Explaining Artificial Intelligence With Tailored Interactive Visualisations

Jeroen Ooge

KU LEUVEN

Explaining Artificial Intelligence With Tailored Interactive Visualisations

Jeroen OOGE

Examination committee:
Prof. em. dr. Bob Puers, chair
Prof. dr. Katrien Verbert, supervisor
Prof. dr. ir. Tinne De Laet
Prof. dr. Vero Vanden Abeele
Prof. dr. Tias Guns
Prof. dr. Denis Parra
(PUC, Chile)

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

October 2023

Preface

This thesis is the result of four years of hard work. To complement the 6 chapters that present my research in a traditional academic format, I also wanted to "humanise" my thesis in a less conventional way. First, Chapters 4 to 8 conclude with anecdotes and photographs that give a glimpse of what happened while I was working on them. Second, the background section in Chapter 2 is written in a non-academic style, hopefully making it easier for people outside my research niche to understand the overall research story.

 For the Einsteins among you, my brain friend will chime in with more technical details from time to time. (Credits to flaticon.com.)

Humanising This Thesis With Glimpses Behind the Scenes

Chapter 2 will explain that AI is often a black box, producing outcomes in an unclear way. In a sense, PhD trajectories are similar: there is typically little context about how PhD researchers spend years working on their thesis. Opening the "PhD black box" does more justice to at least three parties involved:

1. A PhD is extremely demanding for *PhD researchers and their supporters*. Yet, academic texts do not reflect the painstaking hours, sacrifices, adventures, and joyful moments needed to produce them. Theses remain scientific texts, but they shouldn't look as if produced by machines.
2. Sharing experiences and life stories can help *fellow researchers* better process and battle against challenges that put their private lives and mental health under pressure. PhD trajectories generate knowledge beyond research outcomes.
3. Countless times, I have had to explain *people outside academia* what a PhD entails, facing many misconceptions such as "studying for free and

getting long holidays." Academics should better communicate what a PhD encompasses, acknowledging both the tough and enriching parts.

To show a bit of what happened during my PhD, I complemented each chapter with four anecdotes and photographs I took. Furthermore, I acknowledged the many people who supported me throughout.

Humanising This Thesis With A More Accessible Text

Academics sometimes joke that "apart from the examination committee, nobody really reads theses." While this might consolidate PhD researchers who stress about their thesis, it should lead to more reflection: why do people seldom read theses? One of the reasons is that academic writing styles, jargon, and unengaging formats often obscure the fascinating content. More accessible theses can benefit both fellow researchers and non-researchers:

1. *Researchers* are often very specialised, so, in the bigger picture, they are rather unfamiliar with each other's topics. Yet, many researchers want to learn from other fields, which isn't easy when dense papers are scattered over different journals, conference proceedings, and books. Accessible theses are ideal to quickly get a taste of new research areas. This fosters interdisciplinary collaboration and benefits science as a whole.
2. *Non-researchers* have different opinions about inaccessible scientific texts, ranging from "They are Greek to me, so researchers must be really smart" to "Typical for researchers in their ivory tower." Often, there is little engagement with the research itself. This is a failure: the purpose of theses isn't to stroke researchers' egos or reinforce misconceptions of academia, but to share knowledge, spark enthusiasm, and generate discussions. Non-researchers can also be part of this discourse.

To make my thesis more accessible, I wrote the background section in a more fluent style than typical academic writing, included illuminating illustrations, and restricted jargon. This doesn't necessarily make my thesis some light reading, but I hope it lets you understand the overall story and sparks your interest in my research area (and maybe even research overall).

My attempts to add a human touch to my thesis are just an example of how to address the issues I raised. Nevertheless, I hope they inspire you to push further, whether you are a researcher or not. Quickly start reading the next chapter now. Once you get to know my research topic better, you might find my call to "humanise" theses and "open black boxes" quite ironic.

Abstract

The rise of "big data" and artificial intelligence (AI) in countless application domains comes with tremendous opportunities, but also entails challenges concerning transparency and controllability. Well-performing AI models are often "black boxes," which means that understanding how they establish outcomes is hard or even infeasible. Researchers in *explainable AI* (XAI) therefore develop algorithm-centred and human-centred methods that try to give people insights into the reasoning process of AI models. In turn, the expectation is this allows people to better understand and trust AI models, and thus make better-informed decisions. However, the body of experimental human-centred research that backs up these expectations is limited. In addition, it is unclear whether XAI techniques meet the insights required by different user groups across application domains and contexts in the first place. Thus, XAI studies with actual people and real-world data are urgent.

Our work focuses on designing, implementing, and evaluating visualisation-supported explanations for AI systems in healthcare, agrifood, and education. Following human-centred research practices, we study three research questions: (1) How can visual explanations tailored to a target audience and application domain make AI models more transparent?; (2) How can people control AI models with additional feedback, supported by interactive visual explanations?; and (3) How do visual explanations and control affect people's perceptions of AI systems, e.g., in terms of appropriate trust and understanding their outcomes?

Overall, we show how explainability can be established through visual analytics, visualisation-supported justification, and visualisation-supported control. We do this by reviewing the existing literature, developing new visual explanations and control mechanisms in close collaboration with real end-users of AI systems, and conducting user studies to better understand how our explainability methods affect people's perceptions of AI systems. Our work demonstrates the value of human-centred and interdisciplinary research to design XAI solutions that align with people's needs and truly augment human capabilities with AI.

Beknopte samenvatting

De opkomst van “big data” en artificiële intelligentie (AI) in talloze toepassingsdomeinen brengt enorme kansen met zich mee, maar leidt ook tot uitdagingen omtrent transparantie en controleerbaarheid. Goed presterende AI-modellen zijn vaak “zwarte dozen”, wat betekent dat het moeilijk of zelfs onmogelijk is om te begrijpen hoe ze tot resultaten komen. Onderzoekers binnen *verklaarbare AI* (in het Engels: *explainable AI*, ofwel XAI) ontwikkelen daarom algoritme- en mensgerichte methodes die mensen inzicht proberen geven in het redeneerproces van AI-modellen. De verwachting is dat mensen AI-modellen daardoor beter kunnen begrijpen en vertrouwen, en dus beter geïnformeerde beslissingen kunnen nemen. Het experimentele mensgerichte onderzoek dat die verwachtingen ondersteunt, is echter beperkt. Bovendien is het onduidelijk of XAI-technieken überhaupt de inzichten bieden die verschillende gebruikersgroepen in verschillende toepassingsdomeinen en contexten nodig hebben. Er is dus dringend nood aan XAI-onderzoek met echte mensen en gegevens uit de echte wereld.

Ons werk focust op het ontwerpen, implementeren en evalueren van visualisatie-ondersteunde verklaringen voor AI-systemen in de gezondheidszorg, de agro-industrie en het onderwijs. We bestuderen drie onderzoeksvragen op basis van mensgerichte onderzoekspraktijken: (1) Hoe kunnen AI-modellen transparanter worden gemaakt door visuele verklaringen die zijn afgestemd op de doelgroep en het toepassingsdomein? (2) Hoe kunnen mensen AI-modellen controleren met aanvullende feedback, ondersteund door interactieve visuele verklaringen? en (3) Hoe beïnvloeden visuele verklaringen en controle mensen in hun perceptie van AI-systemen, bijvoorbeeld in termen van gepast vertrouwen en begrip van de output?

Samengevat: we laten zien hoe verklaarbaarheid van AI tot stand kan komen via visuele analyse, visualisatie-ondersteunde rechtvaardiging en visualisatie-ondersteunde controle. We doen dit door de bestaande literatuur te bestuderen, nieuwe visuele verklaringen en controlemechanismes te ontwikkelen in nauwe

samenwerking met echte eindgebruikers van AI-systemen, en gebruikersstudies uit te voeren om beter te begrijpen hoe onze verklaringsmethodes de perceptie van mensen over AI-systemen beïnvloeden. Ons werk demonstreert de waarde van mensgericht en interdisciplinair onderzoek om XAI-oplossingen te ontwerpen die aansluiten bij de behoeften van mensen en die menselijke capaciteiten daadwerkelijk versterken met AI.

Contents

Abstract iii

Beknopte samenvatting v

Contents vii

1 Introduction 1

2 Background and Related Work 3
2.1 What Is AI? . . . 3
2.2 Why We Need Explainable AI . . . 6
2.3 XAI, It's Complicated . . . 12
2.4 Algorithmic XAI approaches . . . 14
2.5 Human-Centred XAI Approaches . . . 19
2.6 Visualisation for XAI . . . 25
2.7 Control Mechanisms for XAI . . . 29
2.8 How XAI Can Be Evaluated . . . 31

3 Thesis Overview 33
3.1 Open Research Challenges . . . 33
3.2 Research Goals and Research Questions . . . 37
3.3 Overall Methods . . . 38
3.4 Organisation of the Text . . . 39

I Explainability Through Visual Analytics 41

4 Explaining AI with Visual Analytics in Healthcare 45
4.1 Introduction . . . 45
4.2 Background and Related Work . . . 47
4.2.1 Explainable Artificial Intelligence . . . 47

4.2.2 Visual Analytics for Explainable Artificial Intelligence . 48
4.2.3 Visual Analytics in Healthcare 48
4.3 Paper Collection and Classification Process 50
4.4 Visualising Algorithmic Outcomes in Visual Analytics 50
4.5 Interaction in Visual Analytics 55
4.6 Shepherding Algorithms With Visual Analytics 59
4.7 Directly Explaining Algorithms With Visual Analytics 62
4.8 Observations, Opportunities and Challenges 63
4.8.1 Visualising Outcomes: Many Algorithm-Dependent Possibilities 64
4.8.2 Interacting With Visualisations: Sufficient or Too Much? 65
4.8.3 Shepherding Algorithms: A Higher-Order Interaction . . 65
4.8.4 Direct Explanations: Rare Yet Promising 66
4.9 Conclusion 66

5 Visually Explaining Uncertain Price Predictions in Agrifood 79
5.1 Introduction 79
5.2 Background and Related Work 81
5.2.1 Visualisation for Decision Support Systems 82
5.2.2 Uncertainty Visualisation 82
5.2.3 Visualisation for Explainable Artificial Intelligence . . . 83
5.2.4 Trust in Intelligent Systems 84
5.3 Materials and Methods 84
5.3.1 Visual Decision Support System 84
5.3.2 Study Rationale 85
5.3.3 Study Design 86
5.3.4 Measurement Instruments and Qualitative Analysis . . 88
5.4 Results 90
5.4.1 Usability 90
5.4.2 Usefulness and Needs 94
5.4.3 Model Understanding 99
5.4.4 Trust 101
5.5 Discussion 106
5.5.1 A User-Friendly and Useful Visual DSS 107
5.5.2 Tailoring, Tailoring, Tailoring: Different End Users, Different Needs 107
5.5.3 Gradual Model Understanding through Visual Analysis 108
5.5.4 Trust Is Multi-Faceted and Evolves 108
5.5.5 Fostering Appropriate Trust Through Usefulness and Meeting Needs 109
5.5.6 Taking a Step Back: Increasing Uptake of DSSs in Agrifood with User-Centred Approaches 110
5.5.7 Limitations and Transferability 111

5.6 Conclusions . . . 112

II Explainability Through Visualisation-Supported Justification 123

6 Explaining Recommendations in E-Learning 127
6.1 Introduction . . . 127
6.2 Background and Related Work . . . 129
6.2.1 Explainable Artificial Intelligence . . . 129
6.2.2 Explaining Recommendations . . . 130
6.2.3 Trust in Automated Systems . . . 131
6.2.4 Trust in Explained Recommendations . . . 132
6.2.5 Underexplored Research Areas . . . 132
6.3 Materials and Methods . . . 133
6.3.1 E-learning Platform with an Exercise Recommender . . . 133
6.3.2 Explanations for Recommendations . . . 133
6.3.3 Participant Recruitment . . . 134
6.3.4 Study Design . . . 136
6.3.5 Statistical Analysis . . . 137
6.4 Results . . . 137
6.4.1 Effects of Real Explanations . . . 138
6.4.2 Effects of Placebo Explanations . . . 139
6.4.3 Effects of No Explanations . . . 141
6.4.4 Correlations . . . 141
6.4.5 Recommendation Clicks . . . 141
6.5 Discussion . . . 143
6.5.1 Explanations Increase Multidimensional Initial Trust... 143
6.5.2 ...But Not One-Dimensional Initial Trust . . . 144
6.5.3 Placebo Explanations Are a Useful Baseline . . . 144
6.5.4 Tailoring Explanations Remains Important . . . 145
6.5.5 Taking a Step Back: Recommendations and Explanations in E-Learning . . . 146
6.5.6 Limitations and Future Work . . . 147
6.6 Conclusion . . . 147

III Explainability Through Visualisation-Supported Control 161

7 Steering Recommendations and Visualising Its Impact 165
7.1 Introduction . . . 165
7.2 Background and Related Work . . . 167

7.2.1 Control over Recommendations . . . 167
7.2.2 Explainable AI and Trust . . . 168
7.2.3 Educational Recommender Systems . . . 169
7.2.4 Estimating Mastery and Exercise Difficulty . . . 169
7.3 Materials and Methods . . . 170
7.3.1 E-Learning Platform with Personalised Exercises and a Control Mechanism . . . 171
7.3.2 Study Design . . . 174
7.3.3 Participant Recruitment . . . 176
7.3.4 Data Analysis . . . 176
7.4 Results . . . 176
7.4.1 Effects Without Control or Seeing Its Impact . . . 177
7.4.2 Effects of Controlling Recommendations . . . 178
7.4.3 Effects of Visualising the Impact of Control . . . 181
7.4.4 Correlations . . . 182
7.4.5 Elo Ratings . . . 182
7.5 Discussion . . . 182
7.5.1 Sanity Check for Responses About Control . . . 184
7.5.2 Control Does Not Affect Trust but Stimulates Self-Reflection 185
7.5.3 Seeing the Impact of Control Grows Trust . . . 186
7.5.4 Visualising the Impact of Control is a Kind of Explanation 186
7.5.5 Implications for Explainable AI Research . . . 187
7.5.6 Taking a Step Back: Technology-Enhanced Learning and Control . . . 187
7.5.7 Limitations and Future Work . . . 188
7.6 Conclusion . . . 189

8 Steer, See Impact, Solve 201
8.1 Introduction . . . 202
8.2 Background and Related Work . . . 203
8.2.1 Visual Explanations for AI Models . . . 204
8.2.2 Control Over AI Models . . . 204
8.2.3 Metacognition, Motivation, and Trust . . . 205
8.3 Methods and Materials . . . 206
8.3.1 E-Learning Platform With Learner Control . . . 206
8.3.2 User Studies to Inform Design . . . 209
8.3.3 In-Class Experiment . . . 210
8.4 Design Process . . . 212
8.4.1 Prototype 1 and Informal Feedback . . . 212
8.4.2 Prototype 2 and Think-Aloud Studies . . . 212
8.4.3 Prototype 2 and Focus Groups . . . 215
8.4.4 Prototype 3 and Think-Aloud Studies . . . 219
8.4.5 Final Prototype . . . 222

8.5 Results of In-Class Experiment . . . 222
8.5.1 Validation of Measurements . . . 223
8.5.2 Testing Hypotheses About Perceptions . . . 224
8.5.3 Testing Hypotheses About Interactions . . . 224
8.5.4 Correlation Analysis . . . 226
8.6 Discussion . . . 227
8.6.1 *Why* Explanations not for Adolescents, but for Teachers? 227
8.6.2 Fostering Motivation With *What-If* Explanations and Wise Feedback . . . 229
8.6.3 Learner Control Is Not a Panacea . . . 230
8.6.4 Limitations and Future Work . . . 230
8.7 Conclusion . . . 231

IV Conclusions 243

9 Research Contributions and Future Directions 245
9.1 Research Contributions . . . 245
9.2 Impact . . . 252
9.3 Critical Reflections and Future Directions . . . 253
9.4 Taking a Final Step Back . . . 257

A Questionnaires and Details 261
A.1 Pre- and Post-Study Questionnaires . . . 261
A.2 Elo Rating System . . . 266
A.3 Wise Feedback . . . 267
A.4 Learning Performance Correction . . . 268

Chapter 1

Introduction

We all know it: we live in the era of "big data" and artificial intelligence (AI). The rise of AI is visible in countless application domains. For example, healthcare applies AI to predict the onset of diseases, analyse medical imaging, or help rehabilitate patients with acute and chronic conditions; agrifood uses AI to precisely monitor crop growth, optimise irrigation, or support smart farming; and education adopts AI to recommend learning materials, create new educational content, or automatically assess learners' mastery level. The list goes on and keeps growing daily.

Amidst the AI hype, however, it is often overlooked that AI models do not always behave as expected and that for some models it is even impossible to explain in a human-understandable way how they obtain their outcomes. This can be harmful in situations where people are using AI to make important decisions. Therefore, we need techniques to understand how AI models "reason," how they "behave" in different contexts, and how people can steer them with domain knowledge. Researchers in the field of **explainable AI** (XAI) are developing such techniques. This is a hard yet exciting multidisciplinary challenge, because besides algorithmic solutions, XAI needs to consider what *people* need. In the end, it is namely people who use AI to augment their skills, and who need to be able to rely on it.

Thus, this thesis is about AI, explanations, and people. Essentially, we will study how outcomes of AI models can be explained to people while tailoring different explainability solutions towards people's needs, their experience with AI, and the context in which they use AI. In particular, we will harvest the power of data visualisation and study how explanations supported by visualisations affect people's perceptions of AI systems, for example, their trust in those systems

and how well they understand them. In addition, we will study how people can control AI models while being supported by visual explanations.

In total, this thesis presents 5 elaborate studies about XAI for adolescents and adults, not coincidentally in the domains mentioned before: healthcare, agrifood, and education. These studies investigate three main approaches for explaining AI models: visual analytics, visualisation-supported justification, and visualisation-supported control. The following chapters will clarify these terms. Overall, the work presented in this thesis starts to disentangle the intricate ways in which people calibrate their trust in AI systems, how visual explanations can or cannot meet people's actual needs, and how people interact with control mechanisms for human-AI collaboration. Hopefully, this thesis inspires you and many others to reflect more upon how XAI can or cannot be used to design more trustworthy and controllable AI systems, and more generally, upon the human side of AI.

Chapter 2

Background and Related Work

This chapter is a kind of prequel for the new research in this thesis: it introduces important concepts and existing work. It will first cover some background information about AI itself (Section 2.1) to prepare the motivation for why we need AI to be *explainable* (Section 2.2). Section 2.3 will then discuss why explainable AI is challenging: the problem is interdisciplinary and needs both algorithmic and human-centred approaches (Sections 2.4 and 2.5). Once you grasp what explainable AI is and how complex it is, you might wonder how researchers evaluate explanations. Section 2.8 will answer your questions. The final sections will introduce two approaches that lie at the heart of the research in the next chapters: visualisation (Section 2.6) and control mechanisms (Section 2.7) to facilitate explainability. Prepare for an enlightening start with coffee machines, cute cats and chicken chicks, and risky pyramids!

2.1 What Is AI?

First things first. If we're going to talk about artificial intelligence, we need to agree on what that is. And I mean what is currently possible, not the often dystopian technologies you typically see in science-fiction movies. This section will give you a high-level taste of how AI algorithms work nowadays. We will restrict ourselves to some basic concepts since an introduction to AI can be a book on its own. In fact, there are many accessible examples already, for example (Buijsman, 2020; Domingos, 2015; Mitchell, 2019). For now, I will only present the aspects that are relevant to understand the rest of the story and the motivation for this thesis.

To start off, let me immediately stress that *artificial intelligence* is a kind of buzzword. The word "intelligence" misleads many people into expecting that we're dealing with something similar to human intelligence. That is not the case at all. Essentially, AI is mostly mathematics and finding patterns in data. Furthermore, AI models are specialised in narrow tasks, such as deciding whether an image is a cat or a dog, translating sentences, converting speech to text, playing chess or go, predicting the next numerical values in a time series, and so on. Besides those tasks, they can do literally nothing and they cannot generalise their "knowledge." For example, AI models might excel at distinguishing cats from dogs, but as such do not "learn" anything about mammals, different breeds, the concept of having four legs, the emotional value humans attach to their pets, or distinguishing tigers from wolves. Thus, AI models are the "ultimate idiot savants" (Mitchell, 2019, p. 217). This is completely different from how we as humans reason and learn.

So how does AI work? And what is the difference between an AI algorithm and an AI model? Figure 2.1 shows a high-level representation of the AI lifecycle. First, real world data or knowledge is processed according to some recipe, which is the **AI algorithm**. This recipe then results in an **AI model**, a piece of software that transforms given input into output. You could compare AI models to coffee machines: when you put something in (coffee beans and water), they spit something out (coffee – if you're lucky). Following our metaphor, AI algorithms are like the manufacturing process of these coffee machines. Finally, people use AI models to obtain insights or make informed decisions about something. For example, if an AI model is built to classify photos as cats or dogs, you could give it a photo and it would output 'cat' or 'dog' (also when the photo depicts something completely different, say, a lamp). Obtained insights may lead to new data or knowledge, which can be used to create new AI models.

Some types of AI models, unlike real coffee machines, can create new "knowledge" themselves and thus "learn." For example, the AlphaGo algorithm lets different models play Go against each other to gain "knowledge" about which moves lead to victory (Fu, 2016). This is called *reinforcement learning* (Russell and Norvig, 2021). Bare in mind, though, that this "learning" is still different from human learning and only possible because people actively guide and monitor it.

The AI algorithm is of course the key link in the above cycle. It is the place where researchers apply clever logical and mathematical techniques to build AI models specialised in a specific task. Broadly speaking, there are two main streams in AI: symbolic and subsymbolic AI.

Symbolic AI algorithms combine and process small chunks of human

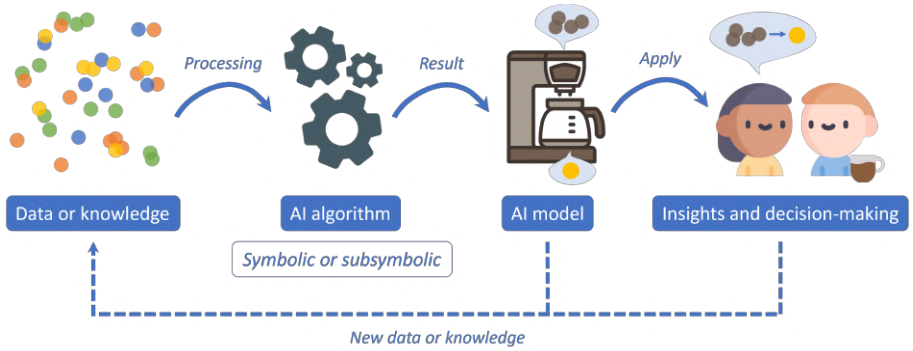


Figure 2.1: Abstract representation of the AI lifecycle. Real-world data or knowledge is processed by a symbolic or subsymbolic AI algorithm to create an AI model, which in turn can be used by people to transform a given input into an output. (Credits: gear, coffee machine, mug, and people by flaticon.com.)

knowledge with logical rules and probabilistic reasoning. A famous example is MYCIN (Shortliffe, 1977), a so-called *expert system* from the 1970s that helped physicians diagnose and treat infections with hundreds of rules based on knowledge collected from physicians. An advantage of symbolic AI models is that they can explain their reasoning process by keeping track of which rules they follow. The downside, however, is that they do not scale up to large or difficult problems (Russell and Norvig, 2021).

