatie werd behandeld. Er werd ook dieper ingegaan op het iteratieve, gebruikersgeoriënteerde ontwikkelingsproces dat gevolgd werd bij de ontwikkeling van het prototype voor de gebruikersstudie. Hierna werden de resultaten van de gebruikersstudie geformuleerd en geanalyseerd. Uit de kwantitatieve resultaten van deze studie blijkt dat het gebruik van SHAP niet leidt tot meer vertrouwen in of meer begrip van het AI-systeem bij AI-novices. Het toevoegen van counterfactual explanations als ondersteuning bij SHAP leidt wel tot meer vertrouwen in het AI-systeem bij AI-novices, maar niet tot een beter begrip van het AI-systeem. Uit een thematische analyse blijkt dat een groot deel van de AI-novices de SHAP-explanations niet begrijpt en onduidelijk vindt. De explanations gaan vaak in tegen de verwachtingen van de gebruikers, wat kan zorgen voor een gebrek aan vertrouwen. De counterfactual explanations blijken duidelijker te zijn voor AI-novices. De gebruikers geven ook aan dat ze inzichten opdoen dankzij de counterfactual explanations, hoewel dit niet blijkt uit de kwantitatieve resultaten. Een verklaring hiervoor kan zijn dat de gebruikers niet voldoende tijd hebben gespendeerd aan de explanations in de gebruikersstudie. Dit leidt meteen ook tot een van de limitaties van dit onderzoek. De gebruikers zijn slechts beperkte tijd in contact gekomen met de explanationinterface uit de gebruikersstudie. In toekomstig werk kan er een meer uitgebreid onderzoek opgezet worden waarbij er gestructureerde diepte-

interviews afgenomen worden bij gebruikers die voor een langere periode gebruik maken van de explanationinterface. De belangrijkste conclusies uit dit onderzoek zijn dat explanations steeds afgestemd moeten worden op het doelpubliek. Ook is het belangrijk om bij onderzoek naar explanations het toepassingsgebied zorgvuldig te bepalen, en om rekening te houden met de verwachtingen van de gebruikers. Tot slot beargumenteert deze masterproef stellig dat het nuttig kan zijn om meer onderzoek te doen naar explanations specifiek gericht op AI-novices. Er moet gewaakt worden over de rechtvaardigheid van explanations en rechtvaardigheid in het gebruik van AI. We kunnen als moderne maatschappij niet toestaan dat mensen uit de boot vallen omdat ze beschikken over een beperkte technische bagage.