Subsymbolic AI "learns" from examples. A subsymbolic algorithm needs tons of human-labelled data, say, photos with a label 'cat' or 'dog', and then uses this *training data* to build a model. This is done iteratively: for each example in the training data, the model outputs a label and compares it to the true label. Then, it modifies its parameters to decrease the difference, so it becomes more likely to perform well on future examples (Russell and Norvig, 2021). Famous subsymbolic algorithms are **neural networks**, for which Figure 2.2 shows a toy example. Roughly, the neural network converts an input image into numbers and feeds them into the input layer, the hidden layers do many computations, and finally, two numbers come out of the output layer. The input image is classified as the label with the highest number; in this case 'cat'. We skip the mathematical details of the computations here, but what's important is that all connections are assigned a weight, that is, a number. These are the parameters that the algorithm updates iteratively. In other words, neural networks try to find weights such that they make the least mistakes for the labelled training data. To conclude, it turns out that neural networks are currently the most performant AI approach for many tasks. One advantage is

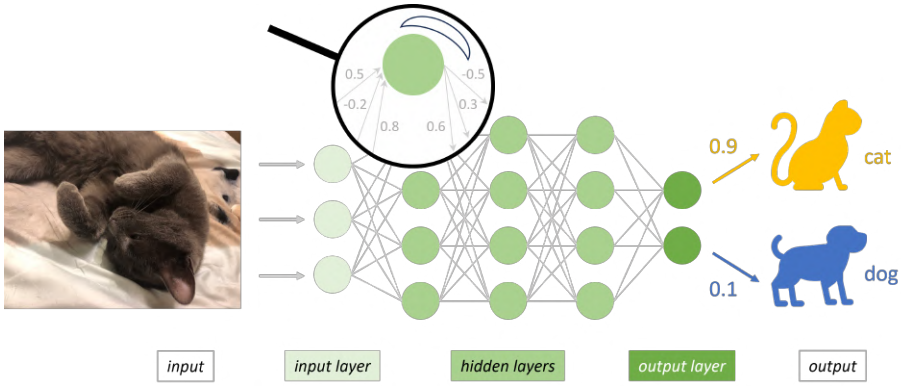


Figure 2.2: A neural network classifies an example image by doing tons of computations, using the weights of its connections. These weights are optimised by the algorithm. (Credits: animals by flaticon.com, and picture of cat Schrödi by Ann De Turck. Schrödi received chicken treats as a reward for his modelling.)

that similar techniques yield performant models in different contexts. Yet, real neural networks can have billions of connections, making it unfeasible for us to understand why they achieve a certain output: we only see billions of weights.

2.2 Why We Need Explainable AI

AI models are often applied as *black boxes*: you have no idea what happens inside; you just put something in and wait until something "magically" comes out (see Figure 2.3 top). If you like the output, you can happily move on with your life. But what if you are suspicious, surprised, or curious about what the model did behind the scenes or why it didn't yield another output (see Figure 2.3 bottom)? In that case, you need an *explanation* for the outcomes. Put differently, the model should be **explainable** or support **explainability**. Getting an explanation is often desirable for at least three reasons: AI models do not always behave as expected, peeking inside black boxes is not always possible or useful, and explainability is becoming a legal right.

AI Does Not Always Behave As Expected

Let's start with a simple example. Say you have an AI model that detects the colour of animals. As shown in Figure 2.4, the model seems to do a perfect

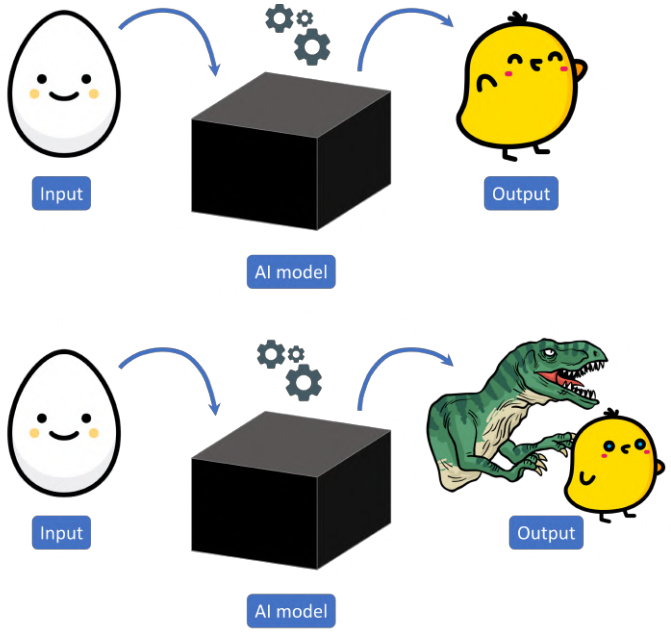


Figure 2.3: AI models are often black boxes: give it some input and wait for the output. **Top**: The rationale behind that output, however, is unclear. **Bottom**: Unexpected outputs sometimes occur without obvious reasons. (Credits: egg by Darius Dan; chick by Smashicons; gear by flaticon.com.)

job: the gorilla is grey, the goat is green, the bear is brown, the bat is blue, and the rabbit is red. But then, all of a sudden it says a black wolf is white. *What's going on?* You might see where this is going: the model wasn't detecting colours at all; it was just generating a colour that starts with the same letter as the given animal. It made many lucky guesses at first but ultimately failed. This toy example illustrates that models can seem highly performant for the wrong reasons. As a consequence, we could fall into the trap of believing that such models are "intelligent" even though they are not.

The example above might seem far-fetched and rather innocent. However, unexpected behaviour of AI models can have severe real-life implications as well. Here are some examples:

- In 2014, Amazon developed an algorithm to screen anonymous resumes of job candidates to predict who was likely to be hired. They found it could still detect the candidates' gender based on their word use and mostly flagged men as suitable (Christian, 2021). This clearly reinforced sexism.

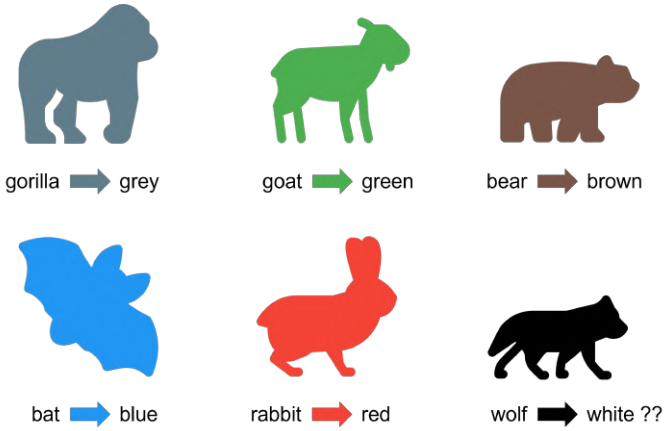


Figure 2.4: A hypothetical AI model correctly detects the colour of five animals, but then makes a strange mistake. (Credits: animals by Vitaly Gorbachev.)

- An algorithm that classified images of melanoma as either cancerous or non-cancerous turned out to classify many non-cancerous images by relying on visible artefacts, for example, medical instruments. During its training, the algorithm learnt these artefacts only show up in images of non-cancerous melanoma (Boggust et al., 2022). If deployed in practice, such an algorithm could have left many cancerous melanomas undetected.
- In 2015, Google started to automatically tag photos in its Photos app. While their algorithm detected Caucasian and Asian faces well, it tagged a selfie of two African Americans as "gorillas" (Mitchell, 2019). It needs no further argumentation that such labelling is completely inappropriate.
- An algorithm to identify objects in an image could be fooled by slightly adapting images with changes invisible to the human eye. After the changes, the algorithm confidently changed its classification from the correct "bus" to "ostrich," for example (Mitchell, 2019; Szegedy et al., 2014). Although funny, this approach could be used to mislead the algorithm for malicious purposes. Similarly, when stickers are attached to traffic signs, self-driving cars might not recognise the signs anymore (Eykholt et al., 2018).

There are many more of these examples and cautionary tales (Branwen, 2011), but you get the point. Both my coloured animals and the more serious examples underline that using AI models as black boxes is not always desirable. Instead, AI models should be able to *explain* their outcomes, such that we can check whether

they work as expected and whether unexpected outcomes are (inevitable) rare side effects or a sign of larger severe shortcomings of the algorithm. In turn, this "explaining to control" could be a stepping stone towards improving the AI model, that is, "explaining to improve" (Adadi and Berrada, 2018).

Peeking Inside Black Boxes Is Not Always Possible or Useful

Maybe you are wondering: why not "peek inside black boxes" to see what they are doing? After all, you recall from Section 2.1 that AI models are just doing complex computations, not magic. Here are two reasons why "peeking" often doesn't work or is not enough.

First, AI models developed and used by companies are often protected by copyrights or intellectual property measures, which does not allow for checking their details and underlying training process. For example, banks might not share how their algorithms determine which clients get a loan, music streaming companies might not share how their algorithms recommend songs that match clients' preferences, and manufacturers of self-driving cars might not share how their cars process sensor inputs to drive autonomously. In other words, "peeking" is literally impossible for outsiders. To still get insights into protected algorithms, there is a need for explanations that focus on algorithms' behaviour, regardless of their technical details. Section 2.4 will discuss how so-called *model-agnostic* XAI techniques deal with this problem.

Second, the currently most performant and thus widely applied AI algorithms are subsymbolic in nature and are being trained on huge amounts of data. Remember from Section 2.1 that such algorithms are *inherently* complex or even infeasible to understand. Take a trained neural network, for example. It is not easy to translate its weights into rules that are understandable by humans because they do not stand for human-interpretable concepts in the first place (Mitchell, 2019). Of course, you might argue that "easier" AI algorithms could bring consolidation. Unfortunately, there seems to be a trade-off today between performance and explainability (Barredo Arrieta et al., 2020; Gunning and Aha, 2019). Figure 2.5 shows how the AI techniques that currently yield the most performant models (neural networks, tree ensembles, and support vector machines) are also the most complex and therefore the least explainable. However, some researchers point out that this trade-off is no definite truth: there might be algorithms that are both very performant *and* interpretable (Liao and Varshney, 2022; Rudin, 2019). In the long run, it might be better to stop explaining black-box algorithms for supporting high-stakes decisions and instead focus on developing performant yet interpretable algorithms (Rudin, 2019). But as long as black-box AI models are being applied in practice, explanations seem

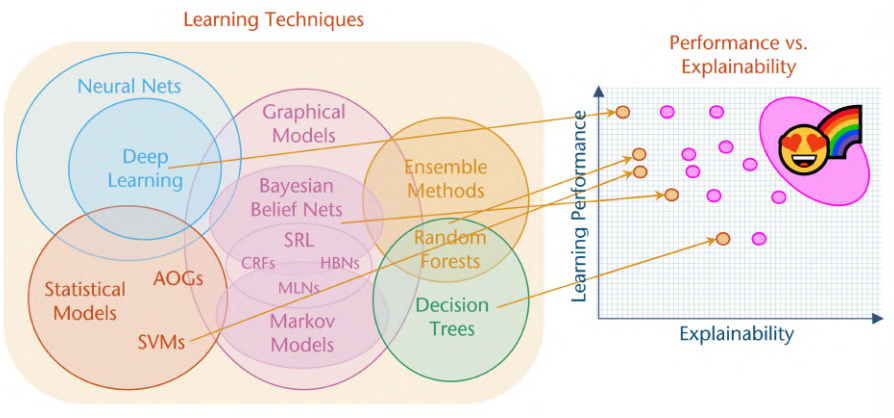


Figure 2.5: The apparent trade-off between performance and explainability: more performant AI algorithms are typically less explainable. XAI tries to push the orange dots towards the pink area, which indicates both high performance and high explainability. (Credits: image based on (Gunning and Aha, 2019).)

crucial to reasonably assess their strengths and weaknesses. At least they push such algorithms more towards the desirable pink area in Figure 2.5.

Researchers sometimes call black-box AI models *opaque*. The two reasons above correspond to two out of three forms of opacity defined in (Burrell, 2016), namely "opacity as intentional secrecy" and "opacity due to scale and how algorithms operate," respectively. Section 2.5 will introduce the third form: "opacity as technical illiteracy."

Explainability Is Becoming a Legal Right

The call for explainability is gradually being reinforced by upcoming legislation, ethical guidelines, and regulations on AI use. The European Union has put itself at the forefront of regulating AI use and automated decision-making in general, protecting people against potentially harmful use of AI technologies. Back in 2016, for example, the adopted General Data Protection Regulation (GDPR) already included a right to explanation (Goodman and Flaxman, 2017; Hamon et al., 2022). So, if an algorithmically made decision significantly affects you, you have the right to ask for an explanation. Very recently, in June 2023, the European Parliament also passed a draft law known as the AI Act (Satariano, 2023). Figure 2.6 shows how this draft law proposes to categorise AI technologies into 3 risk levels: minimal or no risk, high risk, or

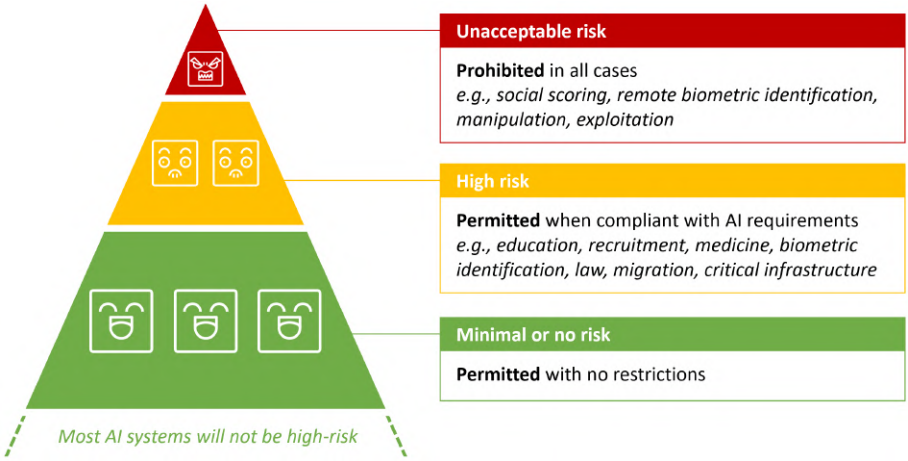


Figure 2.6: Classification of AI systems into 3 risk levels, as proposed in the European AI Act. (Credits: image inspired by (Sioli, 2021); faces by justicon.)

unacceptable risk (Commission, 2023). Most AI technologies will be permitted without restrictions, but AI for sensitive contexts (e.g., education, medicine, and law) will need to comply with specific requirements. AI applications that conflict with EU values will even be banned entirely; for example, social scoring and technologies for manipulation or exploitation. In addition, the AI Act proposes to install supervisory authorities that handle complaints from people affected by AI. Even though these first steps towards AI legislation are sometimes met by criticism (Laux et al., 2022) and concerns about potential restrictions for AI innovation, they underline the urgency of explainable AI.

The AI Act proposes requirements around increasing transparency and supporting human oversight, but this neither enforces XAI nor bans black-box AI (Panigutti et al., 2023). Rather, the transparency requirement demands that AI systems are clearly documented and contain instructions for use, including the system's limitations and capabilities. Furthermore, the requirement of human oversight encompasses that humans should be able to monitor the system's operation, should be aware that they might tend to overly rely on the AI system, and should be able to correctly interpret the system's outcomes. What's important here is that XAI can facilitate all these requirements, but it is not the only solution.

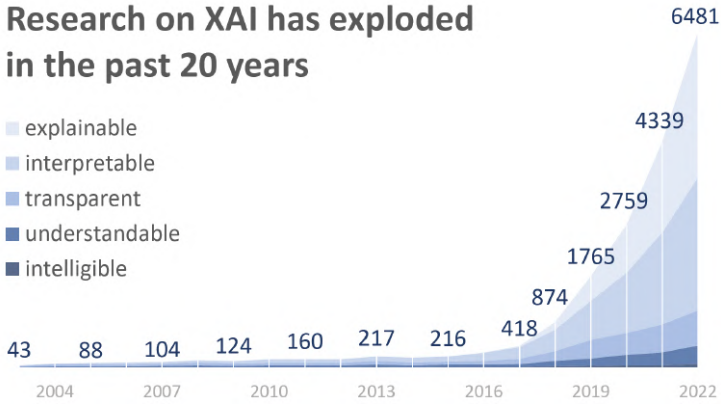


Figure 2.7: XAI-related papers, conference proceedings, and book chapters with the words 'explainable', 'interpretable', 'transparent', 'understandable', or 'intelligible'. Based on a search on Scopus with the query `(understandable OR explainable OR interpretable OR transparent OR intelligible) AND ("artificial intelligence" OR "machine learning" OR "recommender system*" OR "deep learning")`.

2.3 XAI, It's Complicated

The above arguments pro explainability make the challenge clear: we need **explainable AI**, also known as **XAI**. Overall, the goal of XAI is to come up with techniques that allow humans to understand the rationale of AI models, characterise their strengths and weaknesses, and foresee how they will behave in the future (Gunning and Aha, 2019). Researchers poetically call this "opening the black box." The call for human-understandable and simple algorithms is as old as AI itself (Freitas, 2014; Holte, 1993), but especially the past few years were filled with enthusiasm. Figure 2.7 shows how the attention for XAI exploded: XAI research was on the back-burner until 2002, but the number of scientific publications has been increasing dramatically since then. The consensus so far: XAI is a tough nut to crack.

A first challenge for XAI is that there are no widely accepted definitions for terms such as 'explanation' and 'understanding' (Doshi-Velez and Kim, 2017; Lipton, 2018). The same actually holds for the whole of AI: what 'intelligence' means is a deep philosophical question (Legg et al., 2007). Most researchers are pragmatic about this and use different terms interchangeably; some of the most common include 'explainability', 'interpretability', 'transparency', 'understandability', 'intelligibility', 'explicability', and 'comprehensibility' (Adadi and Berrada, 2018;

Barredo Arrieta et al., 2020). Figure 2.7 shows that researchers have historically been using 'interpretability' the longest, but 'explainability' seems to be taking over since DARPA launched its XAI program in 2017 (Gunning and Aha, 2019). These days, researchers seem to typically use 'interpretability' when they are talking about making AI models transparent by design instead of black-box, and 'explainability' when they mean justifying an AI model's behaviour to end-users (Hamon et al., 2022; Panigutti et al., 2023). In this way, interpretability is a passive characteristic: any AI model *is* inherently interpretable or not to a certain degree (Barredo Arrieta et al., 2020). Explainability, however, is an active characteristic: AI models are explainable when they *do* something to clarify or detail their internal functions such that humans can understand them more easily (Barredo Arrieta et al., 2020). Thus, the difference boils down to whether humans are involved. This relates to the next challenge.

A second challenge for XAI is that explainability is a multidisciplinary problem. It can be tackled from at least two perspectives: an algorithmic and a human-centred perspective. Take a look at Figure 2.8. The blue box focuses on the AI system, which involves three parts: the data used for training, the trained AI model, and the outcome. Explanations can focus on each of those parts. Yet, depending on the focus, different explanation techniques are necessary. Mind that these techniques are essentially mathematical in nature. Section 2.4 will discuss these algorithmic XAI approaches in detail. Next, the yellow box focuses on the people using an AI system. Explanations can help them fulfil a specific need, such as assessing the system's fairness or calibrating their trust in the system. The tricky part is that different target audiences have different explainability needs: what computer scientists consider a useful explanation could be incomprehensible for teenagers, for example. In addition, humans are complex creatures who perceive things differently for all kinds of reasons, have different perspectives on what "good" explanations are, have different attitudes towards AI and technology in general, and sometimes hold inconsistent or irrational beliefs. Section 2.5 will discuss how XAI research tries to carve a way through this tricky labyrinth of human perceptions and values.

A third challenge for XAI is a result of the first two: the research field is pretty scattered. Researchers who focus on algorithms have been working rather isolated from researchers who focus on humans, and vice versa (Abdul et al., 2018). In addition, explainability is related to many intertwined topics such as trust, fairness, bias, causality, accountability, privacy, and reasoning (Abdul et al., 2018). Figure 2.9 shows how research into these topics is often isolated. For example, *interpretable machine learning* and *algorithmic fairness* are closely connected because of their focus on algorithms, but there is less overlap with human-centred concepts such as trust and interaction, which are studied more in the context of *recommender systems* and *intelligent agents and systems*.

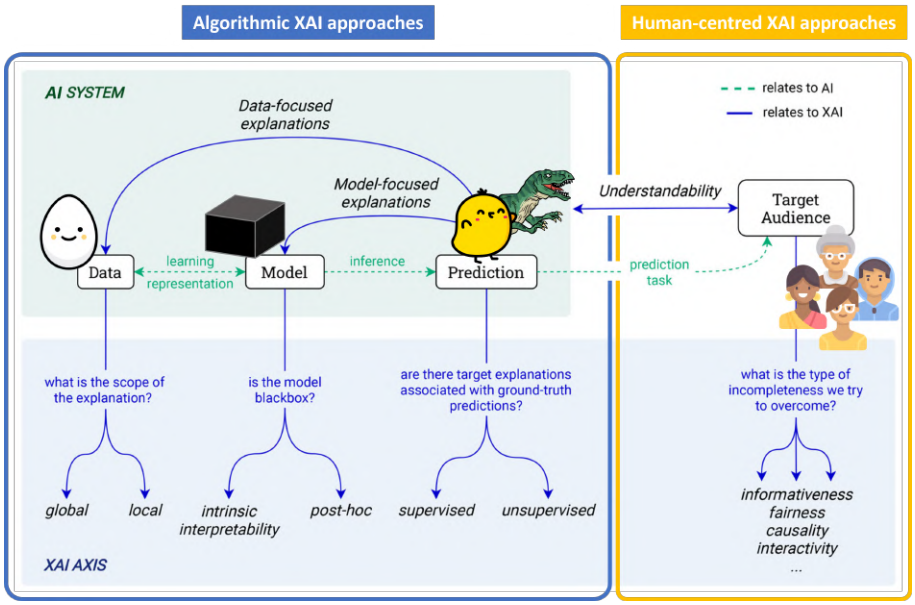


Figure 2.8: XAI is linked to both algorithmic and human aspects. (Credits: image based on (Afchar et al., 2022), egg by Darius Dan; chick by Smashicons; gear, dinosaur, and people by flaticon.com.)

2.4 Algorithmic XAI approaches

Let's start with something you might not realise: although AI techniques such as neural networks are demonstrably powerful, we don't yet fully understand how they work and cannot guarantee they will work in new contexts (Lipton, 2018). And "we" also includes AI experts. That's right, "no one really knows how the most advanced algorithms do what they do" (Knight, 2017). This does not mean AI algorithms are plotting behind our backs to dominate us; it means it is mathematically unclear how neural networks "learn" to generalise. In other words, researchers see their complex AI algorithms yield effective models, but they don't know why. Compare it to anaesthesia for medical operations: while everyone who has been fully sedated for surgery knows it is effective, there is still a lot unknown about why anaesthesia works (TED-Ed, 2015). *But figuring out the mathematics behind AI is hard. Now what?*

Fortunately, researchers have developed tons of algorithmic techniques that give clues about what is happening insides black boxes (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020; Du et al., 2019; Guidotti et al., 2019b; Montavon

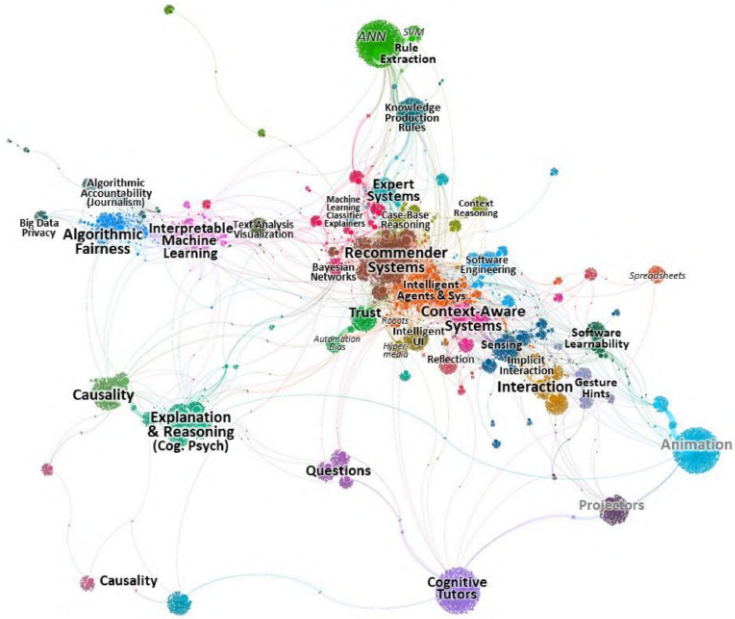


Figure 2.9: Network of different research topics related to XAI. Each circle is a paper and lines connect papers when one cites the other. Tightly connected papers form a cluster and the further clusters are away from each other, the more isolated their topics are. (Credits: image from (Abdul et al., 2018).)

et al., 2018; Stiglic et al., 2020; Vilone and Longo, 2020; Zhang and Chen, 2020). To understand these techniques in detail, you would need technical knowledge about different kinds of AI algorithms, but that would lead us too far. This section will therefore only present the overall ideas. Figure 2.10 shows a general classification of algorithmic XAI approaches: to explain black boxes, we can either turn towards inherently interpretable AI or do reverse engineering. For the latter, we can explain the whole model or single outcomes, either in a model-specific or a model-agnostic way.