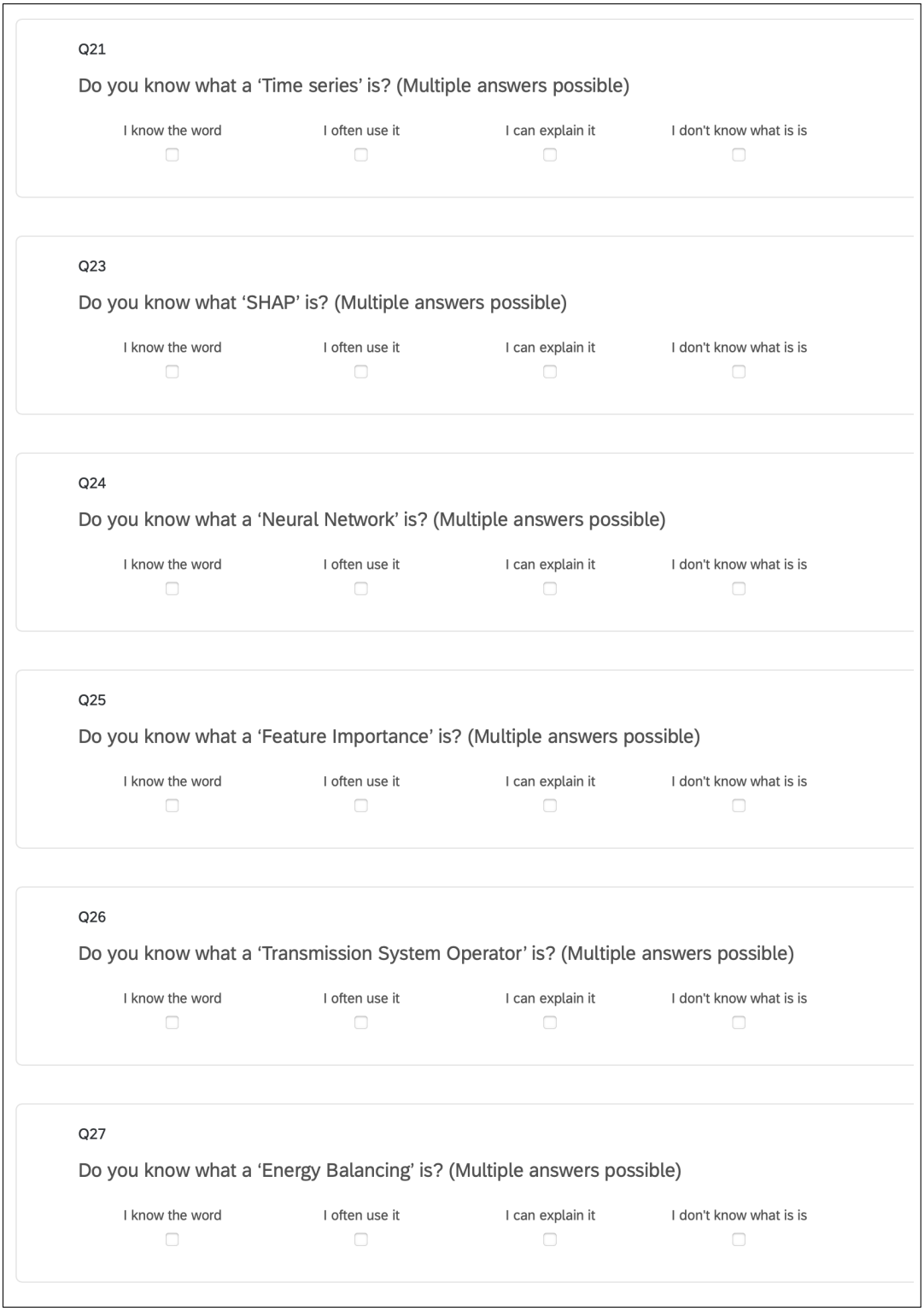
Bijlage A

Vragenlijsten

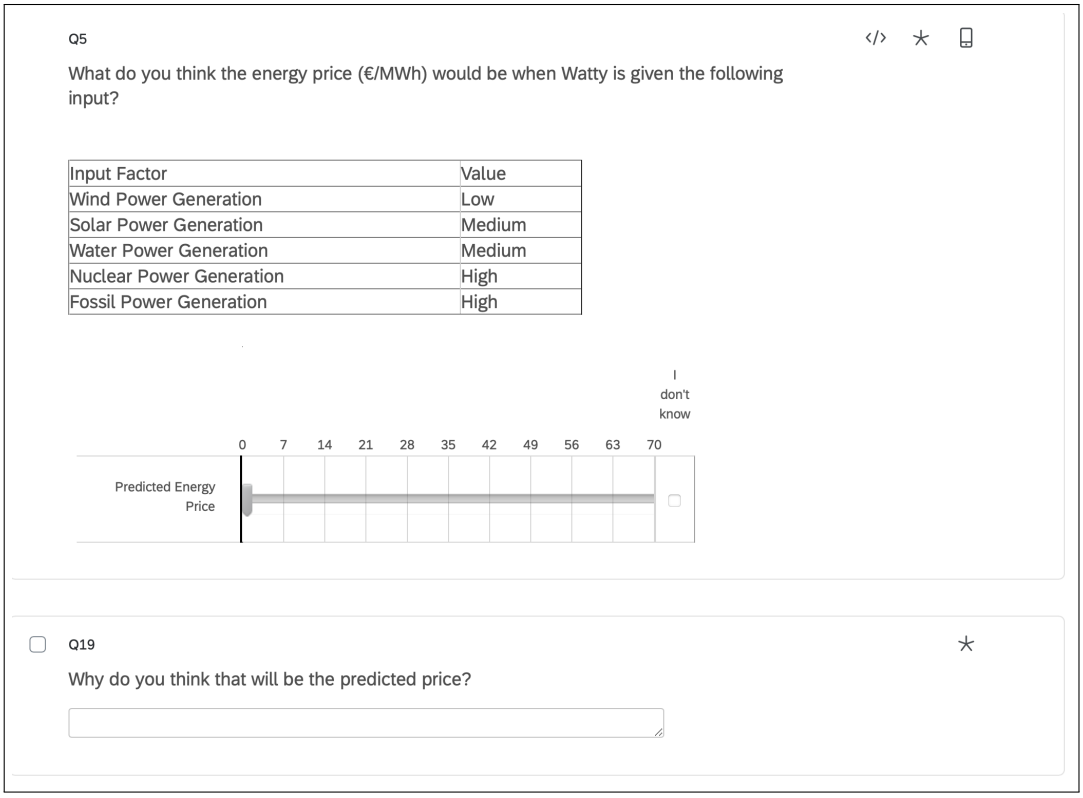
Figuur A.1 geeft de algemene vragen weer die gesteld werden aan de gebruikers. De vragen in Figuur A.2 peilen het kennisniveau van de gebruiker. Het mentaal model van de gebruiker wordt gemeten door de vragen in Figuur A.3. Figuur A.4 geeft de vragen weer die het initieel vertrouwen peilen en de vragen in Figuur A.5 peilen het vertrouwen in het systeem. De open vragen over de explanations worden getoond in Figuur A.6 tot en met fig. A.10 en de suggestievraag in Figuur A.11.



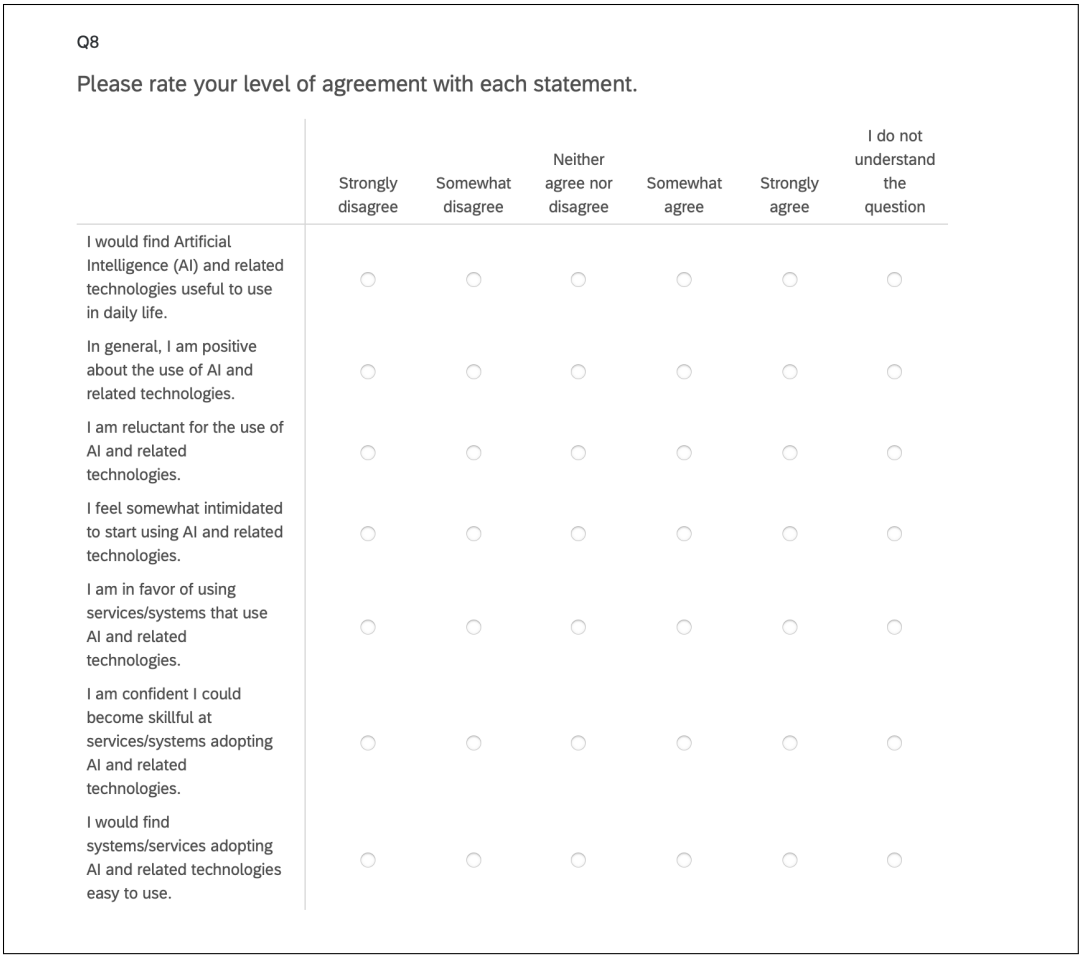
Figuur A.1: Algemene informatievragen.