Inherently Interpretable AI

Remember the suggestion to "peek inside black boxes" in Section 2.2. Sometimes it makes sense to do so: when an AI model is inherently interpretable, it can justify its outcomes and no further explanation is needed. In other words, a model can itself be an explanation (Afchar et al., 2022). Researchers speak of

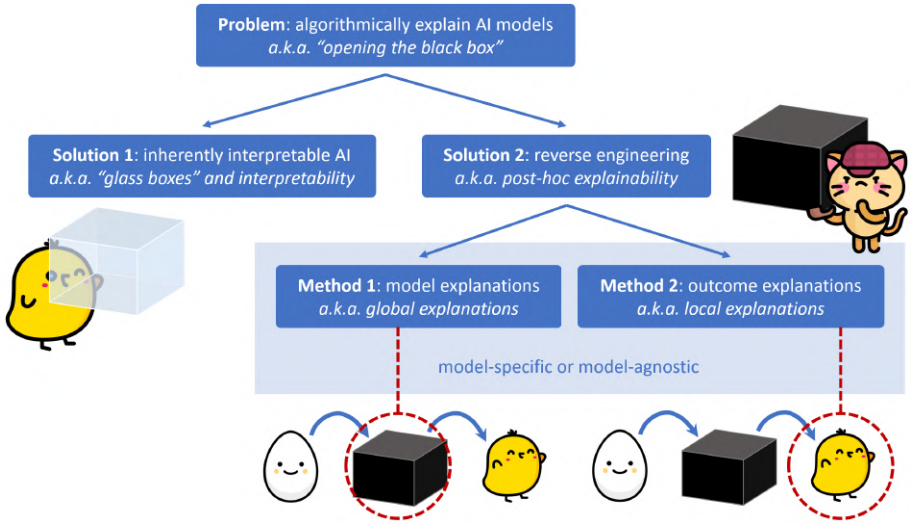
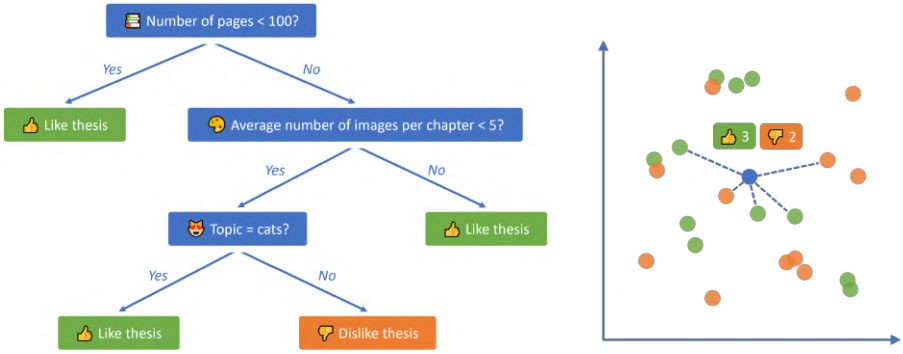


Figure 2.10: Classification of algorithmic XAI approaches. (Credits: image inspired by (Guidotti et al., 2019b); egg by Darius Dan; chicks by Smashicons; cat by flaticon.com.)

"glass" (Abdul et al., 2018; Sokol and Flach, 2018), "white" (Herm et al., 2022; Lundberg et al., 2019) or "transparent boxes" (Barredo Arrieta et al., 2020; Gilpin et al., 2018), and sometimes even "ante-hoc explainability" (Antoniadi et al., 2021; Vilone and Longo, 2020). Two classic examples are decision trees and k-nearest neighbours (Barredo Arrieta et al., 2020).

Decision tree algorithms do what they suggest: they build "trees" to support decision-making. You have already encountered a "tree" in Figure 2.10; to categorise an algorithmic XAI approach, you followed the arrows from the top box until you got to an end. Decision tree algorithms construct such trees based on data. Say you have lots of data about whether readers like thesis texts, together with information about the theses' number of pages, average number of images per chapter, topic, and so on. Then, the algorithm will construct a tree such that it fits the data as well as possible. For example, the end result could look like the decision tree in Figure 2.11a. Given a new thesis text, the decision tree will predict that people like it whenever it has less than 100 pages, studies cats, or has at least five images per chapter on average. I made up this decision tree but it illustrates how it inherently justifies its outcomes: to know what led to a certain outcome, simply follow the path in the tree that led to it.

Similarly, *k-nearest neighbours* algorithms make decisions in an intuitive way.



(a) decision tree **(b)** k-nearest neighbours

Figure 2.11: Two examples of inherently interpretable AI: **(a)** decision trees literally show the path towards decisions, and **(b)** k-nearest neighbours algorithms decides based on the labels of the k most similar data points.

Say you again have a dataset on thesis texts, where each thesis has a label 'like' or 'dislike.' For any new thesis, the algorithm will simply look for the k most similar theses in the dataset. These are the neighbours. Then, the algorithm uses the label that is most frequent among the neighbours as prediction. In general, the number k is fixed and chosen beforehand. In Figure 2.11b, for example, the algorithm uses $k = 5$ and predicts that people will like the new blue thesis text, because most neighbours have a 'like' label (3 out of 5). To conclude, k-nearest neighbours algorithms don't need additional explanations because every outcome is fully determined by its neighbours in the dataset.

Besides decision trees and k-nearest neighbours algorithms, there are 4 more families of inherently interpretable AI techniques: linear or logistic regression, rule-based learners, general additive models, and Bayesian models (Barredo Arrieta et al., 2020). Section 2.5 will make some critical remarks on how transparent all these AI techniques really are for humans.

Reverse Engineering

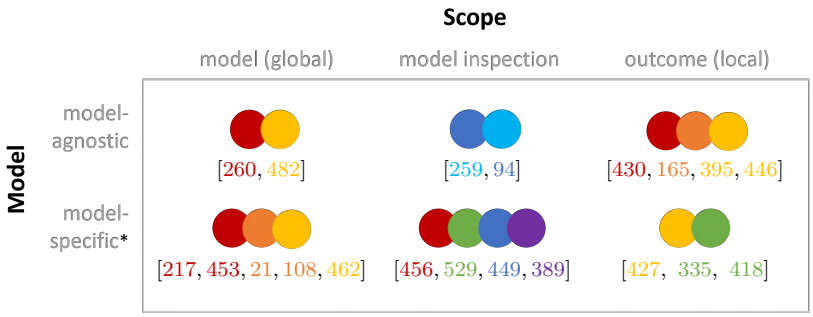
Not all AI algorithms yield inherently interpretable models, however. Section 2.2 explained some AI models are inherently complex, which is why it isn't useful to look at their insides. In such cases, the only thing to work with are inputs and outputs. By studying how these are related, the hope is to learn something about what the AI model is doing. This is called *reverse engineering* (Guidotti et al., 2019b) or *post-hoc explainability* (Adadi and Berrada, 2018; Afchar et al.,

2022; Du et al., 2019). Some researchers, however, object against the latter term because 'explainability' may be misleading: reverse engineering approximates the original model, but there is no guarantee it captures what the model is really doing (Rudin, 2019). Think of the example in Figure 2.4: based on the first few inputs and outputs, you might approximate the model by saying it detects colours in images, but that's not what it does at all. Researchers have developed countless techniques to reverse engineer AI models (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020; Guidotti et al., 2019b; Stiglic et al., 2020), which can be classified in two ways. First, explanations can have different scopes, ranging from single outcomes to the whole model. Second, some explanations only work for models of specific AI algorithms, while others can be applied to any type.

Model vs Outcome Explanations. The first way to classify algorithmic XAI methods relates to their scope, that is, how many outcomes the resulting explanations cover. Explanations covering all possible outcomes are called *model explanations* or *global explanations*; they clarify the algorithm's overall logic. Explanations covering only a single outcome are called *outcome explanations* or *local explanations*; they clarify the algorithm's outcome for a single input. Between these two extremes, it is also possible to learn more about how a model behaves by investigating multiple inputs and corresponding outputs through *model inspection*. This implies that the 'global' and 'local' categories are not strictly separated: you may learn something about a model on a global level by looking at multiple outcome explanations on a local level (Afchar et al., 2022).

Model-Specific vs Model-Agnostic Explanations. The second way to classify algorithmic XAI methods relates to which AI models they can be applied to. Some methods only work for a specific type of model, whereas others work for any model. The former XAI methods are called *model-specific*; the latter *model-agnostic*. Model-agnostic explanations fall into four different types: visualisation, knowledge extraction, influence methods, and example-based methods (Adadi and Berrada, 2018). For example, visualisation gives insights in the AI model's behaviour by showing pairs of inputs and outputs. Visualisation is a strong technique to uncover patterns and will be further discussed in Section 2.6.


Given how we defined 'AI algorithm' and 'AI model' in Section 2.1, the terms 'model-specific' and 'model-agnostic' are slightly confusing. One AI algorithm can generate endless models when given different data, so it would be weird if model-specific explanations only worked for a single model. In reality, model-specific explanations work for any model created by a specific



- **Decision tree**: Approximate the AI model with a decision tree.
- **Decision rules**: Approximate the AI model with decision rules.
- **Feature importance**: How strongly do features determine outcomes?
- **Saliency mask**: Highlight parts of texts or images that influenced the outcome.
- **Partial dependence**: How do outcomes relate to inputs with reduced features?
- **Sensitivity analysis**: How do outcomes change when inputs change?
- **Activation maximisation**: Are there patterns in which neurons are being activated in neural networks for different inputs?

Figure 2.12: Seven general algorithmic XAI techniques together with some examples that realise them, grouped by their scope and which algorithms they explain. Decision trees can be converted to decision rules by listing all decision paths. (Credits: references and classification are from (Guidotti et al., 2019b).)

type of algorithm, so it would be better to call them "*algorithm*-specific." For consistency, we would then talk about "*algorithm*-agnostic" techniques.

Figure 2.12 shows seven algorithmic XAI approaches grouped according to the two categorisations above: decision trees, decision rules, feature importance, saliency masks, partial dependence, sensitivity analysis, and activation maximisation. Most of these approaches occur at different places because they can be implemented in several ways. For example, decision trees can approximate any whole AI model (global, model-agnostic), but also smaller parts of a specific model (model inspection, model-specific).

2.5 Human-Centred XAI Approaches

Research into algorithmic XAI methods is extremely relevant, but it is also important to remember these methods are employed to support *humans*.

Explainability is not a strictly algorithmic characteristic, but lies in how people perceive the explanation (Liao and Varshney, 2022). Many human factors affect how people assess an explanation: the person's technical training and experience with AI, the questions that they want to answer, the context in which an AI-supported decision has to be made, and so on. The following sections will discuss these aspects in more detail.

Sometimes Interpretable Isn't Really Interpretable

Seeing the insides of AI models doesn't necessarily mean *understanding* them (Ananny and Crawford, 2018). We already saw for the case of neural networks that seeing countless weights doesn't help us understand how the networks "reason": we cannot attach meaning to those weights and our minds cannot *simulate* (Lipton, 2018) all computations that involve those weights.

Similarly, algorithmic XAI approaches can make AI models interpretable in principle, but that isn't helpful in practice if people don't find the explanations useful for the insights they are looking for, or if they still cannot simulate them. For example, decision trees are inherently interpretable and are therefore popular to approximate AI models with (see Figure 2.12). However, decision trees can still be impossible to grasp if their decision paths contain hundreds of conditions. The same holds for other "interpretable" AI models such as linear models: if they contain hundreds of parameters, they are not *simulatable* (Lipton, 2018). Overall, if someone cannot simulate an AI model based on an explanation within a reasonable timespan, the explanation isn't really helpful. Yet, what someone considers reasonable is subjective and can only be uncovered with human-centred approaches.

How People Explain Things

Human-centred XAI draws lessons from the social sciences, amongst others, to better align explanations for AI models with how people define, generate, select, present, and evaluate explanations in general (Miller, 2019). Let's briefly discuss three main lessons presented in (Miller, 2019).

First, explanations are typically *contrastive*: people don't ask why a specific event has taken place, but rather why another event didn't take place instead. This inspired XAI researchers to develop algorithmic techniques called *counterfactual explanations*, which compute how much a given input should hypothetically and realistically change to change the original model outcome to a desired one (Dandl et al., 2020; Goyal et al., 2019; Guidotti et al., 2019a; Kaffes et al., 2021; Keane

and Smyth, 2020; Laugel et al., 2019; Moore et al., 2019; Pawelczyk et al., 2020; Poyiadzi et al., 2020; Sharma et al., 2020; Spooner et al., 2021; Wachter et al., 2017; Wang et al., 2021; Yang et al., 2020b). In turn, human-centred XAI researchers study how these counterfactual explanations help decision-making in practice (Barocas et al., 2020; Kasirzadeh and Smart, 2021; Shamma et al., 2022) and meet people's needs (Riveiro and Thill, 2021; Shang et al., 2022).

Second, explanations are *selected*: instead of explaining an event by exhaustively listing all its causes, people typically select one or two causes and consider those to be *the* explanation. For example, if a fan shouts during a tennis rally right before a player hits the ball out, we might say that the miss was caused by the shouting. Doing so, however, we may ignore other contributing causes: the player was extra tensed because they were about to win the championship, there was a slight breeze that blew the ball off course, the player wasn't distracted by the shouting but by the fan's ugly sweater, the fan was shouting the name of the player's secret lover, and so on. In XAI, researchers call selected explanations **justifications**. Justifications explain why specific model outcomes are "good" by providing some easy-to-understand insights about how they were obtained, without covering the full technical reasoning process (Adadi and Berrada, 2018; Ehsan et al., 2019; Vig et al., 2009; Wang et al., 2019a). Not throwing all technical details at "technically illiterate" (Burrell, 2016) people is important because they might be alienating instead of illuminating (Cramer et al., 2008).

Third, explanations are *social*: they are part of a conversation between two parties, where one party is trying to transfer information about an event's cause to another party (Lewis, 1986, p. 217). Important in this conversation is that the explaining party adapts to the other party's current beliefs and knowledge. For example, if AI developers explain their new algorithm to colleagues, they dive into the technical and mathematical aspects because they know their colleagues have the required background for that. But to people with little AI knowledge, AI developers might explain that their algorithm is like the manufacturing process of a coffee machine (assuming they like my metaphor in Section 2.1). For XAI, this means explanations need to be tailored to whoever is receiving them. The next subsection covers how that can be done.

 There are more relevant lessons to be drawn from how people explain things, including that explanations focus on the abnormal, are truthful, and refer to causes instead of probabilities (Miller, 2019; Molnar, 2021).

Different People, Different Needs

Different people have different explainability needs (Ehsan and Riedl, 2020). To address this, Figure 2.13 shows how XAI researchers have proposed to categorise people in at least three broad groups linked to specific explainability needs (Hind, 2019; Langer et al., 2021; Mohseni et al., 2021):

- **AI novices** are people who are impacted by AI systems, but have little to no expertise in the technicalities of AI. These laypeople in terms of AI mainly require explanations to get a better overall understanding of an AI model, so they can assess whether it treats them fairly, they can trust its outcomes, and it protects their data privacy.

  Examples: patients, loan applicants, teachers, regulatory bodies.

- **Data experts** are data scientists and domain experts who use AI systems for analysis, research, or decision-making, but typically lack expertise in the technicalities of AI. Similar to AI novices, they require tools to assess model uncertainty and trustworthiness, but these tools should be more advanced so they can also tune and compare AI models.

  Examples: physicians, loan officers, managers, judges, social workers.

- **AI experts** build and deploy AI models or develop algorithmic XAI techniques. They need to interpret their models to know whether they are working as expected and can be improved.

  Examples: AI researchers, engineers.


Some researchers further refine the user groups and their explainability needs (Hind, 2019; Langer et al., 2021; Suresh et al., 2021). For example, *regulators* such as ethicists, lawyers, and governments supervise how all other groups interact with AI systems and are mainly concerned about trustworthiness and accountability.

The classification above is rather coarse (Liao and Varshney, 2022): groups overlap and especially within the group of AI novices the level of AI expertise can vary quite a lot depending on people's degree or interests. A more fine-grained approach is to directly identify people's explainability needs with questions related to possible insights in AI models (Liao et al., 2020, 2021; Liao and Varshney, 2022). Figure 2.14 shows how these questions can cover what kind of data was used to train the model, what it outputs, how accurate the outcomes are, how they were obtained, how the outcomes relate to the input, why the outcomes weren't different, how the input should change to change the

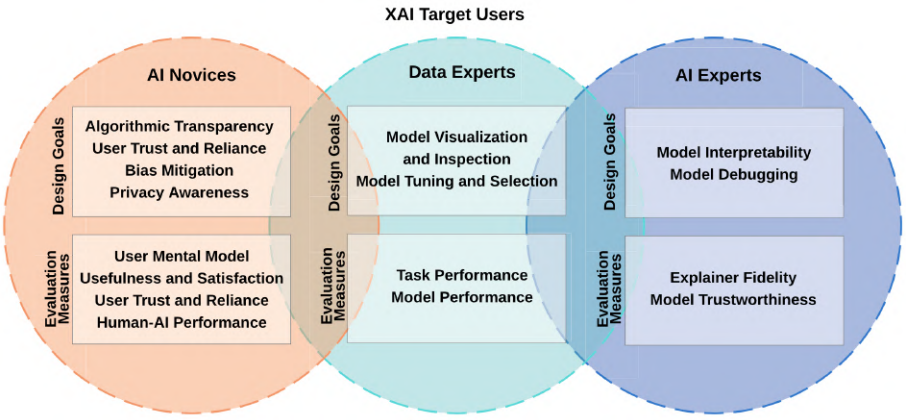


Figure 2.13: Coarse classification of different user groups with respect to AI, together with most common design goals for explainability and evaluation measures. (Credits: image from (Mohseni et al., 2021).)

outcomes, how much the input can change without changing the outcomes, what the outcomes would be for different input, and so on.

Human-Centred Design for XAI

At this point, you might have realised there is no one XAI technique to rule them all. Alas, "explainability is not as simple as providing a nice explanation and all is well" (Weber et al., 2021). And matters are even more complicated. The same people can have different explainability needs in different contexts (Suresh et al., 2021). In sum, **different people in different contexts need different XAI solutions**. For example, during analysis, nurses who use an AI model for monitoring patients at risk might need advanced insights into its performance across the whole pool of patients. However, during a consultation, these insights need to be focused on how one patient can lower their risk and they need to be understandable for the patient too. Thus, it is important to know who needs to know what when, and what explanation types are adequate (Dhanorkar et al., 2021). Moreover, explanations can be represented in many forms, including as a text, a visualisation (see Section 2.6), or a mix of both (Szymanski et al., 2021). To find appropriate explanation techniques and formats for specific people in their specific context, XAI researchers who build explanation interfaces must involve them in a human-centred design process (Abras et al., 2004). In conclusion, "XAI presents as much of a design challenge as an algorithmic challenge" (Liao and Varshney, 2022).

Data

What kind of data was the system trained on?
What is the source of the training data?
How were the labels/ground-truth produced?
What is the sample size of the training data?
What dataset(s) is the system NOT using?
What are the potential limitations/biases of the data?
What is the size, proportion, or distribution of the training data with given feature(s)/feature-value(s)?

Output

What kind of output does the system give?
What does the system output mean?
What is the scope of the system's capability? Can it do...?
How is the output used for other system component(s)?
How should I best use the output of the system?
How should the output fit in my workflow?

Performance

How accurate/precise/reliable are the predictions?
How often does the system make mistakes?
In what situations is the system likely to be correct/ incorrect?
What are the limitations of the system?
What kind of mistakes is the system likely to make?
Is the system's performance good enough for...?

How

How does the system make predictions?
What features does the system consider?
Is [feature X] used or not used for the predictions?
What is the system's overall logic?
How does it weigh different features?
What kind of rules does it follow?
How does [feature X] impact its predictions?
What are the top rules/features that determine its predictions?
What kind of algorithm is used?
How were the parameters set?

Why

Why/how is this instance given this prediction?
What feature(s) of this instance determine the system's prediction of it?
Why are [instance A and B] given the same prediction?

Why not

Why is this instance NOT predicted to be [a different outcome]?
Why is this instance predicted [P instead of a different outcome Q]?
Why are [instance A and B] given different predictions?

How to be that

How should this instance change to get a different prediction?
What is the minimum change required for this instance to get a different prediction?
How should a given feature change for this instance to get a different prediction?
What kind of instance is predicted of [a different outcome]?

How to still be this

What is the change permitted for this instance to still get the same prediction?
What is the range of value permitted for a given feature for this prediction to stay the same?
What is the necessary feature(s)/feature-value(s) present or absent to guarantee this prediction?
What kind of instance gets the same prediction?

What if

What would the system predict if this instance changes to...?
What would the system predict if a given feature changes to...?
What would the system predict for [a different instance]?

Others

How/why will the system change/adapt/improve/drift over time? (change)
Can I, and if so, how do I, improve the system? (improvement)
Why is the system (not) using a given algorithm/feature/rule/dataset? (follow-up)
What does [a machine learning terminology] mean? (terminological)
What are the results of other people using the system? (social)

Figure 2.14: A slightly adapted version of the "XAI question bank," which contains 10 categories of prototypical questions to elicit people's explainability needs. These can then be used to select algorithmic XAI methods that align with the categories (Liao et al., 2020, 2021; Liao and Varshney, 2022).

This is where **human-computer interaction** or **HCI** comes in. Briefly, HCI is an interdisciplinary research field that connects computer science, social sciences, and any other domain that applies technology (Carroll, 1997; Olson and Olson, 2003; Shneiderman et al., 2016). HCI researchers study how interfaces can be designed and tailored to specific end users or application contexts to improve user experience, for example. To do that, HCI researchers work closely together with end users to discover their personal and context-specific needs. In the scope of XAI, this translates to investigating what effective explanations look like and which factors affect their efficacy.

2.6 Visualisation for XAI

So far, we have covered many general examples of explanations. This section focuses on how *visualisation* can compactly represent lots of information in an explanation.

You probably know the saying: "A picture tells more than a thousand words." We humans are incredibly skilled at quickly processing visual information: we can promptly recognise patterns, connect them to meaning, and act upon it. The research domain of **information visualisation** taps into this phenomenon and designs visual representations of data to help people carry out tasks more effectively (Munzner, 2014). Here, 'visualisation' is not just an umbrella term for pictures or graphics such as the schemes in Figure 2.1 and Figure 2.2. It means representing information in a more abstract way; for example, as Venn diagrams (see left part of Figure 2.5), scatter plots (see right part of Figure 2.5 and Figure 2.11b), stacked area charts (see Figure 2.7), networks (see Figure 2.9 and Figure 2.11a), and so on.

Using visualisations for XAI is useful when explanations still contain a lot of information. By representing that information as a well-designed visualisation, you can effectively process it. Yet, the design space for visualisation is huge and whether a visualisation is 'good' depends on the task at hand and the target audience (Munzner, 2014). This is why information visualisation fits well with the philosophy of human-centred design, which we covered in the previous section. The following subsections present how explanations and visualisations can be combined for different target audiences, either with rather simple visualisations or with more complex interactive dashboards. In this thesis, I am using the terms '**visual explanation**' and '**visualisation-supported explanation**' interchangeably; the latter to stress that visual explanations in this thesis are more than highlighted regions in images (Chen et al., 2019) or image descriptions (Hendricks et al., 2016).

Visual Explanations

Figure 2.15 shows five examples of visual explanations. First, Figure 2.15a explains how an AI model predicted which life-insurance plan is suitable for a client by depicting how strongly different parameters were taken into account (Bertrand et al., 2023). Thus, the underlying algorithmic explanation technique is *feature importance*. Here, the model predicted a rather safe plan and the bar chart in the middle shows that this is mainly due to the client wanting to invest a large proportion of their assets. Second, Figure 2.15b is an example of *sensitivity analysis* (Szymanski et al., 2021). An AI model predicts how many seconds a reader would need to finish reading a news article based on parameters such as word count and whether the article contains pictures. The line graph shows how much the prediction would change according to how one of these parameters changes: the predicted time increases when the word count increases and vice versa. Third, Figure 2.15c visualises a *why* explanation: the bars and links show how someone's list of liked songs and the context of those songs led to recommended songs (Bostandjiev et al., 2012). Fourth, Figure 2.15d is similar to Figure 2.15a: it visualises feature importance information for houses (Lundberg and Lee, 2017). However, the bars are replaced by dots and many houses are plotted together, additionally colouring the dots based on their underlying feature value. The visualisation shows, for example, how high values for the second feature (RM = number of rooms) raise the predicted house price. Thus, this example illustrates how similar information can be visualised in different ways. Finally, Figure 2.15e only borderline fits in this list because it isn't really an abstract visualisation of data. Yet, it is an interesting example because it demonstrates an *example-based explanation* (Cai et al., 2019), which we haven't covered before. Specifically, an image recognition model explains why it couldn't recognise someone's drawing by showing the most similar classified training examples it knows and overlays them with the drawing.

Visual Analytics

A specialised subfield of information visualisation is *visual analytics*. Its general goal is to foster analytical reasoning through highly interactive interfaces that combine several visualisations on the same screen (Cui, 2019; Ham, 2010; Keim et al., 2008; Thomas and Kielman, 2009). Concretely, visual analytics is typically meant for data experts and AI experts (Mohseni et al., 2021) (see Section 2.5). It allows them to visually explore large amounts of data so they can discover complex relations, detect biases, and iteratively refine hypotheses. Of course, this requires advanced interactions with the visualisations such as selecting interesting data, exploring different subsets of the data, reconfiguring data by

Figure 2.15: Examples of visual explanations for different AI models. **(a)** Split bar chart for feature importance (Bertrand et al., 2023). **(b)** Line graph for sensitivity analysis (Szymanski et al., 2021). **(c)** Bar charts and network in a *why* explanation (Bostandjiev et al., 2012). **(d)** Bee swarms for feature importances (SHAP website). **(e)** Example-based explanation (Cai et al., 2019).

sorting and rearranging, changing the visual appearance itself, showing more or less details, filtering data on specific conditions, and highlighting connected data in different visualisations (Yi et al., 2007). In addition, given the rise of 'big data', visual analytics is these days often used in combination with AI models that process these huge amounts of data (Chatzimparmpas et al., 2020a,b; Endert et al., 2017; Hohman et al., 2019b; Keim et al., 2010; Liu et al., 2017; Lu et al., 2017). In the context of XAI, data and AI experts use visual analytics to visualise how AI models behave, compare different models, and investigate counterfactual explanations (Chatzimparmpas et al., 2020a,b; Endert et al., 2017; Gomez et al., 2020; Hohman et al., 2019b; Liu et al., 2017; Lu et al., 2017; Zhang et al., 2019). Figure 2.16 shows some impressive examples of how visualisations and interaction can be deeply integrated, and how AI models can be steered through visual control mechanisms. This relates to the next section.

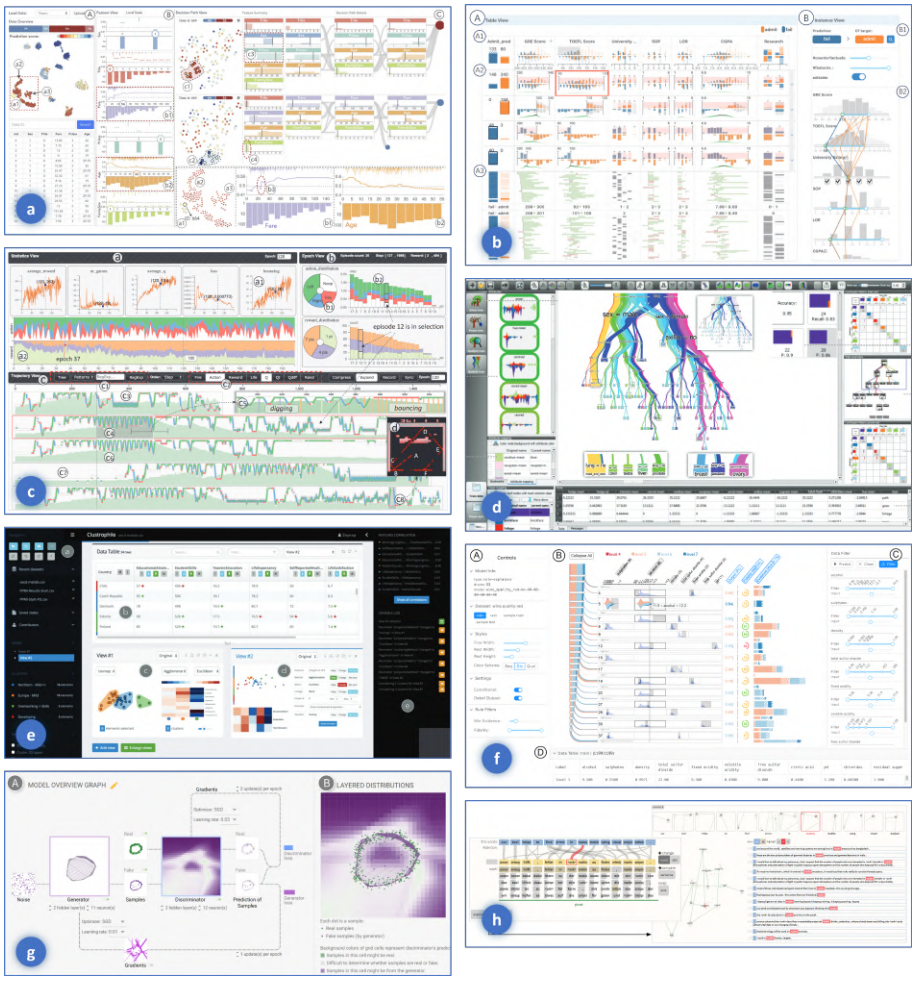


Figure 2.16: Examples of visual analytics systems for **(a)** random forests (Zhao et al., 2019), **(b)** counterfactual explanations (Cheng et al., 2021), **(c)** deep Q-networks (Wang et al., 2019b), **(d)** decision trees (van den Elzen and van Wijk, 2011), **(e)** clustering (Cavallo and Demiralp, 2019), **(f)** decision rules (Ming et al., 2019), **(g)** generative adversarial networks (Kahng et al., 2019), and **(h)** sequence-to-sequence models (Strobelt et al., 2019).

2.7 Control Mechanisms for XAI

So far, we have focused on how XAI can clarify the reasoning process behind the outcomes of AI models. What would be the next step? What should you *do* with this transparency? If you are happy with a model's outcomes and how it works, you might not want to do anything. But if you notice that the model makes faulty inferences, you might want to intervene and correct them (Storms et al., 2022). For example, suppose you like spending me-time on Friday evening while watching romcoms. This week, however, your friend who hates romance comes over for a movie. If your favourite streaming service only recommends romcoms because it infers that's what you like on Friday evenings, you need to somehow tell it about the changed context (Amatriain et al., 2009). Likewise, when an explanation tells you that an AI model comes to the right outcomes for the wrong reasons, you might want to improve its decision process. Imagine your favourite streaming service recommends *the Lord of the Rings*, which you like, but its explanation reveals that it did so because it unrightfully assumes you like fantasy. At that moment, you would need something to tell it you just like *the Lord of the Rings* because you fancy Orlando Bloom. The silver lining in these examples is that transparency might evoke a higher need for control over AI models. Conversely, if you have control over an AI model, you better understand how it uses your feedback to change its decisions. In other words, transparency and control can be two sides of the same coin (Storms et al., 2022) and this is precisely why control mechanisms play such a big role in XAI.

To gain more control over AI models, researchers have developed control mechanisms to actively involve people in the decision process (Jannach et al., 2017). For example, in recommendation systems, you could communicate your initial preferences through forms (Hijikata et al., 2012) or conversational dialogues (Göker and Thompson, 2000) and afterwards, you could steer recommendations through critiquing (Chen and Pu, 2012; Luo et al., 2020; Petrescu et al., 2021), filtering and sorting (Bostandjiev et al., 2012; O'Donovan et al., 2008), interacting with (visual) explanations (He et al., 2016; Schaffer et al., 2015; Tsai and Brusilovsky, 2019b, 2021), or changing the recommendation algorithm itself (Ekstrand et al., 2015). Figure 2.17 shows two examples of how recommender systems can be controlled in combination with visualisations. Yet, how much control and which mechanisms AI systems should incorporate depends on the context, people's personal characteristics, and their mood (Cramer et al., 2008; Jameson and Schwarzkopf, 2002; Jin et al., 2020; Knijnenburg et al., 2011; Konstan and Riedl, 2012; Millecamp et al., 2018; Xiao and Benbasat, 2007). Section 2.6 illustrated this for visual analytics: while data and AI experts might benefit from such highly controllable systems, they are likely too complex for AI novices. This again underlines the importance of human-centred design.

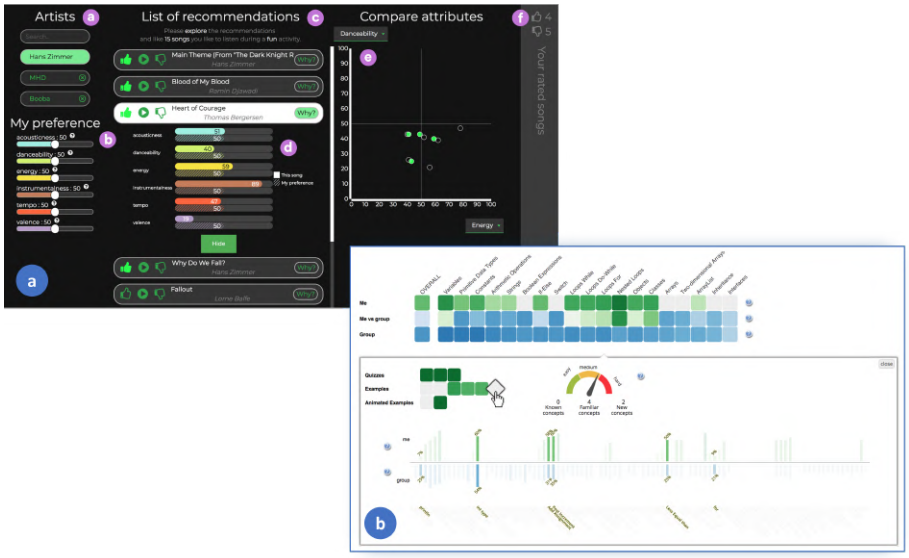


Figure 2.17: Examples of how recommendation systems can be controlled. **(a)** Sliders on the left allow to set preferences for different musical attributes and visualisations show how recommendations match those preferences (Millecamp et al., 2019). **(b)** Learners can follow or override recommended coding exercises by using visualised information about how far they have advanced for different topics and how likely they are to solve exercises correctly (Barria-Pineda, 2020; Barria-Pineda and Brusilovsky, 2019; Barria-Pineda et al., 2018).

Part of the design challenge is how strongly visualisations should be integrated with control over the AI model. There are three integration levels (Turkay et al., 2014). On the first level, visualisations simply present the model outcomes, typically in a static way or with limited interaction possibilities. Thus, you have no control over the model. On the second level, you can modify parameters or the data that the algorithm is using to train the model and the resulting new outcomes are then visualised. The integration is still "semi-interactive," however, because you don't know the model's inner workings and are restricted to changing certain parameters. Finally, on the third level, the model and visualisation are tightly linked: the model can be steered interactively through the visualisation and optionally the model's inner workings are visualised.

2.8 How XAI Can Be Evaluated

Section 2.3 mentioned that 'explainability' is ill-defined. Yet, Section 2.4 showed that researchers nevertheless developed tons of algorithmic XAI techniques, and Section 2.5 stressed how people make everything even more complicated. How do researchers test whether explanations are actually any good? The inconvenient truth is there is no consensus on one overall evaluation method (Vilone and Longo, 2021; Zhou et al., 2021) because of the split between algorithm-centred and human-centred approaches, and different evaluation goals. However, generally speaking, there are three levels of evaluation: functionally-grounded, human-grounded, and application-grounded evaluation (Doshi-Velez and Kim, 2017).

Algorithm-Centred Evaluation

Functionally-grounded evaluation doesn't involve real people in experiments and is therefore algorithm-centred. In this case, "explainability" is optimised according to some formal metric that is supposed to approximate the explanation quality (Doshi-Velez and Kim, 2017). Some widely used metrics are stability, robustness, consistency, sparsity, discriminativeness, and computational efficiency (Afchar et al., 2022). These metrics are intended as a way to translate human wishes for explanations into mathematics. The advantage of functionally-grounded evaluation is that experiments can be run anytime. This is less cumbersome than experiments involving people and also allows to run tests that would be unethical with real humans. For example, giving bad explanations to some test participants and good explanations to others can sometimes be unacceptable. The downside is of course that the chosen metrics and their formal definition define the whole quality assessment and don't necessarily reflect what people think in reality.

Human-Centred Evaluation

Both human-grounded and application-grounded evaluation involve real people in experiments and can assess many different human-centred concepts; for example, overall goodness, satisfaction, understanding, curiosity, trust, reliance, and task performance (Hoffman et al., 2019). Figure 2.13 shows how some concepts are typical for particular user groups. For example, for AI novices, it is common to evaluate explanations in terms of how satisfied people are with them, how they affect people's trust in the AI model, and how well they foster understanding the AI model (for the latter, researchers also use the term

"mental model" (Brachman et al., 2023; Johnson-Laird, 1983; Kulesza et al., 2012, 2013)). These concepts can be measured with a variety of measurement instruments, ranging from carefully constructed questionnaires (Madsen and Gregor, 2000; O'Brien and Toms, 2010; Pu et al., 2011; Vereschak et al., 2021), to interviews (Leech, 2002), to logging what people click on, look at, how long they use the explanation, and so on (Cai et al., 2019). The difference between human-grounded and application-grounded evaluation lies in where the experiments take place.

In *human-grounded evaluations*, researchers assess the quality of explanations based on how participants execute fixed tasks during an experiment in a lab setting (Doshi-Velez and Kim, 2017). These tasks are simplifications of what people might do in real-life applications. Some examples are: participants need to repeatedly choose which explanation they prefer for given pairs (Lundberg et al., 2022); they need to repeatedly guess the output of an AI model for given inputs while seeing an explanation so researchers can assess how well participants understand the explanation (Poursabzi-Sangdeh et al., 2021; Yin et al., 2019); and they need to solve problems or answer questions under different explanation types or formats so researchers can compare them (Bertrand et al., 2023; Bove et al., 2022; Cheng et al., 2019; Gutiérrez et al., 2019b; Szymanski et al., 2021; Wang and Yin, 2021; Yang et al., 2020a).

In *application-grounded evaluations*, participants use an AI system with explanations in real application settings (Doshi-Velez and Kim, 2017). For example, doctors might use the explanations during real consultations with patients, or children might be using them at school to better understand why an AI system is recommending specific exercises. The general idea of this kind of evaluation is that it is best to assess the 'goodness' of explanations directly and in the real applications they were meant for. A big advantage is that explanations can be evaluated from different angles. For example, what explanation types are best suited in specific domains (Afchar et al., 2022). A disadvantage is that such experiments must be carefully planned and executed because many factors may affect the results, making it hard to single out the main effects.

Chapter 3

Thesis Overview

The previous chapter gave a sense of what topics my thesis is covering. Hopefully, you realised how complex it is to explain AI outcomes to humans. As a result, many challenges remain open. Section 3.1 summarises those that inspired the research goals and research questions in this thesis (Section 3.2). Then, Section 3.3 presents the overall human-centred methods we followed to work towards realising those research goals and answering our research questions. The actual research is spread over Chapters 4 to 8. As it might be hard to keep an overview, Section 3.4 clarifies how the rest of this thesis is organised.

3.1 Open Research Challenges

Chapter 2 has introduced XAI, together with many explainability-related concepts and adjacent research fields such as information visualisation, but has only touched upon some of the most pressing research challenges. Even though Chapters 4 to 8 will each start with an in-depth overview of the state-of-the-art and open problems, this section gives a broader view of what lies ahead.

How to Design and Evaluate Explanations With People?

The algorithmic XAI community has developed many techniques to give insights into the reasoning process of AI models (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020; Guidotti et al., 2019b; Montavon et al., 2018; Stiglic et al., 2020). However, it is unclear whether these explanation techniques

meet the insights required by different user groups across application domains and contexts (Mohseni et al., 2021). Furthermore, although it is often claimed that these techniques improve people's understanding of and appropriately trust in AI models, the body of experimental research that backs this up is limited. In general, researchers seem to mainly rely on their intuition of what 'good' explanations are and there is little consensus on how to evaluate them (Doshi-Velez and Kim, 2017). XAI studies with actual people, real-world data, and functional complex models are required (Abdul et al., 2018; Adadi and Berrada, 2018; Gedikli et al., 2014) to investigate how people are affected by explanations, for example in terms of understanding the underlying AI model, trusting it, or feeling satisfied with the explanation. The case of trust is an interesting one, because although transparency is often thought to engender trust, there is little conceptually rich empirical work confirming this (Ananny and Crawford, 2018). Furthermore, trust is a slippery concept because it evolves (Holliday et al., 2016; Nourani et al., 2020), is subject to many factors (Hoff and Bashir, 2015), and can be detrimental when ill-calibrated (Han and Schulz, 2020).

How to Tailor Interactive Visual Explanations?

Information visualisation lies naturally close to XAI since visualisation-supported explanations can effectively communicate complex information. Visual analytics, for example, is a useful technique for data and AI experts to analyse how AI models behave and steer that behaviour accordingly. However, the advanced control possibilities and the typical complex visualisations in visual analytics systems can be overwhelming and do not necessarily align with non-researchers' needs (Kwon et al., 2019). Therefore, an open question is whether design lessons from visual analytics can be transferred to AI novices. Most current visual explanations are namely static (Abdul et al., 2018), even though AI novices might also need to interact with AI systems through visualisations to incorporate their domain knowledge, communicate preferences, or iteratively gain more insights. In this context, designers of explanation interfaces need to make several trade-offs between many desirable explainability goals, such as transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction (Kulesza et al., 2013; Tintarev and Masthoff, 2007b, 2011, 2012). For example, there is a trade-off between effectiveness and efficiency: explanations that present detailed information to help people make good decisions do not necessarily help them make those decisions faster (Tintarev and Masthoff, 2011).

What Are Interesting Application Domains?

To verify whether experimental findings generalise to different contexts, study results should be compared across application domains. This thesis will focus on three domains which have common as well as unique explainability challenges: healthcare, agrifood, and education.

Healthcare. Attempting to further improve human health, healthcare and biomedicine are increasingly collecting large amounts of biological and clinical data in the form of electronic health records, DNA sequence data, imaging, and sensor data, which are then analysed with AI technologies (Litjens et al., 2017; Luo et al., 2016; Miotto et al., 2018; Yu et al., 2018). For example, 'big data' and AI are being used in bioinformatics to study genome-wide associations of diseases, in clinical informatics to increase care for patients (Carriere et al., 2021), in imaging informatics to more efficiently analyse medical imaging (Lee et al., 2021; Li et al., 2007; Liu et al., 2019), and in public health informatics to predict and monitor infectious disease outbreaks (Kopitar et al., 2020; Luo et al., 2016; Stiglic et al., 2018; Viani et al., 2021). However, the black-box nature of complex AI models hampers their adoption in real practice and causal inference (Tu, 1996). While some researchers question whether AI should be held to a higher explanatory standard than physicians (Wang et al., 2020) or hold trust above transparency (Feldman et al., 2019), the general consensus seems that healthcare is in high need of explainable AI models (Ahmad et al., 2018; Holzinger et al., 2019; Stiglic et al., 2020; Vellido, 2020). One reason for this is that medical experts not only have to convince themselves of AI outcomes' validity, but also their patients, who might distrust them if they base their judgement on unexplainable model outcomes (Miotto et al., 2018; Vellido, 2020). Another reason is that healthcare is subject to many medicolegal and ethical requirements because in the extreme case, lives are at stake (Ahmad et al., 2018). Furthermore, medical experts require tools to conduct AI-supported data analysis and need to be integrated more in their design (Vellido, 2020). Finally, it is an open question how model outcomes are best presented to different healthcare stakeholders (Bonnett et al., 2019). Overall, these challenges make healthcare a particularly interesting field to study XAI and visual analytics techniques (Caban and Gotz, 2015; Hu et al., 2016; Preim and Lawonn, 2020; Simpao et al., 2014; West et al., 2015; Wu et al., 2019). There is an especial opening for human-centred research because clinical decision-support systems generally lack explanations which are tailored to clinicians' needs (Antoniadi et al., 2021).

Agrifood. AI and 'big data' are also on the rise in agrifood (Kamilaris et al., 2017), leading to promising research directions such as Agrifood 4.0 (Lezoche et al., 2020), precision agriculture (Cisternas et al., 2020; Linaza et al., 2021; Wachowiak et al., 2017), and smart farming (Ayoub Shaikh et al., 2022; Moysiadis et al., 2021; Wolfert et al., 2017). Example applications include precisely monitoring crop growth (Cisternas et al., 2020; Lindblom et al., 2017) and optimising irrigation (Gil et al., 2021; Kamienski et al., 2018). To process large amounts of data and interact with AI models, agrifood stakeholders increasingly need decision support systems (Zhai et al., 2020). However, even though researchers have proposed many prototypical systems (Gutiérrez et al., 2019a; Zhai et al., 2020), their uptake remains limited so far (McCown, 2002). Possible reasons for this are: current decision support systems lack usability, uncertainty representations, and visualisations; they do not meet end users' needs; and end users often distrust their black-box underlying AI models (Parker, 1999; Parker and Campion, 1997; Rose et al., 2016; Zhai et al., 2020). These challenges could be tackled by combining techniques from XAI, visual analytics, and human-centred design (Lindblom et al., 2017; Parker and Sinclair, 2001; Rose et al., 2017).

Education. The histories of AI and education have always been deeply intertwined (Doroudi, 2022), but especially in recent years, education is embracing technology-enhanced learning for personalised learning (Verbert et al., 2012) and learning is shifting away from traditional classrooms to e-learning environments (Salau et al., 2022). These evolutions make large-scale data collection possible, which in turn inspires *learning analytics* to better understand and support learners based on data (Bodily et al., 2018b). Furthermore, it allows increasing adoption of AI technologies for recommending learning materials (Drachsler et al., 2015; Khanal et al., 2020; Salau et al., 2022; Wu et al., 2020), assessing learners' mastery level (Galici et al., 2023; Torkamaan and Ziegler, 2022), creating educational content (Bitew et al., 2022; Khosravi et al., 2023; Kurdi et al., 2020; Ni et al., 2022), evaluating the quality of learning materials (Conijn et al., 2023), and so on. Similar to other domains, calls for XAI and control mechanisms are emerging in the field (Khosravi et al., 2022). Interestingly, education has a long tradition in both aspects. First, to provide transparency, education has long been studying open learner models, which show learners what the system knows about them (Bull, 2020; Bull and Kay, 2007; Bull and McKay, 2004; Rahdari et al., 2020). Second, to foster metacognitive skills such as self-knowledge and reflection, learners have been given control over all learning aspects, including their learner model, the way learning materials are being selected and presented, and learning materials' difficulty (Brusilovsky, 2023; Bull and Pain, 1995; Kay, 2001; Mabbott and Bull, 2006; Papoušek and Pelánek, 2017). Yet, research typically doesn't include

needs studies of end users (Bodily et al., 2018a) and there is a lack of research on control mechanisms for selecting learning materials (Brusilovsky, 2023). Addressing these challenges is especially hard for education, because learners might not always be ready to control or collaborate with AI due to insufficient knowledge (Brusilovsky, 2023).

3.2 Research Goals and Research Questions

Our research focuses on designing, implementing, and evaluating visualisation-based explanations for systems that integrate AI models such as prediction models and recommendation algorithms. We follow a human-centred approach and thus tailor our explanation interfaces to specific target audiences and application domains. Some of our research objectives are the following:

O1. Evaluate visual explanations in healthcare, agrifood, and education, for example in terms of their ability to foster appropriate trust in AI models and understanding their outcomes;
O2. Study the trade-offs between completeness and complexity for explanation interfaces during human-centred design processes;
O3. Design interaction techniques that allow people to incorporate their domain knowledge into AI systems.

These objectives are complemented by the following broad research questions:

RQ1. How can visual explanations tailored to a target audience and application domain make AI models more transparent?
RQ2. How can people control AI models with additional feedback, supported by interactive visual explanations?
RQ3. How do visual explanations and control affect people's perceptions of AI systems in terms of, for example, appropriate trust and understanding their outcomes?

Given the focus on visualisations and human perceptions, the AI models we implemented in this thesis are not optimised in terms of performance and are less advanced than, for example, the neural networks introduced in Section 2.1. Specifically, Chapter 5 uses a linear regression model and Chapters 6 to 8 use recommendation algorithms based on an Elo rating system.

3.3 Overall Methods

To tackle the research goals and questions, we applied various research methods that each have their own intricacies and learning curves.

First, we conducted a **systematic review** of the literature (Grant and Booth, 2009) to get an overview of the existing research on visual analytics in the scope of XAI. This required carefully constructing a search query (Rethlefsen et al., 2014), tediously screening thousands of papers according to the PRISMA guidelines (Moher et al., 2009), coding the collected papers in a huge Excel spreadsheet, synthesising the coded papers into a coherent story, and finally making recommendations for future research (Bakken, 2019).