Figuur A.2: Peiling van het kennisniveau van de gebruiker, naar Ooge et al. [53].



Figuur A.3: Evaluatievraag voor het mentaal model van de gebruiker, naar Mohseni et al. [47].

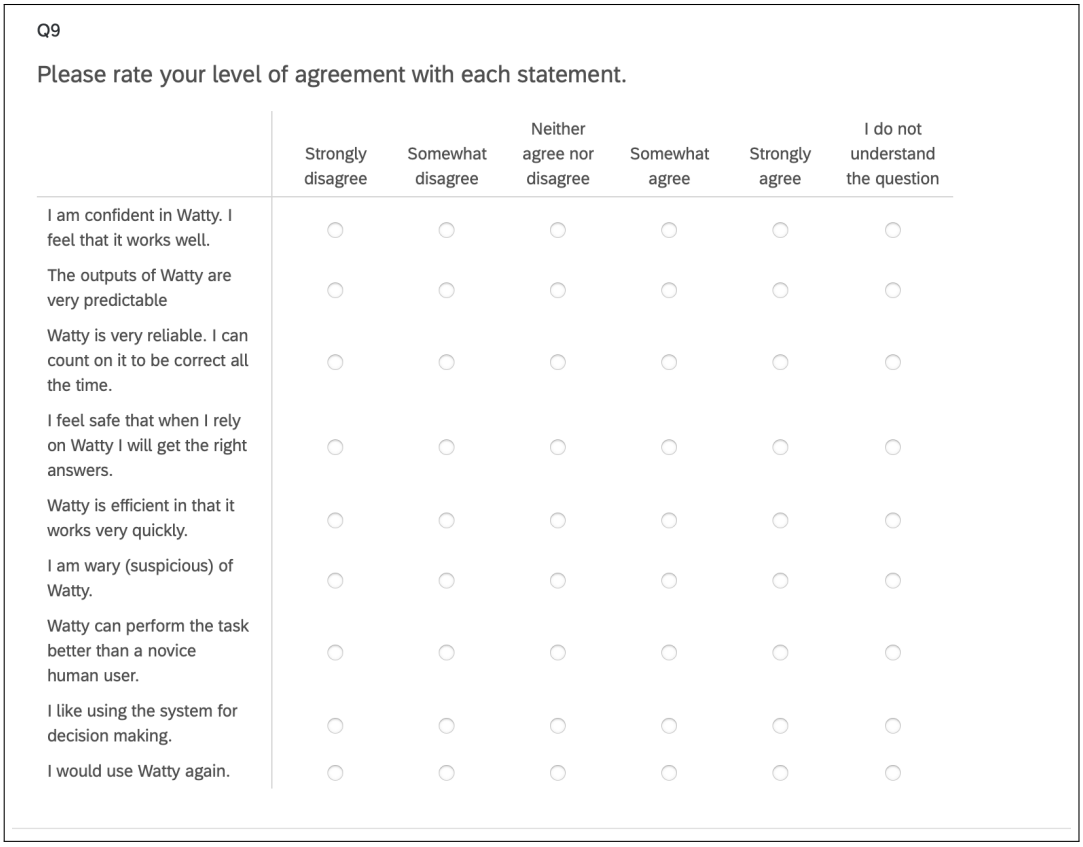


Q8

Please rate your level of agreement with each statement.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	I do not understand the question
I would find Artificial Intelligence (AI) and related technologies useful to use in daily life.	○	○	○	○	○	○
In general, I am positive about the use of AI and related technologies.	○	○	○	○	○	○
I am reluctant for the use of AI and related technologies.	○	○	○	○	○	○
I feel somewhat intimidated to start using AI and related technologies.	○	○	○	○	○	○
I am in favor of using services/systems that use AI and related technologies.	○	○	○	○	○	○
I am confident I could become skillful at services/systems adopting AI and related technologies.	○	○	○	○	○	○
I would find systems/services adopting AI and related technologies easy to use.	○	○	○	○	○	○

Figuur A.4: Evaluatievraag voor het initieel vertrouwen van de gebruiker, naar Andrews et al. [5]

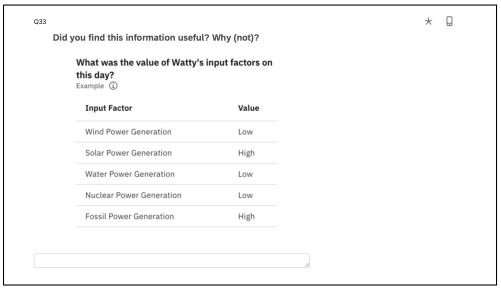


Q9

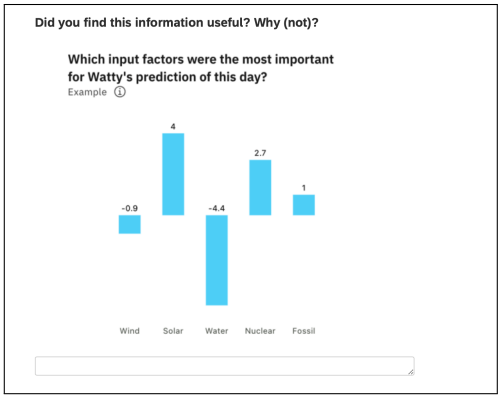
Please rate your level of agreement with each statement.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	I do not understand the question
I am confident in Watty. I feel that it works well.	○	○	○	○	○	○
The outputs of Watty are very predictable	○	○	○	○	○	○
Watty is very reliable. I can count on it to be correct all the time.	○	○	○	○	○	○
I feel safe that when I rely on Watty I will get the right answers.	○	○	○	○	○	○
Watty is efficient in that it works very quickly.	○	○	○	○	○	○
I am wary (suspicious) of Watty.	○	○	○	○	○	○
Watty can perform the task better than a novice human user.	○	○	○	○	○	○
I like using the system for decision making.	○	○	○	○	○	○
I would use Watty again.	○	○	○	○	○	○

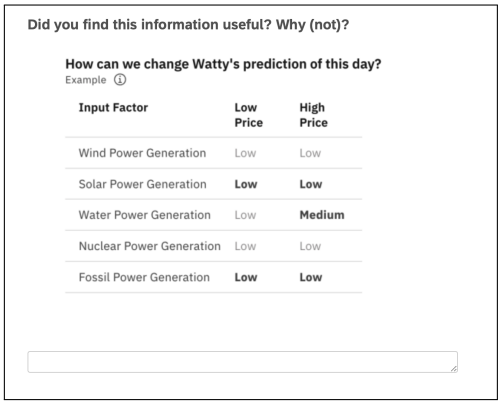
Figuur A.5: Evaluatievraag voor vertrouwen van de gebruiker in het systeem, naar Hoffman et al. [28].



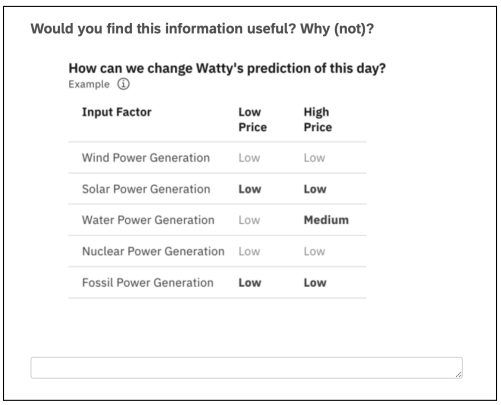
Figuur A.6: Open vraag over inputfactoren.