Second, we designed explanation interfaces following a **human-centred design** approach (Abras et al., 2004) to ensure people in our target audiences can use our explanation interfaces as intended and can learn how to use them with little effort. Concretely, we collected target-users' needs and iterated over low-fidelity prototypes during multiple *focus groups* (Hennink, 2014) and *think-aloud studies* (Abras et al., 2004). This iterative approach implied we often needed to start over designing parts or even entire interfaces.

Third, we conducted **in-depth semi-structured interviews** (Leech, 2002) and **randomised controlled experiments** (Glennerster and Takavarasha, 2013) to rigorously evaluate our explanation interfaces. In these studies, we collected and analysed data both quantitatively and qualitatively, typically combining both to benefit from both their advantages. Quantitative data, such as log data on how people use our systems and self-reported Likert-type questionnaires based on validated scales, were analysed statistically with parametric and non-parametric approaches (Creswell and Creswell, 2017; Everitt and Hothorn, 2011; Siegel and Castellan, 1988; Snedecor and Cochran, 1969). The most challenging parts here are wrangling the collected data into manageable formats for analysis, and selecting the appropriate statistical methods depending on the data. Qualitative data, such as interview transcriptions and written responses to open questions, were analysed thematically (Braun and Clarke, 2012; Braun et al., 2018). This analysis was arguably the most energy-consuming as I first had to manually transcribe dozens of hours of interviews, code the resulting dozens of pages of text, and then bring all codes together into a coherent story.

3.4 Organisation of the Text

During my PhD, I contributed to 14 papers in total (see Figure 3.1). The rest of this thesis only presents some of them, divided into four parts. The first three each investigate a different approach towards explainability: Part I discusses visual analytics, Part II addresses visualisation-supported justification, and Part III delves into visualisation-supported control. Part IV presents the overall conclusions. The research in Chapters 4 to 8 has been published in scientific journals and conferences or will be submitted there soon. This has two implications.

First, while I am the principal author of the chapters presented in this thesis, the research is the result of close collaboration with multiple colleagues. To acknowledge their contributions, I will use "we" throughout the chapters unless I'm making personal statements. (You might have noticed I already started doing that in this chapter.)

Second, the writing style is an academic one. Phrasings are therefore quite dense: a lot of information needs to be communicated within limited space because publication venues enforce length restrictions and readers have limited attention spans. In addition, it is regularly assumed that readers are familiar with related research and jargon. If you are inexperienced with scientific texts, the upcoming chapters may therefore be more challenging to understand. Still, I invite you to give it a try. If the text really gives you the shivers, it's okay to just focus on the pictures and background stories. Hopefully, they prevent you from running away before you get to the overall conclusions and acknowledgements.

7 **Research publications as first author**

- **Ooge, J.**, Dereu, L., and Verbert, K. (2023). Steering Recommendations and Visualising Its Impact: Effects on Adolescents' Trust in E-Learning Platforms. In Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI'23, pages 156–170, New York, NY, USA. Association for Computing Machinery.
- **Ooge, J.***, Kato, S.*, and Verbert, K. (2022). Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In 27th International Conference on Intelligent User Interfaces, IUI'22, pages 93–105, New York, NY, USA. Association for Computing Machinery.
- **Ooge, J.**, Stiglic, G., and Verbert, K. (2022). Explaining artificial intelligence with visual analytics in healthcare. WIREs Data Mining and Knowledge Discovery, 12(1):e1427.
- **Ooge, J.** and Verbert, K. (2022a). Explaining Artificial Intelligence with Tailored Interactive Visualisations. In 27th International Conference on Intelligent User Interfaces, IUI'22 Companion, pages 120–123, New York, NY, USA. Association for Computing Machinery.
- **Ooge, J.** and Verbert, K. (2022b). Visually Explaining Uncertain Price Predictions in Agrifood: A User-Centred Case-Study. Agriculture, 12(7):1024.
- **Ooge, J.** and Verbert, K. (2021). 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, New Orleans, LA, USA. IEEE.
- **Ooge, J.**, De Croon, R., Verbert, K., and Vanden Abeele, V. (2020). Tailoring Gamification for Adolescents: A Validation Study of Big Five and Hexad in Dutch. In Proceedings of the Annual Symposium on Computer-Human Interaction in Play, pages 206–218, Virtual Event Canada. ACM.

4 **Research publications as co-author**

- Gosak, L., **Ooge, J.**, Fijačko, N., Kamenšek, J., Kocbek, P., Debeljak, N., Verbert, K., and Štiglic, G. (2023). Self-Care Oriented Smartphone Apps for Type 2 Diabetes: A Comparative Analysis. In Proceedings of the Central European Conference on Information and Intelligent Systems, Dubrovnik, Croatia.
- Donoso-Guzmán, I., **Ooge, J.**, Parra, D., and Verbert, K. (2023b). Towards a Comprehensive Human-Centred Evaluation Framework for Explainable AI.
- Bhattacharya, A., **Ooge, J.**, Stiglic, G., and Verbert, K. (2023). Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What- If Explorations. In Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI'23, pages 204–219, New York, NY, USA. Association for Computing Machinery.
- Htun, N.-N., Rojo, D., **Ooge, J.**, De Croon, R., Kasimati, A., and Verbert, K. (2022). Developing Visual-Assisted Decision Support Systems across Diverse Agricultural Use Cases. Agriculture, 12(7):1027.

3 **Research publications under review**

- **Ooge, J.***, Szymanski, M.*, Vanneste, A., and Verbert, K. Steer, See Impact, Solve: How Learner Control and Visual Explanations Impact Learning, Motivation, and Trust. Submitted to CHI 2024.
- Szymanksi, M.*, **Ooge, J.***, Verbert, K. Feedback, Control, or Explanations?: Interaction Mechanisms for Domain Experts in AI-Based Decision-Support Systems. Submitted to LAK 2024.
- Bhatt, S., **Ooge, J.**, Van Den Noortgate, W., Verbert, K. Inferring teacher competencies for personalized learning from teaching and learning analytics on i-Learn. Submitted to Journal of Learning Analytics.

Figure 3.1: Overview of the 14 publications I contributed to during my PhD.

Part I

Explainability Through Visual Analytics

Chapter 4 presents a systematic review on existing visual analytics systems in healthcare. This chapter was published as a journal paper (Ooge et al., 2022b):

> ***Ooge, J.**, Stiglic, G., and Verbert, K. (2022). Explaining artificial intelligence with visual analytics in healthcare.* WIREs Data Mining and Knowledge Discovery*, 12(1):e1427*

As the first author, I conducted the whole review process, classified and analysed the collected papers, wrote the paper, and collected unpublished screenshots of the visual analytics systems. The methods, results, and text were discussed with both co-authors.

Chapter 5 presents an uncertainty-aware visual analytics system for agrifood. This chapter builds on a pilot study published as a workshop paper (Ooge and Verbert, 2021) and was published as a journal paper (Ooge and Verbert, 2022):

> ***Ooge, J.** and Verbert, K. (2021). 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, New Orleans, LA, USA. IEEE*

> ***Ooge, J.** and Verbert, K. (2022). Visually Explaining Uncertain Price Predictions in Agrifood: A User-Centred Case-Study.* Agriculture*, 12(7):1024*

As the first author of both papers, I designed and implemented the visual anaytics system, conducted all interviews, transcribed and analysed them, and wrote the papers. I also presented the first paper during the *TREX 2021* workshop. The methods, results, and text were discussed with Katrien Verbert. Finally, the chapter contributed to the following paper:

> *Htun, N.-N., Rojo, D., **Ooge, J.**, De Croon, R., Kasimati, A., and Verbert, K. (2022). Developing Visual-Assisted Decision Support Systems across Diverse Agricultural Use Cases.* Agriculture*, 12(7):1027*

Chapter 4

Explaining AI with Visual Analytics in Healthcare

To make predictions and explore large datasets, healthcare is increasingly applying advanced algorithms like artificial intelligence. However, to make well-considered and trustworthy decisions, healthcare professionals require ways to gain insights in these algorithms' outputs. One approach is visual analytics, which integrates humans in decision-making through visualisations that facilitate interaction with algorithms. Although many visual analytics systems have been developed for healthcare, a clear overview of their explanation techniques is lacking. Therefore, we review 71 visual analytics systems for healthcare, and analyse how they explain advanced algorithms through visualisation, interaction, shepherding, and direct explanation. Based on our analysis, we outline research opportunities and challenges to further guide the exciting rapprochement of visual analytics and healthcare.

4.1 Introduction

Healthcare professionals are increasingly acquiring vast amounts of electronic health records, analysing these data with advanced algorithms like artificial intelligence (AI), and basing decisions on the algorithmic outcomes (Miotto et al., 2018). Countless examples illustrate the rise of AI in healthcare: Stiglic et al. (2018) and Kopitar et al. (2020) built predictive models for chronic diseases, Liu et al. (2019) detected diseases from medical imaging with deep learning, Viani et al. (2021) and Carriere et al. (2021) applied natural language

processing to extract disease onset from textual health records and to assist with rehabilitation assessment and treatment, etc.

The shift towards "big data" and AI comes with tremendous opportunities for healthcare, but also entails important challenges (Ahmad et al., 2018). A prominent challenge is that well-performing techniques such as deep learning generally yield "black box" models: understanding how they establish outputs is hard or even infeasible. Many healthcare stakeholders deem it unacceptable to fully rely on "black boxes", and call for explaining algorithmic decision processes. This call is further reinforced by medico-legal and ethical requirements, and regulations on AI use like the European GDPR, which endorses a right to explanation (Goodman and Flaxman, 2017).

Constructing explanations for AI models is the holy grail in *explainable artificial intelligence* (XAI), a melting pot of research fields like cognitive psychology, human-computer interaction, and computer science (Hind, 2019). A promising approach for XAI is *visual analytics*. This subfield of information visualisation fosters analytical reasoning through interactive visual interfaces (Cui, 2019; Ham, 2010; Keim et al., 2008): by visually exploring data and iteratively refining hypotheses, users can discover complex relations in large datasets, detect biases, and get insights in how algorithms work (for example through *shepherding*, i.e. controlling the algorithmic process).

Many visual analytics systems have been developed for healthcare, but a clear overview of their explanation techniques is lacking. Therefore, we review visual analytics systems that incorporate advanced algorithms, and that were either specifically designed for a healthcare context, or evaluated therein. Our contribution is twofold. First, we showcase the potential of visual analytics for explaining algorithms according to four perspectives, enclosed in our research questions:

- **RQ1**. How do visual analytics systems visualise the outcomes of advanced algorithms?
- **RQ2**. Which interactions do visual analytics systems support?
- **RQ3**. How do visual analytics systems support shepherding of advanced algorithms?
- **RQ4**. How can visual analytics systems explain advanced algorithms directly?

Second, we analyse main trends, opportunities and remaining challenges for visual analytics in healthcare. Along the way, we report which advanced algorithms are incorporated in visual analytics systems for healthcare, and for which purposes they are used. We present our findings with an interdisciplinary

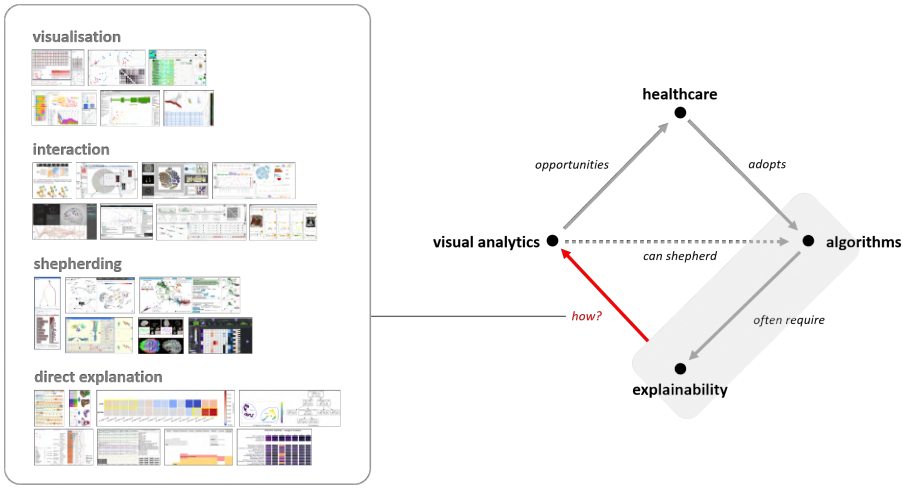


Figure 4.1: Healthcare increasingly adopts advanced algorithms, and often requires explanations for the algorithmic process. Visual analytics can provide insights in algorithms through visualisation, interaction, shepherding and direct explanations. Thus, visual analytics holds important opportunities for healthcare.

audience in mind, and thus hope to further strengthen the bridge between visual analytics, AI and healthcare.

4.2 Background and Related Work

Our review touches upon healthcare, advanced algorithms, explainable AI, and visual analytics. This section presents relevant work in the intersection of these domains.

4.2.1 Explainable Artificial Intelligence

XAI encompasses a huge collection of intertwined topics, including trust, fairness, bias, causality, accountability, privacy and reasoning (Abdul et al., 2018). One side-effect of this rich mix is that researchers have not yet agreed upon a rigorous definition for explainability, and often interchange it with interpretability, understandability, or intelligibility (Gilpin et al., 2018; Lipton, 2018).

Human-computer interaction recognises that the meaning of explainability and its requirements depend on the target user and application context. Mohseni et al. (2021) classified target users into AI novices, data experts, and AI experts, each needing unique design goals and evaluation measures. Wang et al. (2020) and Ahmad et al. (2018) pointed out that the importance of explanations depends on the healthcare application: they are crucial when care is affected, but less pressing for treatment cost prediction. To determine a suitable explanation level, Vellido (2020) argued to integrate healthcare experts in the design of data analysis interpretation strategies.

Even though explainability lacks a formal definition, the AI community has developed many explanation techniques for AI models (Adadi and Berrada, 2018). Guidotti et al. (2019b) categorised these techniques according to how they open the "black box" problem: by explaining the model itself, by explaining the outcomes, or by inspecting the model. Stiglic et al. (2018) and Du et al. (2019) categorised explanation techniques by scale (local vs global) and type (model-specific vs model-agnostic): local explanations focus on a single instance, whereas global explanations try to explain the entire model; model-specific explanations are only applicable for particular models (e.g., deep learning models (Montavon et al., 2018)), whereas model-agnostic explanations can explain any model (Ribeiro et al., 2016).

4.2.2 Visual Analytics for Explainable Artificial Intelligence

Many authors have surveyed visual analytics systems in the scope of AI. Some surveys mainly focus on the machine learning aspect. For example, Liu et al. (2017) classified visual analytics systems by whether they are intended to understand, diagnose or refine machine learning models; Endert et al. (2017) considered the machine learning type and the interaction intent. Other surveys rather focus on the visualisation aspect. For example, Lu et al. (2017) categorised predictive visual analytics systems based on their interaction methods and prediction tasks; Hohman et al. (2019b) discussed the why, who, what, how, when, and where of visualising deep learning models. Chatzimparmpas et al. (2020a) covered both machine learning and visualisation aspects in a review on enhancing trust with interactive visualisations: their fine-grained classification covers interaction type, machine learning model, and trust level.

4.2.3 Visual Analytics in Healthcare

Ever since visual analytics emerged, healthcare has been recognised as one of its most promising application areas (Keim et al., 2008; Thomas and Kielman,

Table 4.1: Query used for paper selection. Categories were combined with an *AND* operator.

Category	Keywords
algorithm	ai OR algorithm* OR artificial intelligence OR automated OR big data OR data mining OR deep learning OR machine learning OR predict*
analytics	analytics OR data analy* OR decision support OR electronic health records
healthcare	*medic* OR bioinformatics OR clinic* OR health*
interaction	explor* OR interact*
visualisation	dashboard OR graphic* OR interface OR visual*

2009) because of the many opportunities for clinicians, patients, researchers, and other healthcare stakeholders (Caban and Gotz, 2015). The fruitful interplay between healthcare, medical visualisation, and visual analytics produced an extensive jargon, including visual intelligent decision support systems (Ltifi and Ayed, 2016), clinical informatics (Simpao et al., 2014), and health informatics (Wu et al., 2019).

Despite the diverging terminology, a lot of interesting work has been presented in different healthcare areas, for example population health services (Chishtie et al., 2020), prevention of disease outbreaks (Preim and Lawonn, 2020), and cancer-related genomics (Qu et al., 2019). In biomedics, Sturm et al. (2015) categorised existing work on interactivity level vs analysis type and visualisation technique, and Turkay et al. (2014) classified visual analysis tools by their analytical task and integration of computational methods. Finally, Rostamzadeh et al. (2020) and West et al. (2015) reviewed interactive visualisations of electronic health records, and Wang et al. (2011) presented case studies and design guidelines based on their experiences with *Lifelines2*.

To conclude, a rich set of surveys highlights the importance of explaining machine learning. However, the general surveys on visual analytics in Section 4.2.2 are not framed in a healthcare context and do not discuss its specific requirements. In contrast, the reviews in Section 4.2.3 are healthcare-oriented, but do not focus on explanations or only cover a specific healthcare subdomain. To shed light on the intersection of healthcare, visual analytics and explanation techniques, we review visual analytics systems for healthcare that facilitate algorithmic explainability. We also extend the scope from machine learning to advanced algorithms in general.

4.3 Paper Collection and Classification Process

Starting from the key reviews in Section 4.2, we iteratively compiled the search query in Table 4.1 to target interactive visualisations that were specifically developed for or applied in a healthcare context, and that involve at least one advanced algorithm. Our query also considers diverging terminology for similar concepts: for example, "decision support systems" in healthcare can fit with what the visualisation community considers as interactive visual dashboards.

In January 2021, the first author queried Scopus, and then screened 1908 matches based on their abstract (285 remaining) and full-text (83 remaining, including 12 overview papers). We excluded papers that present fully static or interaction-limited visualisations (i.e. first level of integration in (Turkay et al., 2014)), solely discuss image processing outcomes, do not allow for data analysis, present statistical or visualisation software, or do not involve advanced algorithms. The latter condition excluded dashboards like *LifeLines2* (Wang et al., 2011). Finally, the first author classified all included papers in Tables 4.2 and 4.3, inspired by existing frameworks for activity types (Rostamzadeh et al., 2020), algorithmic classes (Endert et al., 2017), interaction types (Yi et al., 2007), and explanation techniques (Guidotti et al., 2019b; Mohseni et al., 2021).

Sections 4.4 to 4.7 present Tables 4.2 and 4.3 in detail. Each section treats one technique to obtain insights in advanced algorithms: visualisation, interaction, shepherding, or direct explanation.

4.4 Visualising Algorithmic Outcomes in Visual Analytics

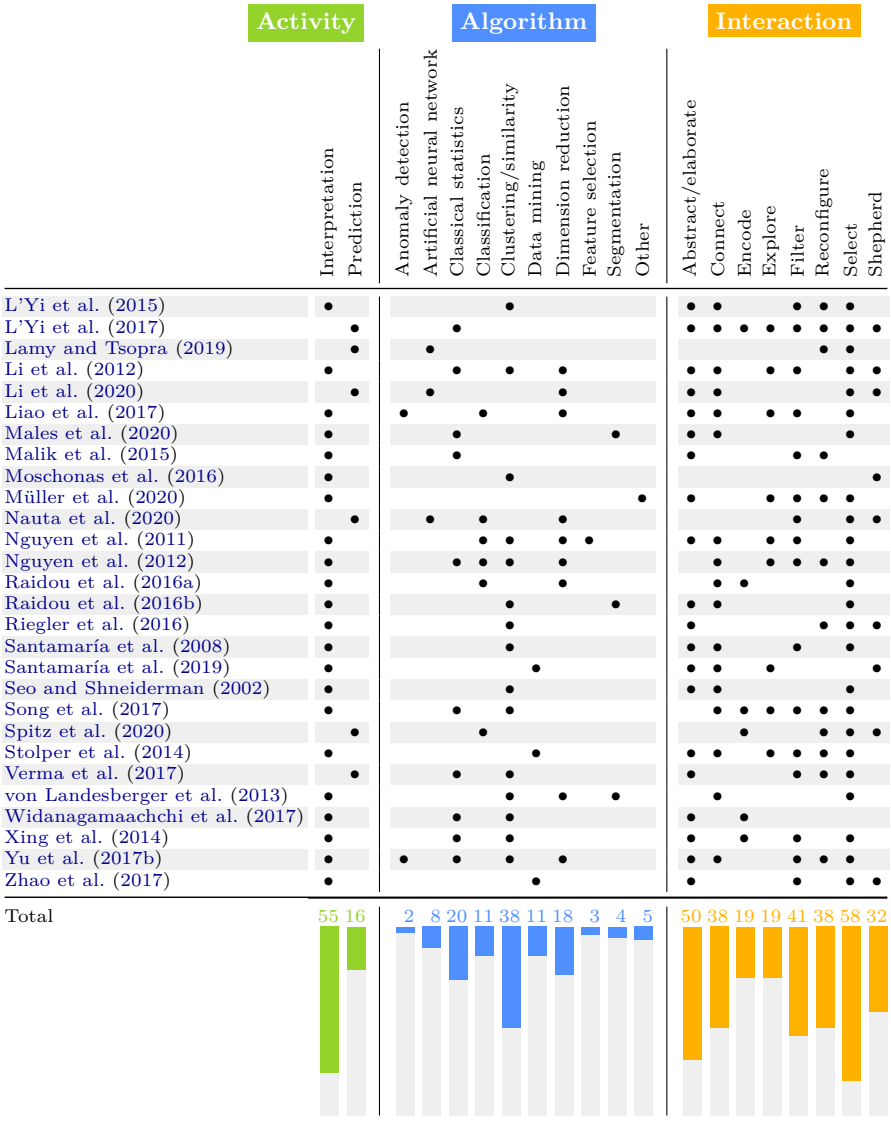
A first way to gain insights in advanced algorithms is to visualise their outcomes. Based on the first two column groups in Table 4.2, we discuss several visualisation approaches (cfr. RQ1) for distinct algorithmic families, and uncover the healthcare activities for which visual analytics systems in our sample are used.

Table 4.2: Classification of 71 visual analytics systems in our sample according to healthcare activity types, present algorithms, and interaction types.

	Activity		Algorithm										Interaction							
	Interpretation	Prediction	Anomaly detection	Artificial neural network	Classical statistics	Classification	Clustering/similarity	Data mining	Dimension reduction	Feature selection	Segmentation	Other	Abstract/elaborate	Connect	Encode	Explore	Filter	Reconfigure	Select	Shepherd
Abbasloo et al. (2019)	●				●									●	●	●			●	●
Abdullah et al. (2020)	●						●		●				●	●			●	●	●	●
Afzal et al. (2011)		●										●	●			●			●	●
Alsaad et al. (2019)		●		●		●							●							
Barlowe et al. (2013)	●						●								●			●	●	●
Behrisch et al. (2018)	●						●					●	●		●	●				
Borland et al. (2020)	●						●						●		●					●
Brunker et al. (2019)	●				●		●							●	●		●	●	●	●
Cao et al. (2011)	●						●						●	●				●	●	●
Clark et al. (2017)	●				●								●						●	●
Dang et al. (2015)	●						●						●	●				●	●	
Dingen et al. (2019)		●			●								●					●	●	●
Dixit et al. (2017)	●					●		●							●		●	●	●	●
Fang et al. (2017)	●							●					●				●	●		
Farag et al. (2015)	●						●		●					●			●	●	●	
Feller et al. (2018)	●						●								●		●			●
Geurts et al. (2015)	●						●				●		●						●	
Gotz et al. (2011)	●						●						●	●				●	●	●
Gotz et al. (2014)	●							●					●				●			●
Gotz et al. (2020)	●				●		●						●		●		●	●	●	●
Guo et al. (2020)	●			●			●		●				●	●					●	
Guo et al. (2018)	●						●						●	●			●	●	●	●
Herold et al. (2010)	●					●													●	
Hinterberg et al. (2015)	●				●												●		●	
Huang et al. (2015)	●						●										●	●	●	●
Huang et al. (2019)		●					●						●						●	
Hund et al. (2016)	●						●		●				●	●	●	●	●	●	●	●
Hur et al. (2020)		●		●				●					●				●		●	
Ji et al. (2017)	●								●				●	●		●	●	●	●	●
Ji et al. (2019a)	●						●		●				●	●		●		●	●	
Ji et al. (2019b)	●			●	●		●		●						●		●	●	●	●
Jönsson et al. (2019)	●				●									●		●	●			
Kakar et al. (2019)	●							●					●	●			●	●	●	
Klemm et al. (2014)	●						●							●					●	
Klimov et al. (2015)	●							●					●				●	●		
Kovalerchuk et al. (2012)		●						●					●		●	●		●	●	
Krause et al. (2014)		●				●				●			●	●	●	●	●	●	●	●
Krause et al. (2016)		●					●					●	●	●			●	●	●	●
Krause et al. (2018a)		●				●							●				●			
Kumar et al. (2015)	●				●		●			●			●	●			●	●	●	
Kwon et al. (2018)	●				●		●		●				●	●	●	●	●	●	●	●
Kwon et al. (2019)		●		●					●				●	●			●	●	●	●
Kwon et al. (2021)	●							●				●		●			●	●	●	

Continued on next page

Table 4.2 – Continued from previous page



Activity Rostamzadeh et al. (2020) divided healthcare activities into interpreting, predicting, and monitoring. ■ *Interpretation* is the most frequently supported activity in our sample (55/71). Thus, most visual analytics systems are geared towards exposing patterns in data, and discovering relations among features. Less common is ■ *Prediction* whereby outcomes are anticipated

or hypotheses are formed based on the available data (16/71). Finally, ■ *Monitoring* is absent in our sample: no visual analytics systems help to manage recurrent or chronic diseases.