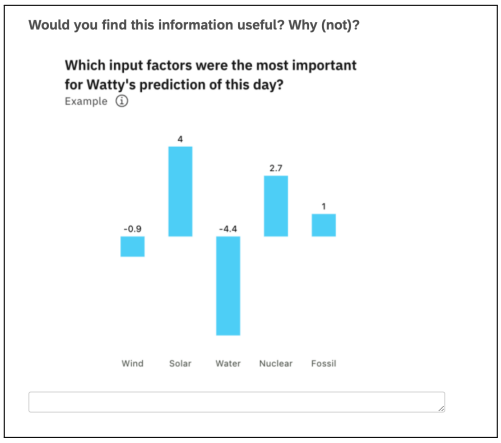
Figuur A.7: Open vraag over SHAP.



Figuur A.8: Open vraag over CF.



Figuur A.9: Open vraag over SHAP.



Figuur A.10: Open vraag over CF.

Figuur A.11: Open vraag voor suggesties en opmerkingen.

Bijlage B

Think-aloud 1

Probleem	Freq.	Oplossing	Categorie
Betekenis van inputfactoren onduidelijk	3	Kolomtitel toevoegen aan tooltip; korte samenvatting van betekenis van inputfactor toevoegen (bv. Wind = hoeveelheid opgewekte windenergie op die dag); veranderen van "laagïn "lage vraag"	**Explanation**
Verschil tussen inputfactoren en SHAP onduidelijk	2	Kolomtitel en meer gedetailleerde uitleg over SHAP toevoegen	
Uitlegscherm niet grondig gelezen	2	Belangrijke aspecten van de uitleg benadrukken	
Onzekerheidsinterval onduidelijk	1	Betere tekstuele uitleg van onzekerheidsinterval	
kWh moeilijk te interpreteren	1	Andere metriek toevoegen (bv. Hoeveel ladingen van een wasmachine,...)	
CF: raar om te kiezen voor een hoge prijs	1	Beide opties tonen zonder interactie	
Niet voorbereid op Model Output Prediction	1	De gebruiker vooraf op de hoogte brengen van de verwachtingen (bv. In een overzichtsscherm of een instructiescherm)	**Informatie**
Niet duidelijk waar data uit verleden stopt en voorspelling begint	2	Verander kleuren en voeg "huidige dag"label toe	**Stijl**
"Completed"knop is verwarrend	1	Vervangen door "Ik ben klaaröf inactief maken tot hoofd UI wordt getoond	
"About your knowledge survey" geen optie ïk ken het niet" "Would you like... using for decision making".	1	Optie toevoegen	**Vragenlijsten**
Niet duidelijk welke beslissingen gemaakt kunnen worden	1	Vraag verwijderen	
Model Output Prediction moeilijk met enkel inputfactoren	1	Toon ook tooltip voor voorspelling op MOP-scherm; laat gebruiker beslissen hoe zeker hij/zij is van eigen voorspelling	
Niet voldoende tijd besteed in de explanationinterface	2	Voeg specifieke taken toe; bij voltooiing: toon popup "ben je zeker dat je genoeg verkend hebt?"	**Timing**

Tabel B.1: Feedback verzameld in Think-Aloud 1

Bijlage C

Think-aloud 2

Probleem	Freq.	Oplossing	Categorie
Periode van voorspelling niet afgestemd op periode van te kiezen contract	3	Periodes gelijkstellen	**Explanation**
Niet duidelijk waar data uit verleden stopt en voorspelling begint	3	Voeg scheidingslijn en label toe	
Niet voldoende data in lijngrafiek	2	Meer data geven	
Extra uitleg nodig bij explanations	2	Info-popups toevoegen	
Onduidelijk wat eenheid is op SHAP grafiek	1	Geen oplossing (deel van SHAP explanation)	
Geen trendlijn op prijsgrafiek	1	Geen oplossing (geen deel van explanation)	
Onduidelijk of SHAP gaat over specifieke voorspelling of algemene data	1	Info-popups toevoegen	
Data en explanations komen niet altijd overeen met verwachtingen	1	Geen oplossing (limitatie van data & model)	**Data**
Trainingdata wordt niet gegeven	1	Geen oplossing (geen deel van explanation)	
Niet alle labels van SHAP grafiek zijn zichtbaar	1	Labels herwerken	**Stijl**
Onduidelijk wat woorden in het vet betekenen bij CF	1	Info-popups toevoegen	
Tabel van CF visueel niet aanlokkelijk vergeleken met SHAP grafiek	1	Kleur toevoegen aan tabel	
Model Output Prediction moeilijk	1	Ander evaluatiecriteria gebruiken	**Vragenlijsten**

Tabel C.1: Feedback verzameld in Think-Aloud 2

Bibliografie

[1] *Notes on the N-Person Game — II: The Value of an N-Person Game*. RAND Corporation.