Algorithm The visual analytics systems in our sample incorporate a multitude of algorithmic families. *Clustering/similarity* is the most common family (38/71): its algorithms are mainly used for interpreting, which explains that activity's prevalence. ■ *Clustering* algorithms group similar data points, say similarly nutritious meals (Feller et al., 2018) or genomes (Seo and Shneiderman, 2002). Results from k-means and k-nearest neighbours are usually visualised in scatter plots with dots coloured according to their cluster (e.g., (Ji et al., 2019b; Klemm et al., 2014), Figure 4.2a), or in projection plots after dimension reduction, e.g., (Guo et al., 2020; Ji et al., 2017). To avoid dimension reduction, Abdullah et al. (2020) and Kwon et al. (2018) (Figure 4.2a) use parallel sets: axes that represent features are connected by ribbons to show the proportional distribution of feature combinations for each cluster. Alternatively, Gotz et al. (2011) and Cao et al. (2011) (Figure 4.4a) visualise clusters of patients as Voronoi treemaps, where cells represent features of patients. Interestingly, L'Yi et al. (2015) apply parallel sets to compare clustering algorithms. Two more clustering algorithms are biclustering and hierarchical clustering. Santamaría et al. (2008) visualise biclustered microarray data as a Venn diagram. Cluster hierarchies are typically visualised as dendrograms next to a heat map matrix, e.g, (Farag et al., 2015; Yu et al., 2017b) in Figure 4.2d. Other visualisations are: an expandable list of cluster representatives (Raidou et al., 2016b), inductively grouped circles (Behrisch et al., 2018), and a time line of clusters at a given hierarchy level (Widanagamaachchi et al., 2017).

We classified ■ *Similarity* measures like cosine, Jaccard and Hellinger distance together with clustering, because they are often used to group data. For example, Borland et al. (2020) (Figure 4.5d) hierarchically aggregate similar events in an icicle plot to track selection bias in patient cohorts; Barlowe et al. (2013) (Figure 4.2b) rank protein flexibility plots by similarity to a target.

Next comes ■ *Classical statistics* (20/71), which illustrates that current systems still heavily rely on non-machine learning algorithms. *Classical statistics* includes three techniques. (1) Correlation analysis: cells in correlation matrices are typically colour-coded to reveal patterns and outliers, e.g., (Males et al., 2020; Song et al., 2017). (2) (Non)parametric hypothesis testing: Abbasloo et al. (2019) and Jönsson et al. (2019, 2020) (Figure 4.3b) highlight brain regions that significantly differ for people with and without a condition; Malik et al. (2015) visualise patients that survived or died after certain event sequences as back-to-back bar charts, and indicate significant differences. (3) Regression: Males et al. (2020) (Figure 4.2c) compare two groups' colon morphology in

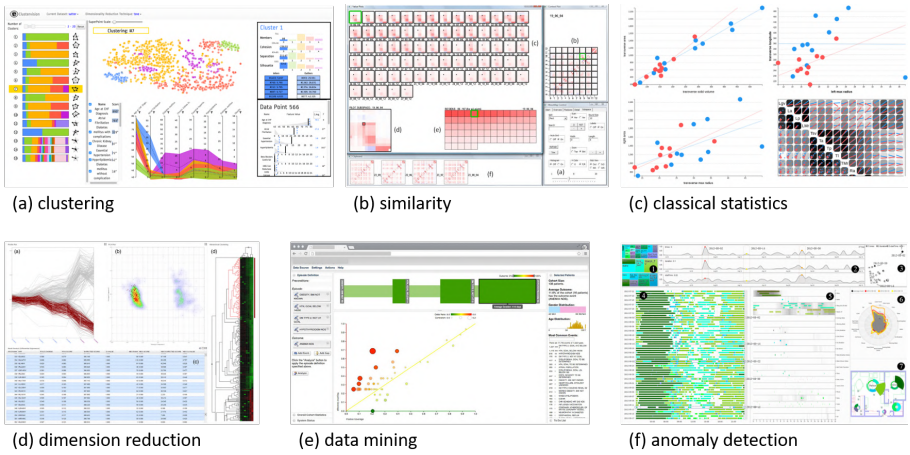


Figure 4.2: Examples of visualising algorithmic outcomes. **(a)** Clustering outcomes in a scatter plot and parallel coordinates (Kwon et al., 2018); **(b)** Plots ranked according to similarity to the first one (Barlowe et al., 2013); **(c)** Classical regression lines in scatter plots (Males et al., 2020); **(d)** Top: parallel coordinates and projection plot with heat map after dimension reduction. Right: heat map matrix and dendrogram of hierarchical clustering (Farag et al., 2015); **(e)** Results of sequential pattern mining in a scatter plot (Gotz et al., 2014); **(f)** Anomalous activities highlighted in several visualisations (Liao et al., 2017).

overlaid scatter plots and regression lines; Verma et al. (2017) predict adverse drug reactions with logistic regression, and visualise the confidence as ribbon width in a Sankey diagram.

The following techniques are more commonly associated with AI: *Dimension reduction* (18/71), *Data mining* and *Classification* (both 11/71). ■ *Dimension reduction* projects multidimensional data to 2D or 3D with principal component analysis (PCA), singular value decomposition, multiple factor analysis, t-SNE, UMAP, or self-organising maps. In our sample, PCA is the most popular technique, for example for omics data analysis (Farag et al., 2015; Nguyen et al., 2012) or feature extraction on bacteria's infrared spectroscopy spectra (Ji et al., 2019a). Reduced dimensions are typically visualised in a 2D scatter plot, which can be augmented with a density heat map, or accompanied by a parallel coordinates visualisation of the original data (Farag et al., 2015; Ji et al., 2019a) (Figure 4.2d). Only (Nguyen et al., 2012, 2011) use 3D scatter plots.

■ *Data mining* algorithms extract patterns from data. With sequential pattern mining, Fang et al. (2017) highlight similarities in line graphs of health sensor

data, Gotz et al. (2014) (Figure 4.2e) query event sequences and correlate them with positive and negative outcomes, Santamaría et al. (2019) highlight nucleosome patterns in sequenced chromosomes, and Klimov et al. (2015) identify risk factors for kidney damage in parallel coordinates. Furthermore, Dixit et al. (2017) optimise care pathways with process mining, and Zhao et al. (2017) mine association rules for cancer causes and visualise them in parallel coordinates.

■ *Classification* algorithms like random forest and k-nearest neighbours assign data points to a class, for example to predict rupture risk of aneurysms (Spitz et al., 2020). Their overall performance is typically visualised in a confusion matrix, e.g., (Krause et al., 2018a), and data points' classes are often indicated with colours, e.g., (Herold et al., 2010; Nguyen et al., 2012, 2011).

The remaining five algorithmic families include specialised techniques. ■ *Artificial neural network* mainly consists of recurrent neural networks with long short-term memory (LSTM) to handle long-range temporal dependencies. ■ *Feature selection* contains techniques like information gain, which identify the most relevant features in a dataset. ■ *Segmentation* algorithms partition medical images into multiple segments. ■ *Anomaly detection* identifies unusual data points, for example anomalous activities in smart homes (Liao et al., 2017) (Figure 4.2f). Finally, ■ *Other* contains algorithms like partial dependence, epidemiological models, Bayesian networks and hidden Markov models, which do not fit in any of the previous families.

Overall, algorithmic outcomes can be visualised in many typical and alternative ways, depending on the algorithmic family and the desired insights. These insights are in turn related to the healthcare activity which is most often interpreting data, rather than predicting or monitoring.

4.5 Interaction in Visual Analytics

While static visualisations may already provide interesting insights in advanced algorithms, adding interaction makes them more powerful as users can then test hypotheses, focus on particular insights, or view information from different angles. The **Interaction** group in Table 4.2 classifies our sample into seven interaction types proposed by Yi et al. (2007) (cfr. RQ2). *Select* (58/71) and *Abstract/elaborate* (50/71) are the most frequently supported types, whereas *Encode* and *Explore* (both 19/71) the least. Most visual analytics systems support multiple interactions and as *Select* co-occurs with all other interaction types, we present those first. Note that Lu et al. (2017) introduced shepherding as an interaction type, yet we discuss it separately in Section 4.6.

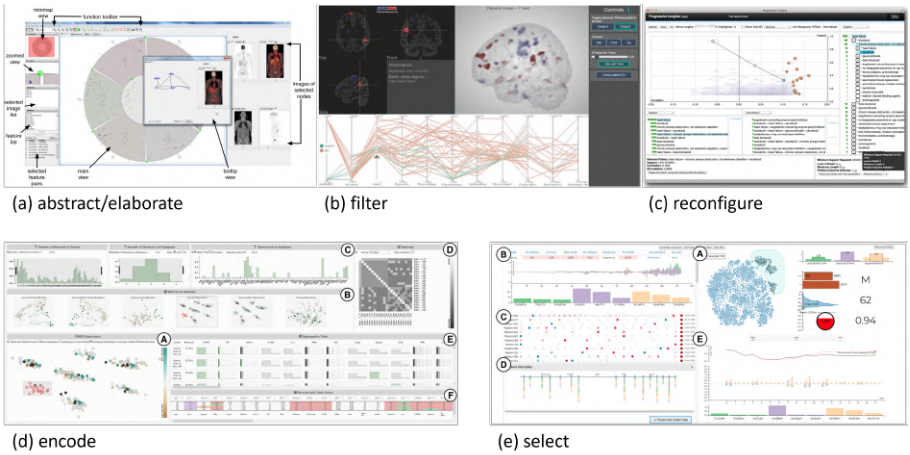


Figure 4.3: Examples of interaction with visualisations. **(a)** Abstract or elaborate information with zooming (Kumar et al., 2015); **(b)** Filter data with sliders in parallel coordinates (Jönsson et al., 2019, 2020); **(c)** Reconfigure a scatter plot by changing axes' features, and connect a selected data point with related points (Stolper et al., 2014); **(d)** Encode data differently by adding spikes to points in a scatter plot, and brush-and-link (Hund et al., 2016); **(e)** Select nodes in a scatter plot with lasso selection (Kwon et al., 2019).

■ *Abstract/elaborate* interactions show or hide details in four ways. (1) Tooltips that pop up when hovering or clicking a visualisation can show raw data, e.g., (Malik et al., 2015; Xing et al., 2014), or additional visualisations, e.g., (Afzal et al., 2011; Gotz et al., 2014). (2) Collapsing components removes visual clutter, for example lines in line graphs (Afzal et al., 2011) or parallel coordinates (Huang et al., 2019), identical rows in matrices (Dang et al., 2015), or similar points in scatter plots (Kwon et al., 2018). Conversely, expanding components shows extra information like individual lines instead of ribbons in parallel sets (Kwon et al., 2018), subsequences of sequential health records (Malik et al., 2015), or groups in icicle plots (Borland et al., 2020). (3) Zooming enlarges a visualisation, e.g., (Kumar et al., 2015; Males et al., 2020) (Figure 4.3a). An interesting zooming variant is lensing, which enlarges a specific area and compresses the rest: Dang et al. (2015) use it in a large matrix of protein-biomolecule reactions. (4) To change the visualised level of a clustering hierarchy, Widanagamaachchi et al. (2017) and Behrisch et al. (2018) provide a slider, and Seo and Shneiderman (2002) a bar that can be dragged along a dendrogram.

■ *Filter* interactions allow to focus on insights of interest by setting conditions on the data with check-boxes, radio buttons or sliders. Examples are: filtering

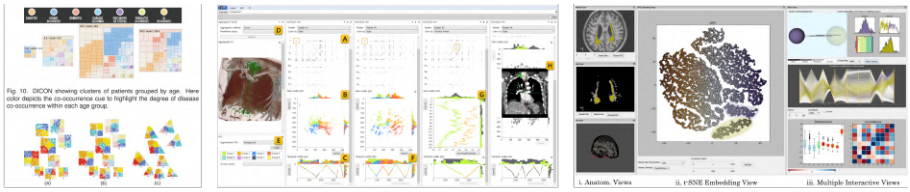


Figure 4.4: Examples of interaction with visualisations. **(a)** Connect a hovered cell to related cells with highlighting (Cao et al., 2011); **(b)** Explore 3D plots with scrolling (Song et al., 2017); **(c)** Select data in a scatter plot and update other visualisations (Raidou et al., 2016a).

network connections above a correlation threshold (Xing et al., 2014), filtering results that are statistically significant (Jönsson et al., 2019, 2020) (Figure 4.3b), and adjusting the range of attributes in parallel coordinates by brushing axes, e.g., (Yu et al., 2017b), or manipulating sliders on the axes, e.g., (Jönsson et al., 2019; Santamaría et al., 2008).

■ *Reconfigure* interactions change the spatial arrangement of data items in at least three ways. (1) Sorting matrices, lists and tables can reveal patterns, e.g., (Dang et al., 2015), and anomalies like violations in diagnostic rules for breast cancer (Kovalerchuk et al., 2012). (2) Changing attributes. Axes in scatter plots can be configured with buttons or drop-down menus to represent different attributes, e.g., (Abdullah et al., 2020; Krause et al., 2014; Stolper et al., 2014) (Figure 4.3c). Checkboxes can change available attributes in parallel sets, e.g., (Abdullah et al., 2020), or line graphs, e.g., (Fang et al., 2017). (3) Repositioning data points manually can reduce occlusion (L'Yi et al., 2017; Verma et al., 2017), and automatic repositioning according to a chosen similarity metric can group similar data points (Brunker et al., 2019; Ji et al., 2019a).

■ *Connect* interactions highlight associations between data items, and are often triggered by hovering. Hovering one part of a visualisation can highlight other parts in the same visualisation, for example connected ribbons and sets in parallel sets, e.g., (Abdullah et al., 2020), neighbours in a network (Brunker et al., 2019; Li et al., 2020; L'Yi et al., 2017), or other features from the hovered data point (Cao et al., 2011; Gotz et al., 2011) (Figure 4.4a). Hovering can also highlight related entities across multiple visualisations, e.g., (Kakar et al., 2019; Kumar et al., 2015).

■ *Explore* interactions bring new data subsets into view. 2D visualisations can often be panned, e.g., (Afzal et al., 2011; Behrisch et al., 2018; L'Yi et al., 2017). In 3D, images can be rotated with sliders or buttons, e.g., (Abbasloo et al.,

2019; Jönsson et al., 2019), or the y-direction can be explored by scrolling up and down the x-z plane (Song et al., 2017) (Figure 4.4b). Other exploration interactions are: picking different clustering results for visualisation (Kwon et al., 2018), and drilling down a Bayesian network by clicking nodes (Müller et al., 2020).

■ *Encode* interactions alter the visual representation of data, which may concern the overall visualisation or the colour encoding. First, some visual analytics systems switch between entirely different visualisation types, e.g., (Behrisch et al., 2018; Borland et al., 2020; Krause et al., 2014). Others extend a visualisation, for example by adding colour-coded links to a similarity scatter plot (Brunker et al., 2019), or spikes to dots in a scatter plot to represent all data dimensions (Hund et al., 2016) (Figure 4.3d). Second, recolouring nodes in scatter plots or networks can reveal similarity to a specific node (Ji et al., 2019b); association strength with a particular entity (Xing et al., 2014); and cluster specifics like compactness, size and distribution (Hund et al., 2016).

■ *Select* interactions mark data items, either manually through brushing, e.g., (Guo et al., 2018), lasso selection, e.g., (Kwon et al., 2019; Raidou et al., 2016a) (Figure 4.3e), clicking on a legend (Guo et al., 2020), or pinning (Dingen et al., 2019; Kakar et al., 2019); or automatically based on a chosen metric, e.g., (Stolper et al., 2014). Selected data are typically coloured prominently to easily focus on them. For example, Herold et al. (2010) and Hinterberg et al. (2015) respectively highlight cells on fluorescence micrographs and significant phenotype-gene expression associations that match set thresholds.

Select often precedes other interactions, hence its prominence. Geurts et al. (2015) compare the quality of several segmentation algorithms for selected segments (*Abstract/elaborate*). Liao et al. (2017) only show selected items in a radar map (*Filter*). Lamy and Tsopra (2019) reposition selected rainbow boxes, and Nguyen et al. (2012, 2011) reposition points in a scatter plot by similarity to a selected point (*Reconfigure*). Upon selecting multiple clustering results, Kwon et al. (2018) convert nodes in a scatter plot into small pie charts that reflect to which clusters the nodes belong in the different clusterings (*Encode*). Lastly, selecting data points can connect them to related data (Stolper et al., 2014) (Figure 4.3c), or update other visualisations, e.g., (Hund et al., 2016; Klemm et al., 2014; Raidou et al., 2016a, 2015) (Figure 4.3d, Figure 4.4c, *Connect*).

To conclude, visual analytics systems support many interaction types that often co-occur. These interactions facilitate insights in algorithmic outcomes, which can in turn strengthen a user's mental model of how an advanced algorithm works.

4.6 Shepherding Algorithms With Visual Analytics

So far, we have covered two methods to gain insights in advanced algorithms: visualising their outcomes, and interacting with the visualisations. A special interaction type is ■ *Shepherding*: guiding or controlling the algorithmic process to show algorithms' behaviour under different settings (cfr. RQ3). Shepherding bridges interactive visualisations and direct explanations, because it is an example of what Mohseni et al. (2021) call "what-if explanations". Table 4.2 shows that less than half (32/71) of the visual analytics systems in our sample allows shepherding. This section groups those systems by their "level of integration" (Turkay et al., 2014), which indicates how seamlessly they integrate algorithms. Figure 4.5 shows that we found examples in the full spectrum between level two (semi-interactive) and three (tight integration). Recall that our review excluded level one systems (static visualisations or limited interaction).

Visual analytics systems of integration level two can only modify parameters or the data domain through menus or pop-up windows that obscure the visualisation, e.g., (Brunker et al., 2019; Santamaría et al., 2019). An illustrative example is (L'Yi et al., 2017) (Figure 4.5a), where users can configure prediction models for miRNA-mRNA interaction in a tab completely separate from the visualisation. This approach hinders swift shepherding as users constantly need to switch between configuration and visualisation.

To better integrate the algorithm configuration, visual analytics systems can fix settings panels along the visualisation, or use pop-ups that minimally obscure the visualisation. Through radio buttons, checkboxes, text fields or sliders, users configure algorithms and rerun them by pressing a button. Examples of modifiable aspects are: the parameter k in k-nearest neighbours clustering (Spitz et al., 2020), the number of clusters (Ji et al., 2019b) (Figure 4.5b), the applied algorithm (Riegler et al., 2016), the attributes for analysis (Abbasloo et al., 2019; Dixit et al., 2017; Zhao et al., 2017), the query for pattern mining (Gotz et al., 2014) (Figure 4.2e), and the feature weights that best distinguish ill and healthy people (Moschonas et al., 2016).

Visual analytics systems with a settings panel can integrate algorithms more tightly by rerunning them automatically after reconfiguration. For example, Clark et al. (2017) rerun statistical tests for drug dose-response analysis after altering features or the tests' sidedness, Barlowe et al. (2013) rerank histograms after modifying the number of bins, Feller et al. (2018) reapply clustering after changing the number of clusters, Jeong et al. (2009) (Figure 4.5c) and Ji et al. (2017) change the contribution of dimensions in weighted PCA, and (Guo et al., 2018) adjust the clustering level of medical event sequences to find meaningful

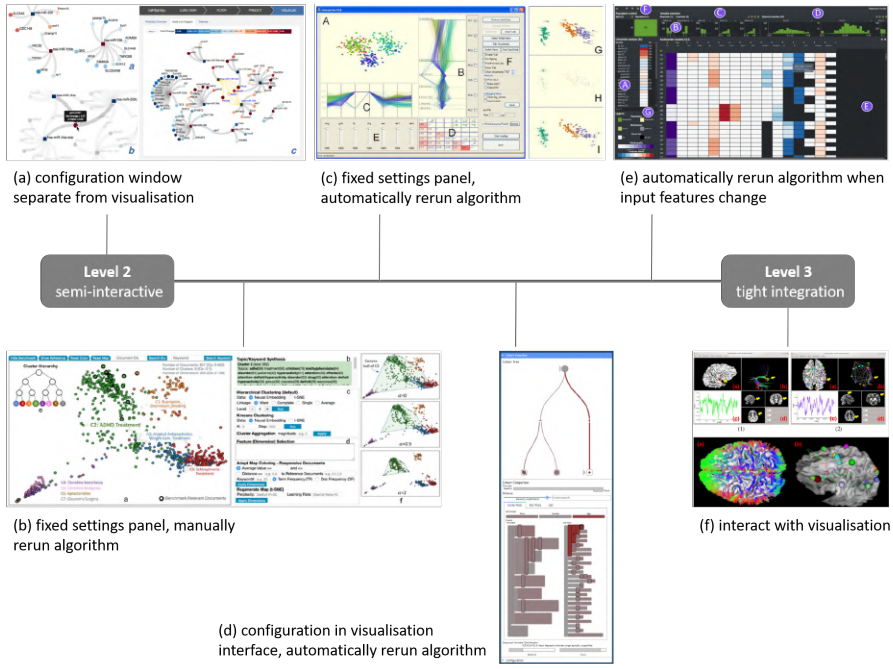


Figure 4.5: Examples of shepherding, ordered by the level of integration in the visual analytics system (Turkay et al., 2014). Semi-interactive examples are situated left; the more to the right, the tighter the integration. **(a)** L'Yi et al. (2017), **(b)** Ji et al. (2019b), **(c)** Jeong et al. (2009), **(d)** Borland et al. (2020), **(e)** Dingen et al. (2019), **(f)** Li et al. (2012).

groupings and to understand their sensitivity.

To further integrate algorithms into visual analytics systems, visualisations can themselves incorporate configuration functionalities like sliders to adjust how aggressively sequential events are grouped (Borland et al., 2020; Gotz et al., 2020) (Figure 4.5d), drop-down menus to configure dimension reduction techniques (Abdullah et al., 2020; Kwon et al., 2018), and textfields to set the maximal cohort size for cohort clustering (Huang et al., 2015).

Four visual analytics systems approach the highest integration level: they automatically update algorithmic outcomes when input changes. First, Li et al. (2020) show predicted risk of heart failure in a line chart, and add a line for the updated risk after removing or adding drugs. (Kwon et al., 2019) is similar, though its predictions need to be rerun manually. Next, Afzal et al. (2011) compare diseases' mortality and infection rates under different epidemiological

Table 4.3: Classification of ten visual analytics systems with direct explanations in our sample according to explanation type, explanation scale, explanator and target user.

	Type		Scale			Explanator								User		
	Model-agnostic	Model-specific	Global	Local	Model inspection	Activation maximisation	Contextual decomposition	Decision rules	Decision tree	Feature importance	Partial dependence plot	Sensitivity analysis	Visualisation	AI novices	Data experts	AI experts
Alsaad et al. (2019)		•		•			•								•	•
Hur et al. (2020)		•			•	•										•
Krause et al. (2014)		•		•						•					•	•
Krause et al. (2016)	•			•	•						•				•	•
Krause et al. (2018a)		•	•	•				•		•					•	•
Kwon et al. (2018)		•	•							•					•	•
Lamy and Tsopra (2019)		•	•										•	•	•	•
Müller et al. (2020)		•		•	•					•		•			•	•
Nauta et al. (2020)		•	•						•						•	•
von Landesberger et al. (2013)		•		•									•			•
Total	1	9	4	6	3	1	1	1	1	4	1	1	2	1	8	10

parameters and measures. Last, Dingen et al. (2019) (Figure 4.5e) allow to drag variables into a dedicated panel to automatically generate logistic regression models, and compare those models across groups.

Finally, visual analytics systems of the third integration level shepherd algorithms through direct interaction with visualisations. First, Li et al. (2012) (Figure 4.5f) allow to relocate and edit regions of interest in a 3D image of the brain, after which the strength of connections between them is recomputed. Second, Hund et al. (2016) (Figure 4.3d) project clustering results under different distance measures, and rerun the algorithms when users filter the data. Third, Gotz et al. (2011) and Cao et al. (2011) (Figure 4.4a) provide rich interactions to refine clusters of patients: users can filter features, merge clusters with lasso selection, drag patients out of clusters, and automatically remove patients far from the cluster centre.

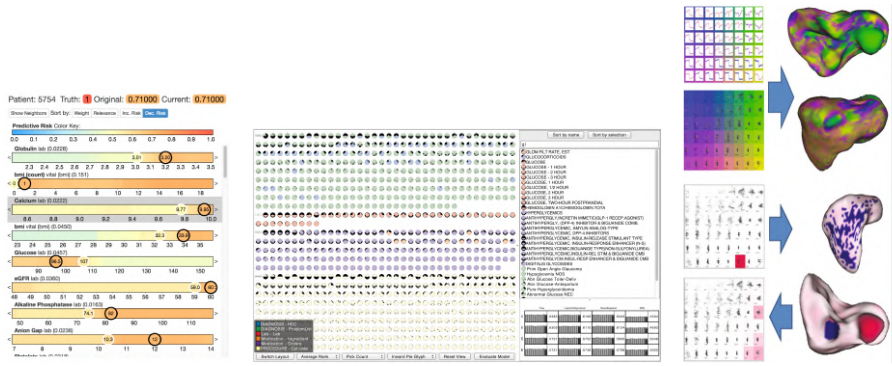
4.7 Directly Explaining Algorithms With Visual Analytics

Besides visualising and interacting with algorithmic outcomes, a final technique to better understand advanced algorithms is to directly explain how they work (cfr. RQ4). Our sample includes ten visual analytics systems with explicit explanations: Table 4.3 classifies them by type, scale, explanator, and target user.

Type All but one explanation techniques are model-specific. Some of them can be applied to other algorithms of the same family (e.g., all feature selection algorithms), but only *Prospector* (Krause et al., 2016) (Figure 4.6a) is fully model-agnostic. *Prospector* shows how a prediction is affected when feature values are perturbed with colour-coded sliders that correspond to partial dependence plots.

Scale Zooming in on explanations' scale, our sample mainly contains local explanations (6/10). Three systems explain artificial neural networks (ANNs) on a different scale. (1) Nauta et al. (2020) globally explain how an ANN predicts coma outcome: for a fixed epoch and a fixed hidden layer, all input activations are projected onto a scatter plot, and users can then select clusters to train a decision tree that distinguishes them. (2) Alsaad et al. (2019) use contextual decomposition to locally explain how a long short-term memory (LSTM) network predicts asthma based on clinical visits: each visit's contribution to the prediction is visualised in a heat map matrix that also highlights the most predictive subset of visits. (3) Hur et al. (2020) apply model inspection to explain a LSTM that predicts heart failure or heart surgery based on medical pathways: for the average or a specific patient, an attention heat map shows the variable weights in all time steps of the LSTM.

Explanator Four visual analytics systems in our sample use feature importance for explanations: they assign scores to features to indicate how strongly they impact the algorithmic outcome. (1) Müller et al. (2020) predict suitable cancer treatments with a Bayesian network, and determine the global and local relevance of evidence with sensitivity analysis. Users can also see the impact of updating or adding evidence in donut charts. (2) To select features that best predict diabetes, Krause et al. (2014) (Figure 4.6b) score feature importance with information gain, Fisher score, odds ratio, and relative risk. These four scores are computed in ten cross-validation folds, and visualised as quadrants of a circular bar chart. (3) Krause et al. (2018a) predict hospital admission with binary classifiers, and explain them with decision rules, which consist of feature sets that change the prediction outcome when removed. Users can inspect these decision rules in a matrix of data items (rows) and features



(a) (b) (c)

Figure 4.6: Examples of direct explanations in visual analytics systems. **(a)** Sliders show how perturbed inputs impact the prediction (Krause et al., 2016); **(b)** Circular glyphs with quadrants of feature importance scores (Krause et al., 2014); **(c)** Segmented mesh colour-coded by similar landmark movements (von Landesberger et al., 2013).

(columns, ordered by gini feature importance). (4) After clustering, Kwon et al. (2018) apply ANOVA to identify significant relationships between features and clusters, and use the F-values as importance values to rank features.

User Most of the collected explanation methods (8/10) are designed for both data experts and AI experts. von Landesberger et al. (2013) (Figure 4.6c) specifically target AI experts, who can detect drawbacks of segmentation algorithms through visualisations of landmark movements. Lamy and Tsopra (2019) also explain through visualisation, which seems the only approach suitable for AI novices: they visualise ANNs without hidden layers as rainbow boxes. These boxes symbolise connections in the ANN, and their height equals the weight of those connections.

4.8 Observations, Opportunities and Challenges

Our review investigated four methods to obtain insights in advanced algorithms: (1) visualising their outcomes and (2) interacting with these visualisations, (3) shepherding them, and (4) directly explaining them. These methods are not always clearly separable as shepherding can be a kind of interaction with visualisations, and visualisations can also act as direct explanations. Put differently, there is a fine line between getting insights in algorithmic outcomes

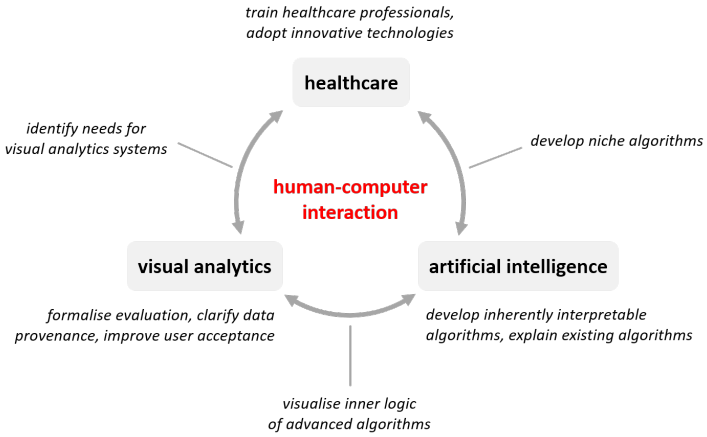


Figure 4.7: Opportunities and challenges for visual analytics, artificial intelligence, and healthcare. Human-computer interaction can mediate the dialogue between these three communities.

and in the inner logic of algorithms themselves. This section answers our research questions, and positions them in the broader, multidisciplinary context of Figure 4.7.

4.8.1 Visualising Outcomes: Many Algorithm-Dependent Possibilities

Section 4.4 showed that algorithmic outcomes can be visualised in many ways (RQ1), ranging from basic scatter plots to original custom glyphs. Some visualisations often occur together with specific algorithms, for example dendrograms with hierarchical clustering, and projection plots with dimension reduction. This co-occurrence seems to hold in general: visualisation approaches strongly depend on the algorithmic family and the healthcare activity. We observed that visual analytics systems for a healthcare context mainly rely on mainstream interpretative algorithms such as clustering, dimension reduction and classical statistics, resulting in few visual analytics systems for predictive activities, and even none for monitoring chronic diseases such as diabetes. On the one hand, this suggests room for adopting more specialised state-of-the-art AI techniques such as transformer neural networks, Bayesian networks, and natural language processing. On the other hand, the absence of monitoring systems in our sample suggests a gap for further research. This gap is related to a second observation: most of the visual analytics systems in our sample

target healthcare professionals, but as technology increasingly facilitates self-monitoring, laypeople are likely to become an important target group too. Future studies could thus investigate how visual analytics may fit the needs of laypeople.

4.8.2 Interacting With Visualisations: Sufficient or Too Much?

Section 4.5 demonstrated that existing visual analytics systems incorporate all seven interaction types from Yi et al. (2007) (RQ2): *Abstract/elaborate*, *Connect*, *Encode*, *Explore*, *Filter*, *Reconfigure* and *Select*. The frequent appearance of *Abstract/elaborate* presumably originates from the widespread details-on-demand mantra in information visualisation (Shneiderman, 2003). All interaction types exist in different forms, and are often combined in highly interactive visualisations. While this is encouraged to facilitate insights in algorithmic outcomes, some developers of visual analytics systems in our sample, e.g., (Kwon et al., 2019), intensively collaborated with healthcare professionals and note that caution is needed: end-users are not always looking for highly exploratory, information-heavy interfaces that are interesting from a visualisation perspective, but are too complex for their needs. Instead, some healthcare contexts may require visual analytics systems that act as fellow healthcare experts and point out interesting cases in the data. Thus, tailoring the amount of interaction in visual analytics systems is part of the broader challenge to involve end-users in every stage of the design process and identify their needs. This may improve user acceptance and enhance trust in the proposed system (Abdullah et al., 2020).

4.8.3 Shepherding Algorithms: A Higher-Order Interaction

Section 4.6 showed that algorithms can be shepherded by tuning algorithmic parameters or modifying input (RQ3). Visual analytics systems can integrate shepherding along a continuum that ranges from separating shepherding and visualisations to tightly connecting both. Despite its potential to provide what-if explanations, optimise existing models, or build new meaningful models, shepherding was relatively uncommon in our sample. This could be due to a trade-off that rises when the amount of transparency and shepherding freedom is determined: giving too much control to non-trained users can overwhelm them or incur misleading outcomes resulting from overfit models. Blindly playing around with parameters may be harmless for exploratory contexts such as clustering similar documents (Ji et al., 2019b) or medical images (Riegler et al., 2016), but can have severe consequences in delicate predictive contexts. Therefore, AI

experts should inform healthcare professionals about the applicability, strengths and weaknesses of AI models, and visual analytics developers should help in training healthcare professionals to use their systems. User-studies in our sample endorse the need for such training, e.g., (Cao et al., 2011; Gotz et al., 2020; Kwon et al., 2021). Another possibility is to develop different versions of visual analytics systems and only provide shepherding functionalities in specific contexts.

4.8.4 Direct Explanations: Rare Yet Promising

Section 4.7 revealed two interesting observations about direct explanations (RQ4). First, few examples were present in our sample, yet all of them involved visualisations. Second, AI novices are so far often neglected. Although directly explaining advanced algorithms to AI novices seems challenging, future research could for example investigate whether conversational design and example-based explanations give solace (Ribera and Lapedriza, 2019). In that way, patients may better understand personalised health plans proposed by their clinician, potentially by comparing their health measurements against similar patients. Of course, AI experts remain an important target group as well: direct explanations in interactive visual analytics systems can help them explain existing black-box models, obtain information about the inner logic of advanced algorithms, and design algorithms that are better interpretable. Regarding the latter, collaborating with healthcare professionals is essential to learn what interpretability means in their context.

4.9 Conclusion

XAI is extremely relevant in our current age of algorithms and massive data collection, and visual analytics has proven to play an important role in the quest for explainability. Healthcare has acknowledged the value of visual analytics in many applications, but may not have taken enough advantage of this exciting field yet (West et al., 2015). For example, Bonnett et al. (2019) showed that risk prediction is still dominated by simple static visualisations like point score systems and nomograms. In addition, our review suggests a lack of interactive visual analytics systems for monitoring, and systems that target laypeople.

Our review also reveals that visual analytics holds many opportunities for XAI in healthcare by providing insights in advanced algorithms through visualisation, interaction, shepherding, and direct explanations. Yet, complex challenges remain: many healthcare stakeholders are involved in the visual analytics

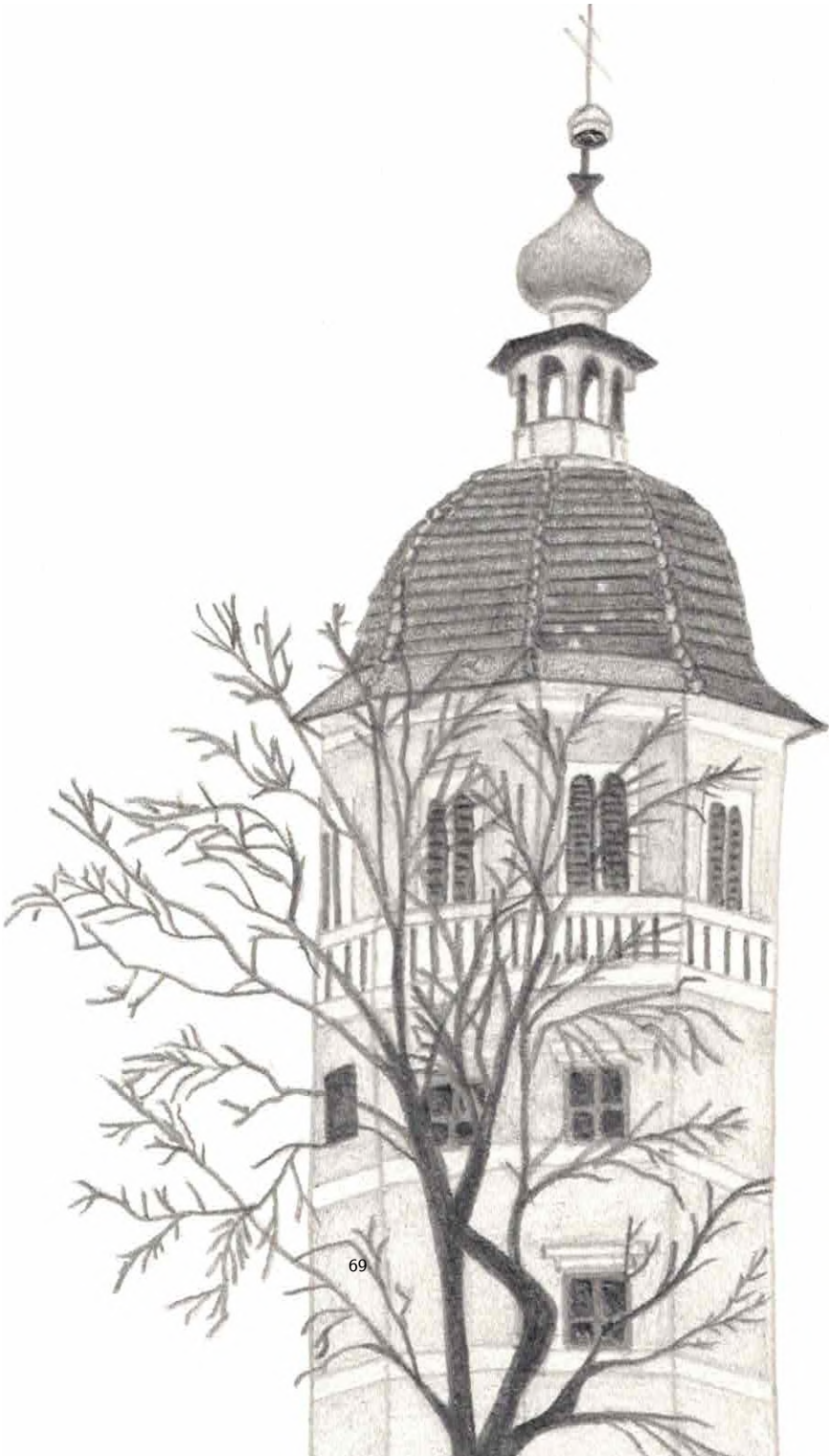
process (Kolyshkina and Simoff, 2021), domain practices should be respected, domain expertise is often required to correctly interpret algorithmic results, and explanation techniques should be tailored to the application domain and target users. These XAI challenges cannot be solved in isolation, so we encourage the visual analytics, AI and healthcare communities to further reach out to each other, and we invite the human-computer interaction community to help mediating this fascinating interdisciplinary dialogue.

Acknowledgements

This work was supported by the Research Foundation-Flanders (FWO) grant G0A3319N and the Slovenian Research Agency grant ARRS-N2-0101. Thanks to Martijn Millecamp and Robin De Croon for providing helpful feedback on earlier versions of the text.

The Human Side of Chapter 4

The Three Towers



I visited Gregor's lab in Maribor (Slovenia) with Katrien and Robin and presented a preliminary literature review on visual analytics and XAI for healthcare, which became the foundation for this chapter. I remember everyone's enthusiasm and how I secretly felt like an absolute fraud who didn't deserve the warm welcome. After the research visit, Robin and I travelled to Graz (Austria), where we spent several days walking for miles. One day, we climbed the Schlossberg to visit the famous Uhrturm tower. Compared to the spectacular view on the city, I found the Uhrturm slightly underwhelming, but I was intrigued by how a nearby clock tower was trying to hide in plain sight. Never thought I could feel like a tower.

Songs on repeat:

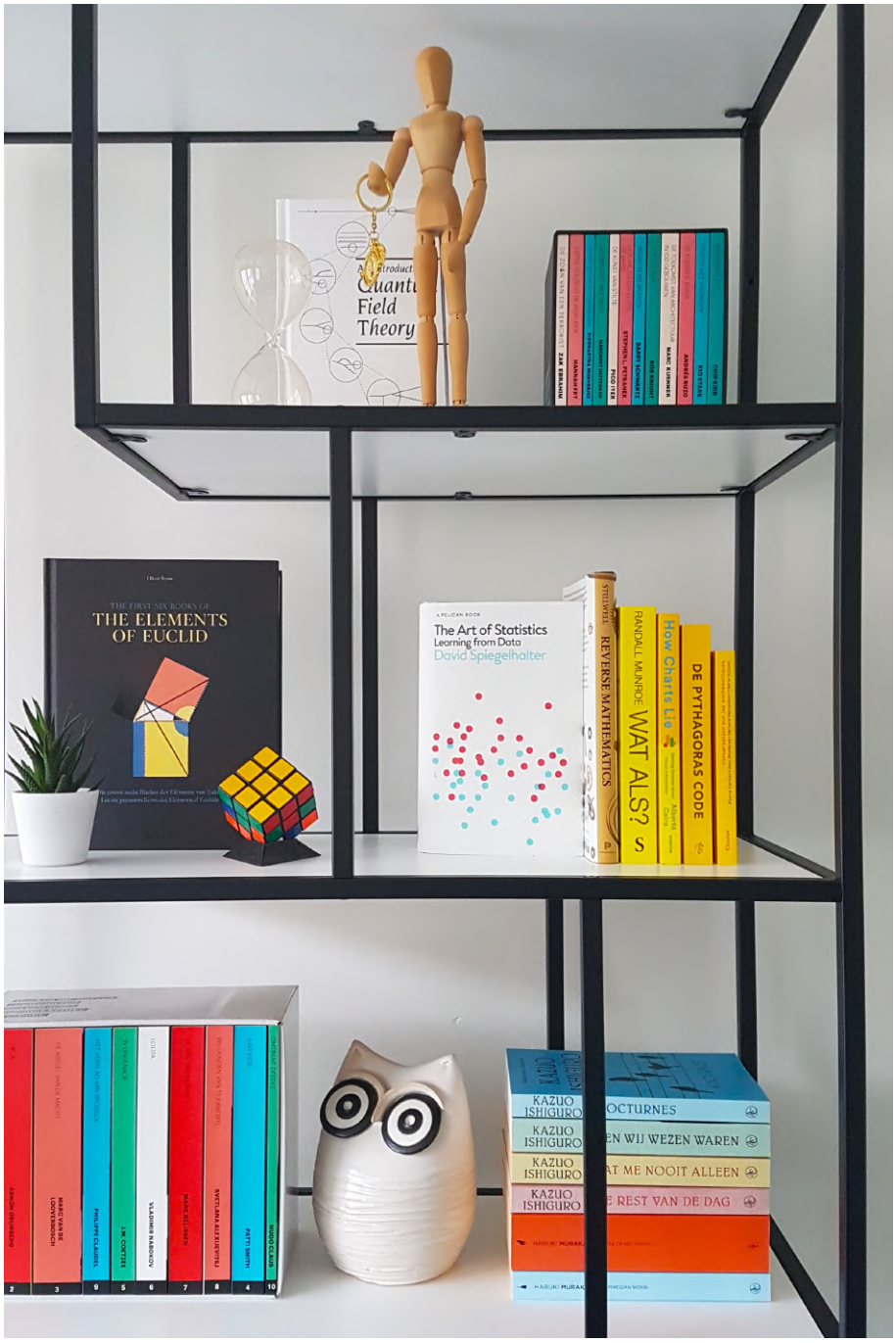
- *Underdog* by BANKS
- *Alibi* and the rest of the *Goddess (Deluxe)* album by BANKS
- *Doin' Time* by Lana Del Rey

Clock tower in Graz – November 2019

In December 2019, Oscar gifted me *The Art of Statistics: Learning from Data* by David Spiegelhalter during a Secret Santa dinner with the Augment lab in which I worked. His kind words and everyone's company made me feel accepted by an incredible international team. I was happy and could not have wished for a better team at the start of my PhD. Until today, Oscar's book has a prominent place in my living room, reminding me of that unforgettable evening. Oh, and I loved the book, by the way. Read it.

Songs on repeat:

- *Stroke* and the rest of the *III* album by BANKS
- *Gemini Feed* and the rest of the *The Altar* album by BANKS
- *Late Night Feelings* by Mark Ronson and Lykke Li



The Art of Statistics – October 2023

When I started collecting papers in November 2019, I was intimidated by the vast amount of existing work and other meticulously executed systematic reviews. Paralysed by the fear of missing relevant work and constructing a suboptimal classification, I fled into working on other papers (Ooge et al., 2020; Ooge and Verbert, 2021, 2022). In January 2021, I regained courage and screened about 2000 paper abstracts in a couple of days, after which I painstakingly started to screen the full papers and classify them in a ginormous Excel sheet (118 columns). Driven by the rush to finish as many papers as possible each day, I worked late hours at my studio desk with frontal view on the imec tower in Heverlee. Every evening, I watched the tower light up, its office lights painting abstract patterns in the night sky. I still wonder whether the office in the left upper corner, typically lit until 3 am, was also housed by a crazy workaholic.

Songs on repeat:

- *Love Hangover* by Diana Ross
- *Everything i wanted* by Billie Eilish

imec tower – January 2021

In the spring of 2021, I was finishing the categorisation of papers and started writing the actual review. Tired of staring at the imec tower and working from home under the Covid restrictions, I started working outside. My favourite spot was a bench in front of the Arenberg castle in Heverlee, next to what I baptised the "frog pool." I think the Hogwarts vibes that the castle was giving me made me believe that something magical would bring together my categorisation and notes into a coherent review. The croaking frogs in the deliciously warm sun brought me into the perfect trance for making that magic happen. By June 2021, I had been writing while sneezing my brains out in the high grass and doing my laundry at the laundry centre. But whenever I see Table 4.2, I can't help hearing "croak"s and "ribbit"s.

Songs on repeat:

- *Technicolour* by Montaigne
- *Thunder* by Catnapp
- *Hello?* by Clairo

Arenberg castle in Heverlee – May 2021

Chapter 5

Visually Explaining Uncertain Price Predictions in Agrifood

The rise of 'big data' in agrifood has increased the need for decision support systems that harvest the power of artificial intelligence. While many such systems have been proposed, their uptake is limited, for example because they often lack uncertainty representations and are rarely designed in a user-centred way. We present a prototypical visual decision support system that incorporates price prediction, uncertainty, and visual analytics techniques. We evaluated our prototype with 10 participants who are active in different parts of agrifood. Through semi-structured interviews and questionnaires, we collected quantitative and qualitative data about four metrics: usability, usefulness and needs, model understanding, and trust. Our results reveal that the first three metrics can directly and indirectly affect appropriate trust, and that perception differences exist between people with diverging experience levels in predictive modelling. Overall, this suggests that user-centred approaches are key for increasing uptake of visual decision support systems in agrifood.

5.1 Introduction

Under the impulse of success stories in other domains, artificial intelligence and 'big data' are on the rise in agrifood (Kamilaris et al., 2017), leading to promising research directions such as *Agriculture 4.0* (Zhai et al., 2020) and the broader *Agrifood 4.0* (Lezoche et al., 2020), *precision agriculture* (Cisternas et al., 2020; Linaza et al., 2021; Wachowiak et al., 2017), and *smart*

farming (Ayoub Shaikh et al., 2022; Moysiadis et al., 2021; Wolfert et al., 2017). While the adoption of such technologies is still modest in real-life agrifood applications (Osinga et al., 2022), it is expected that the wide availability of cloud computing and remote sensing (Navarro et al., 2020) will further boost their spread (Liakos et al., 2018). To process the explosive amount of information in this era of growing digitisation and to make data-grounded decisions, agrifood stakeholders increasingly need the assistance of *decision support systems* (DSSs) (Zhai et al., 2020) that facilitate learning and allow to modify decision processes by integrating domain knowledge, rather than systems that merely prescribe actions (McCown, 2002; Rojo et al., 2021).