[2] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). 6:52138–52160. Conference Name: IEEE Access.

[3] D. Alvarez-Melis and T. S. Jaakkola. Towards robust interpretability with self-explaining neural networks.

[4] N. Amjady and M. Hemmati. Energy price forecasting - problems and proposals for such predictions. 4(2):20–29.

[5] J. E. Andrews, H. Ward, and J. Yoon. UTAUT as a model for understanding intention to adopt AI and related technologies among librarians. 47(6):102437.

[6] D. W. Apley and J. Zhu. Visualizing the effects of predictor variables in black box supervised learning models.

[7] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Mueller. How to explain individual classification decisions.

[8] R. Binns. Fairness in machine learning: Lessons from political philosophy.

[9] D. Bissing, M. T. Klein, R. A. Chinnathambi, D. F. Selvaraj, and P. Ranganathan. A hybrid regression model for day-ahead energy price forecasting. 7:36833–36842.

[10] F. Bre, J. M. Gimenez, and V. D. Fachinotti. Prediction of wind pressure coefficients on building surfaces using artificial neural networks. 158:1429–1441.

[11] R. M. J. Byrne. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 6276–6282. International Joint Conferences on Artificial Intelligence Organization.

[12] R. M. J. Byrne. Précis of *The Rational Imagination: How People Create Alternatives to Reality*. 30(5):439–453.

[13] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, Apr. 2017.

[14] M. Campbell, A. Hoane, and F.-h. Hsu. Deep blue. 134(1):57–83.

[15] P. Chaudhury, A. Tyagi, and P. K. Shanmugam. Comparison of various machine learning algorithms for predicting energy price in open electricity market. In *2020 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE)*, pages 1–7. IEEE.

[16] M. Chromik and M. Schuessler. A taxonomy for human subject evaluation of black-box explanations in xai. In *ExSS-ATEC@IUI*, 2020.

[17] Council of European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 6. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[18] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, Mar. 2017. arXiv: 1702.08608 version: 2.

[19] C. Fernández-Loría, F. Provost, and X. Han. Explaining data-driven decisions made by AI systems: The counterfactual approach. Number: arXiv:2001.07417.

[20] X. Ferrer, T. v. Nuenen, J. M. Such, M. Cote, and N. Criado. Bias and discrimination in AI: A cross-disciplinary perspective. 40(2):72–80.

[21] J. H. Friedman. Greedy function approximation: A gradient boosting machine. 29(5).

[22] J. Fürnkranz, T. Kliegr, and H. Paulheim. On cognitive preferences and the plausibility of rule-based models. 109(4):853–898.

[23] F. Gedikli, D. Jannach, and M. Ge. How should i explain? a comparison of different explanation types for recommender systems. 72(4):367–382.

[24] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE.

[25] A. Gosiewska and P. Biecek. Do not trust additive explanations. Number: arXiv:1903.11420.

[26] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):93:1–93:42, Aug. 2018.

[27] J. Hatwell, M. M. Gaber, and R. M. A. Azad. gbt-HIPS: Explaining the classifications of gradient boosted tree ensembles. 11(6):2511.

[28] R. Hoffman, S. Mueller, G. Klein, and J. Litman. *Metrics for Explainable AI: Challenges and Prospects*. Dec. 2018.

[29] D. Holliday, S. Wilson, and S. Stumpf. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 164–168. ACM.

[30] A. Hudon, T. Demazure, A. Karran, P.-M. Léger, and S. Sénécal. Explainable artificial intelligence (XAI): How the visualization of AI predictions affects user cognitive load and confidence. In F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, A. B. Randolph, and G. Müller-Putz, editors, *Information Systems and Neuroscience*, volume 52, pages 237–246. Springer International Publishing. Series Title: Lecture Notes in Information Systems and Organisation.

[31] P. Humphreys. The philosophical novelty of computer simulation methods. *Synthese*, 169(3):615–626, Aug. 2009.

[32] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14. ACM.

[33] M. T. Keane, E. M. Kenny, E. Delaney, and B. Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques.

[34] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697. ACM.

[35] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness.

[36] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, and F. Doshi-Velez. An evaluation of the human-interpretability of explanation.

[37] J. Larson, L. Kirchner, S. Mattu, and J. Angwin. Machine Bias.

[38] T. Laugel, M.-J. Lesot, C. Marsala, and M. Detyniecki. Issues with post-hoc counterfactual explanations: a discussion. Number: arXiv:1906.04774.

[39] Z. C. Lipton. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, June 2018.

[40] O. Loyola-Gonzalez. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. 7:154096–154113.

[41] H. Lu, X. Ma, M. Ma, and S. Zhu. Energy price prediction using data-driven models: A decade review. 39:100356.

[42] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions.

[43] M. G. M. Mehedi Hasan and D. Talbert. Mitigating the rashomon effect in counterfactual explanation: A game-theoretic approach. 35.

[44] T. Miller. Explanation in artificial intelligence: Insights from the social sciences.

[45] T. Miller, P. Howe, and L. Sonenberg. Explainable AI: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences.

[46] S. Mohseni, F. Yang, S. K. Pentyala, M. Du, Y. Liu, N. Lupfer, X. Hu, S. Ji, and E. D. Ragan. Trust evolution over time in explainable ai for fake news detection. 2020.

[47] S. Mohseni, N. Zarei, and E. D. Ragan. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv:1811.11839 [cs]*, Aug. 2020. arXiv: 1811.11839.

[48] C. Molnar. *Interpretable Machine Learning*.

[49] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

[50] T. D. Mount, Y. Ning, and X. Cai. Predicting price spikes in electricity markets using a regime-switching model with time-varying parameters. 28(1):62–80.

[51] M. A. Nielsen. Neural Networks and Deep Learning. 2015.

[52] I. Nunes and D. Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. 27(3):393–444.

[53] J. Ooge and K. Verbert. Trust in prediction models: a mixed-methods pilot study on the impact of domain expertise. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pages 8–13. IEEE.

[54] N. Pindoriya, S. Singh, and S. Singh. An adaptive wavelet neural network-based energy price forecasting in electricity markets. 23(3):1423–1432.

[55] E. Puiutta and E. M. Veith. Explainable reinforcement learning: A survey.

[56] G. Ras, M. van Gerven, and P. Haselager. Explanation methods in deep learning: Users, values, concerns and challenges.

[57] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning.

[58] M. T. Ribeiro, S. Singh, and C. Guestrin. "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*, KDD '16, pages 1135–1144, New York, NY, USA, Aug. 2016. Association for Computing Machinery.

[59] M. Riveiro and S. Thill. 'that's (not) the output i expected!' on the role of end user expectations in creating explanations of AI systems. 298:103507.

[60] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez. Explainable artificial intelligence (XAI) on TimeSeries data: A survey.

[61] J. Schneider and J. Handali. Personalized explanation in machine learning: A conceptualization.

[62] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: from theory to algorithms*. Cambridge University Press.

[63] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. 529(7587):484–489.

[64] B. F. Skinner. *Science And Human Behavior*. Simon and Schuster, Mar. 1965. Google-Books-ID: Pjjknd1HREIC.

[65] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186. ACM.

[66] M. Sundararajan and A. Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[67] Venkatesh, Morris, Davis, and Davis. User acceptance of information technology: Toward a unified view. 27(3):425.

[68] S. Verma, J. Dickerson, and K. Hines. Counterfactual explanations for machine learning: A review.

[69] S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. 7(2):76–99.

[70] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Number: arXiv:1711.00399.

[71] G. Warren, M. T. Keane, and R. M. J. Byrne. Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI.

[72] H. J. P. Weerts, W. van Ipenburg, and M. Pechenizkiy. A human-grounded evaluation of SHAP for alert processing.

[73] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM.

[74] Y. Zhang, P. Tino, A. Leonardis, and K. Tang. A survey on neural network interpretability. 5(5):726–742.