Yet, even though the need for DSSs in agrifood has been acknowledged for over two decades (McCown, 2002) and many prototypes have been proposed (Gutiérrez et al., 2019a; Zhai et al., 2020), the uptake of these systems has been limited so far. Parker (1999); Parker and Campion (1997), Zhai et al. (2020), and Rose et al. (2016) discussed several reasons for this low uptake: user interfaces of DSSs are not always user-friendly and lack visualisations, DSSs are not necessarily relevant when they do not meet end users' needs or decision-making styles, outputs often miss uncertainty representations, and end users often distrust DSSs with opaque underlying algorithms. In other words, developers of DSSs for agrifood face important design challenges such as increasing usability, guarding usefulness for end users, and raising appropriate trust in underlying decision models.

Tackling these challenges requires human-centred approaches, which lie at the core of *human–computer interaction* (HCI), an interdisciplinary field that connects computer science, social sciences, and technology-applying domains such as agrifood. Specifically, HCI studies how interfaces can be designed and tailored to specific end users or application contexts to improve user experience, for example (Carroll, 1997; Olson and Olson, 2003; Shneiderman et al., 2016). Two subdomains of HCI specialise in visualising complex information and explaining artificial intelligence, respectively. The first subdomain, *visual analytics*, fosters analytical reasoning with visual dashboards that support advanced interaction and visual exploration to discover hidden patterns in data (Cui, 2019; Ham, 2010; Keim et al., 2008). The second subdomain, *explainable artificial intelligence* (XAI), seeks techniques that give insights into outcomes of artificial intelligence models, and studies interrelated topics such as trust, fairness, bias, causality, accountability, privacy, and reasoning (Abdul et al., 2018).

Visual analytics and XAI are relevant in agrifood because DSSs increasingly include predictive models and benefit from visualising information. Yet, current DSSs in agrifood often lack uncertainty representations and are rarely designed in a user-centred way (Rose et al., 2017). To enable informed decision-making

by different end users, researchers and practitioners have called for adopting more user-centred and HCI practices in agrifood (Lindblom et al., 2017; Parker and Sinclair, 2001; Rose et al., 2017).

We address this call by presenting a visual DSS that shows predicted food product prices and uncertainty in the predictions. We evaluated our prototype with 10 participants who are active in different parts of agrifood; collecting and analysing both qualitative and quantitative data. In particular, we focused on the following research questions:

RQ1 **Usability**: How user-friendly are the interaction functionalities and the visualisation in our visual DSS?
RQ2 **Usefulness and needs**: How useful is our visual DSS and how does it accommodate the needs of people active in agrifood?
RQ3 **Model understanding**: How does visualising uncertain predictions affect people's understanding of the prediction model underlying our visual DSS?
RQ4 **Trust**: How does visualising uncertain predictions affect people's trust in the prediction model underlying our visual DSS?

Our research contribution consists of extensively evaluating our visual DSS from two perspectives. First, considering our prototype as a product, we assessed its usability and usefulness. Sections 5.4.1 and 5.4.2 show that participants were generally very positive about our prototype's usability (RQ1) and expressed needs regarding control, comparison, and explanations (RQ2). Second, considering our prototype as an XAI research tool, we dived deeper into what affected participants' understanding of and trust in the prediction model underlying our DSS, and the relation with uncertainty visualisation. Sections 5.4.3 and 5.4.4 show that participants' understanding was affected on an algorithmic and an outcome level (RQ3), and that trust in the prediction model evolved under several factors (RQ4). In both perspectives, we considered the impact of participants' experience with predictive modelling, observing different responses for different experience levels. Finally, we made our prototypical visual DSS open-source so that the community can use it as a flexible basis for more advanced dashboards tailored to specific contexts.

5.2 Background and Related Work

To contextualise our research, we first discuss visualisation for DSSs and uncertainty representation. Then, we turn towards XAI and focus on trust.

5.2.1 Visualisation for Decision Support Systems

Visualising information augments people's abilities to get insights into complex data and more effectively fulfil tasks that cannot be automated (Munzner, 2014). Presenting decision-making information visually has also been found to make DSSs more user-friendly (Rose et al., 2016). Hence, it is no surprise that DSSs often incorporate visualisations to facilitate decision-making across application domains, e.g., healthcare (Botha et al., 2012; Rind, 2013; West et al., 2015), learning analytics (Verbert et al., 2013; Vieira et al., 2018), finance (Savikhin et al., 2011), and supply chain analytics (Basole et al., 2017; Khakpour et al., 2021). In many of these domains, decision-making is supported by visual analytics, which combines powerful visualisations with advanced interaction techniques (Yi et al., 2007) and automated data analysis. This allows people to iteratively generate and test hypotheses (Cui, 2019; Ham, 2010; Keim et al., 2008, 2010). In healthcare, for example, visual analytics has been applied to personalise medical treatments by analysing electronic health records, modelling diseases and medical prediction, optimising care pathways, and so on (Hu et al., 2016; Preim and Lawonn, 2020).

In agrifood, many visual DSSs have been proposed too, for example in dairy farming (Di Silvestro et al., 2014), crop control (Armstrong and Nallan, 2016; Machwitz et al., 2019), land assessment (Ochola and Kerkides, 2004), irrigation management (Accorsi et al., 2014), and climate monitoring (Jarvis et al., 2017). Yet, Gutiérrez et al. (2019a) found that most visual DSSs include maps, contain a single visualisation, and are intended for farmers to manage crops or assess land suitability. This suggests room for dashboards with multiple visualisations in other application areas such as livestock monitoring and sales. In addition, it suggests that current visual DSSs in agrifood are less advanced than visual analytics approaches in terms of varied visualisations and interaction possibilities.

5.2.2 Uncertainty Visualisation

Visual DSSs are subject to uncertainties in the data and uncertainties propagated during the data processing, modelling, and visualisation (Sacha et al., 2016; Skeels et al., 2010). These uncertainties can be visualised in many ways (Demmans Epp and Bull, 2015; Spiegelhalter et al., 2011), but there are two challenges. First, visualising uncertainty entails a trade-off: showing too much uncertainty may overload or confuse people, whereas showing too little uncertainty feigns accuracy and may mislead people (Sacha et al., 2016). Second, some approaches for uncertainty visualisation may be clearer or less misleading than others.

Tackling these challenges is hard, which unfortunately often results in simply omitting uncertainty (Franconeri et al., 2021; Hullman, 2020). This is currently the case in agriculture: visual DSSs rarely consider uncertainty (Gutiérrez et al., 2019a; Zhai et al., 2020). One exception, for example, is CropGIS (Machwitz et al., 2019), which predicts produced biomass of maize under different meteorological conditions. CropGIS then visualises the mean prediction in a line chart, together with the minimum, maximum, and 1σ-confidence interval, resembling a *fan chart* (Britton et al., 1998) with a single fan.

Researchers in information visualisation face the above two challenges by studying the pros and cons of different uncertainty visualisation techniques. For example, in the case of predicted time series, studies have shown that (a) similar to fan charts, uncertainty intervals around a prediction line are best distinguished with different opacity levels (Seipp et al., 2019); (b) fan charts are a good compromise between accuracy and uncertainty (Gutiérrez et al., 2019b); and (c) compared to ensemble charts, fan charts lead to higher acceptance of predictions (Leffrang and Müller, 2021).

5.2.3 Visualisation for Explainable Artificial Intelligence

As visual DSSs often incorporate complex algorithms, end users typically need explanations to understand the algorithmic decision-making, appropriately trust it, and detect potential biases (Gunning and Aha, 2019). There is no one-size-fits-all explanation, however. Human-centred XAI researchers therefore study how explanations can be effectively designed, considering factors such as the application context (Dhanorkar et al., 2021; Suresh et al., 2021; Vellido, 2020), human reasoning processes (Wang et al., 2019a), and end users' goals (Mohseni et al., 2021) or personal characteristics (Millecamp et al., 2019; Suresh et al., 2021).

XAI and visual analytics largely intersect. Visualisations can namely serve as explanations when people get visual insight in model outcomes and model behaviour, actively interact with them, and steer the underlying algorithms (Ooge et al., 2022b). Given the wide interest in visualisation for XAI, many surveys have discussed the state-of-the-art in visual analytics for machine learning (Endert et al., 2017; Liu et al., 2017), deep learning (Hohman et al., 2019b), predictive modelling (Lu et al., 2017), and enhancing trust in machine learning (Chatzimparmpas et al., 2020a) from different perspectives. A meta-analysis of all these surveys confirmed the key role of visualisation in interpreting machine learning (Chatzimparmpas et al., 2020b).

5.2.4 Trust in Intelligent Systems

Many application domains call for increasing end users' trust in algorithmic decision-making of DSSs, including agrifood (Gutiérrez et al., 2019a; Rose et al., 2016). In the scope of explaining black-box algorithms, trust is thus heavily studied in XAI and visual analytics. However, trust is a slippery concept for at least two reasons. First, there is no widely accepted definition for trust in intelligent systems, although many definitions have been proposed (Jacovi et al., 2021; Madsen and Gregor, 2000; Vereschak et al., 2021). Second, measuring trust is very challenging because it evolves (Holliday et al., 2016; Nourani et al., 2020; Ooge and Verbert, 2021) and is affected by many factors (Hoff and Bashir, 2015), for example, domain expertise (Nourani et al., 2020; Ooge and Verbert, 2021), visualised information and uncertainty (Mayr et al., 2019; Sacha et al., 2016), model accuracy (Papenmeier et al., 2022; Yin et al., 2019), and level of transparency (Kizilcec, 2016). In addition, there is growing consensus among XAI researchers that optimising trust is not always desirable; rather, the stress should lie on *appropriate* trust (Gunning and Aha, 2019) and trust calibration (Han and Schulz, 2020; Solhaug et al., 2007). Some researchers even argue that XAI research should move away from trust and focus on utility instead (Davis et al., 2020).

5.3 Materials and Methods

This section presents how we conducted our user-centred study. We first describe our visual DSS, study rationale, and overall study design. Then, we provide more details on how we measured usability, trust, and experience with predictive regression.

5.3.1 Visual Decision Support System

We developed a prototypical visual DSS for exploring product prices in various countries. Besides visualising historical price evolutions, our system visualises predicted future prices and the prediction model's uncertainty. Rather than building an advanced standalone interface with an accurate prediction model, we aimed to create a simple and flexible proof of concept for which the underlying dataset and prediction model could easily be replaced. To encourage future adaptations, we built our prototype with the open-source Meteor, React, and D3 frameworks, and made our code publicly available at `https://github.com/JeroenOoge/explaining-predictions-agrifood` (accessed on 9 July 2022).

In our proof of concept, the dataset contained price evolutions in European countries over the past 3 decades for over 400 food products, including fruits, vegetables, dairy, meat, and cereals. For each country separately, price predictions were generated by fitting a third-degree polynomial to the country's past price data with linear regression and least-squares estimation, extrapolating the fit for five years from the last known data point on. Uncertainty consisted of 55–99%-prediction intervals with increments of 5%.

Figure 5.1 shows our dashboard. At the top, two search fields with dropdown menus allow selecting a desired food product and countries available for that product. In the middle, the price evolution for selected countries is visualised in a line graph; each country is represented by a differently coloured full line. At the bottom, five checkboxes allow to enable or disable visual components: the first is enabled by default (*Past data*); the others are related to the prediction outcome and model (*Future prediction*, *Future uncertainty*, *Past fit*, and *Past uncertainty*). The future prediction and past fit are visualised as dashed lines, and the prediction intervals as stacked bands (i.e., fans), where larger intervals gradually become lighter. Finally, hovering over the chart and its visual components shows details-on-demand in the form of a tooltip with the exact price values or additional information.

5.3.2 Study Rationale

Adapting to economic uncertainty and predicting market fluctuations are important challenges in Agrifood 4.0 (Zhai et al., 2020). To meet these challenges, we framed our study in the context of predicting food product prices and built upon an earlier pilot study (Ooge and Verbert, 2021), which showed that four people experienced with predictive modelling had different trust evolutions while using our visual DSS. To investigate the transferability of our preliminary results, we recruited via email 10 end users who are active in agrifood or finance. Then, we evaluated our prototypical visual DSS according to four metrics: usability, usefulness and user needs, model understanding, and trust. With the former two, we considered our prototype as a *product*: we wanted to identify issues with the visualisation and the interaction possibilities and find out whether our prototype matches participants' needs. With the latter two, we considered our prototype as an *XAI research tool*: we set out to discover how the visual components in our visual DSS impact participants' understanding of the prediction model and what affects participants' trust in the model. For all four metrics, we also considered the effect of participants' profession and experience with predictive modelling.

In addition, we were interested in whether our visual DSS would allow

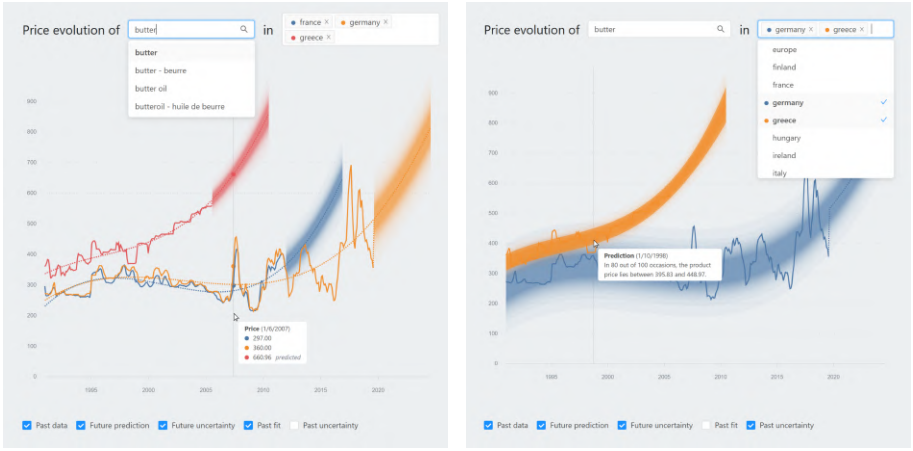


Figure 5.1: Screenshots of our responsive visual DSS during interaction. **Left**: selecting a food product in the upper left search field and getting details about the price and date upon hovering over the line chart. **Right**: selecting countries in the upper right search field and getting a description of the hovered fan (*"In 80 out of 100 occasions, the product price lies between A and B"*. where A and B are the lower and upper bounds of the prediction interval at the indicated date, respectively).

participants to identify the limitations of our simple prediction model. We assumed that obvious prediction failures, for example, an almost flat regression line for clearly periodic price evolutions, would not evoke lively discussions. Therefore, we deliberately built our study around a specific case of butter prices in France (data available for 1991–2011) and the Netherlands (data available for 1991–2019), with two not too obvious shortcomings. First, the model fit the past data rather poorly (high RMSEA). Second, even though France and the Netherlands had historically similar prices, the prediction for France largely diverged from the real data in the Netherlands, suggesting poor prediction performance.

5.3.3 Study Design

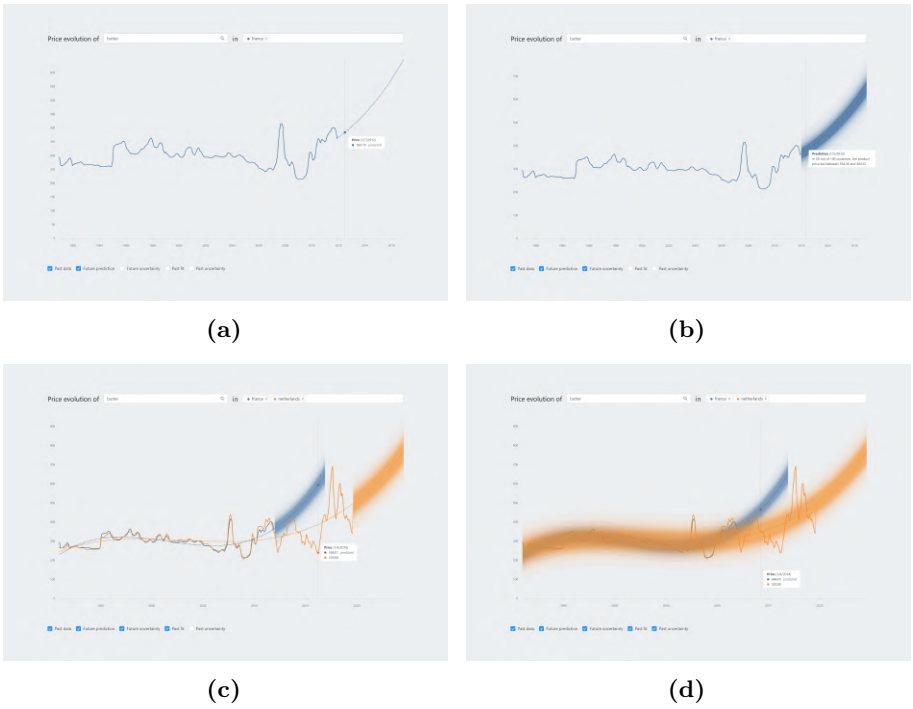
In July–October 2020, we collected qualitative data on our four evaluation metrics with online semi-structured interviews, quantitative data from Likert-type questions on trust, and observational data on how participants interacted with our visual DSS (participants shared their screen during the study). Figure 5.2 shows the overall structure of our study.

Figure 5.2: The flow of our study, including 5 phases: an introduction, four scenarios with one country, four scenarios with two countries, a questionnaire, and additional questions.

First, participants introduced themselves and we familiarised them with our visual DSS: we explained how they could compare past butter prices in France and the Netherlands and see details-on-demand in the visualisation, and we introduced the price prediction functionality without revealing details about the underlying prediction model.

Next, participants went through eight scenarios, enabling the *Future prediction*, *Future uncertainty*, *Past fit*, and *Past uncertainty* checkboxes one by one, first for a setting with one country (France; Scenarios 1–4) and then for a setting with two countries (France and the Netherlands; Scenarios 5–8). Figure 5.3 shows some representative screenshots. Each scenario consisted of three phases: (1) we asked participants to explore the visualisation while thinking out loud (*Explore the new component in the visualisation. Explain what you see. What grabs your attention?*); (2) we asked them about their **trust** and **model understanding** (*Do you trust the prediction model? Do you understand how the prediction model works? Which parts of the visualisation made you say that?*); and (3) we quantitatively measured their trust.

Finally, after completing all scenarios, participants reported their experience with four concepts related to predictive modelling and answered additional questions about **model understanding** and **usefulness** (*Which combination(s) of components do you find most useful to get insights into the prediction*



(a) **(b)**

(c) **(d)**

Figure 5.3: Our visual DSS with different sets of enabled visual components. **(a)** Scenario 1: the future prediction for France is visualised as a dashed line. **(b)** Scenario 2: the future uncertainty for France is visualised as fans. **(c)** Scenario 7: the past fit for France and the Netherlands is visualised as dashed lines. **(d)** Scenario 8: the past uncertainty for France and the Netherlands is visualised as fans.

model? Would you like to investigate or explore other things to get insights into the prediction model? Would you use this visualisation for your job activities?). In the post-study discussion, we asked participants how they experienced the study and stressed that our prediction model was not meant for making real-life decisions.

5.3.4 Measurement Instruments and Qualitative Analysis

To assess **usability**, we observed participants' interactions with our visual DSS and analysed their think-aloud feedback during exploration. As such, we could study whether participants easily found the information they were looking

for; understood filtering, clicking and hovering functionalities; and had further suggestions. In contrast to Likert scales for overall usability (Bangor et al., 2008; Brooke, 1996), this approach gives concrete insights into how, why, and which parts of visualisations should be adapted to improve usability.

To quantitatively measure **trust** in each scenario, we averaged responses to four Likert-type questions rated on a 7-point range (0–not at all to 6–extremely). These questions were inspired by a widely-used scale for trust in automated systems by Jian et al. (2000). Yet, as we considered it unfeasible for participants to answer all 12 items in this scale 8 times, we selected and adapted the 4 items that seemed most relevant for prediction models:

1. *I am suspicious of the prediction model's outputs* (reverse-scored);
2. *I am confident in the prediction model*;
3. *I can trust the prediction model*;
4. *The prediction model is deceptive* (reverse-scored).

To measure participants' experience with predictive regression, we combined self-reported data and indirect experience indicators. First, participants self-reported their experience with the concepts *prediction interval*, *linear regression*, and *time series prediction* through checkboxes *I know the word* (K), *I often use it* (U) and *I can explain it* (E). For each concept, we assigned a score between 0 (very inexperienced) and 5 (very experienced) based on their answers ($K = 1$, K & $U = 3$, K & $E = 4$, K & U & $E = 5$); the average E_s served as a final estimate for self-reported experience. Second, we scored participants' experience between 0 and 5 based on their background (E_b) and use of jargon related to statistics or predictive modelling during the interview (E_j). Then, we used the average of E_s, E_b and E_j as an estimate for experience with predictive regression.

Finally, to qualitatively analyse participants' feedback, we recorded the interviews, which lasted 70–130 min, depending on the amount of feedback. We then thematically analysed 120 pages of transcription, following the 6 phases from Braun and Clarke (Braun and Clarke, 2012). Specifically, we first coded our data deductively (i.e., starting from our four metrics) and then inductively for each metric (i.e., driven by the data instead of preset topics). To guard the originality of participants' feedback and respect participants' efforts to speak English, we only corrected language mistakes in quotes below when clarification was needed.

5.4 Results

This section presents the findings of our study with 10 participants whose specifics are shown in Table 5.1. First, approaching our visual DSS as a product, we focus on usability and usefulness. Then, taking an XAI research perspective, we turn towards model understanding and trust. Throughout, as summarised in Table 5.2, we also highlight differences between participants who have low, medium, and high experience with predictive regression.

5.4.1 Usability

Our semi-structured interviews brought up four themes on usability: *Understanding the visualisation*, *Visual encoding of information*, *Interacting with the visualisation*, and *Workflow*.

> **Understanding the visualisation**: most participants understood the overall goal, but some visual components need clarification.

Overall, participants were very positive about the visualisation and understood its main goal. For example, P4 found the visualisation *"very readable"* and complimented it for being a *"very simple instrument"* with a clear aim; P5 described the visualisation as *"very easy, simple, clear, and [without] any frills"*; and P8 stated: *"The dashboard I like. It's very simple and easy to use, so it's not too complex or anything like this. [...] It's just easy to use, gives you all the information [...] in a very sort of simple way"*. Most participants understood the visual components sufficiently and could use them without further clarification.

Specifically, participants described the future uncertainty fans as *"area[s] in which the price is statistically expected"* (P1), which *"shows the spread of [...] the predicted values around the [prediction] line"* (P9). In more economical terms, P5 talked about *"buffer points, which [indicate] the minimum and maximum of the variation of the future price"* and considered the fans' percentages to be *"the likelihood to be in these buffers"*. Many participants furthermore observed that uncertainty fans enlarge for larger percentages, entailing a trade-off between precision and correctness: *"[If you restrict a 90%-fan to a 50%-fan, then] you have more accuracy but you don't have a good prediction"*.

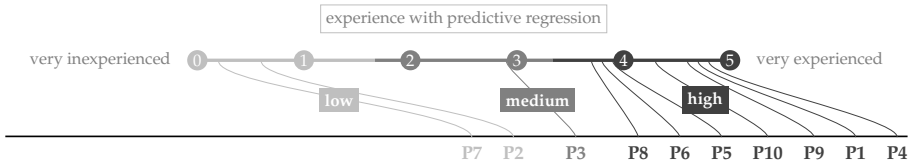
In addition, participants correctly interpreted the past fit as the *"fit between the model and the real data"* (P5), *"normalization of the slope"* (P3), *"average trend"* (P3, P6), *"natural evolution of the curve"* (P4), or *"total, general shape*

Table 5.1: Participants' background information, including their experience with predictive regression (l low, m medium, h high) as an average of self-reported experience (E_s), background (E_b), and jargon use (E_j). All participants identified as male and had a post-graduate education level.

ID	Profession	Country	Age	Experience (E_s, E_b, E_j)
P1	*Industry*: quality manager in a biscuit factory; deals with food safety issues, supply simulations	Greece	45–54	h 4.7 (4, 5, 5)
P2	*Industry*: food safety auditor for a certification body; audits companies on food safety and fraud	Greece	35–44	l 0.6 (0.3, 1, 0.5)
P3	*Industry*: quality manager in a biscuit factory; deals with food safety issues, supply simulations	Greece	35–44	m 2.9 (2.7, 3, 3)
P4	*Academia*: professor in mechanical engineering; expertise in food quality and life cycle assessment	Italy	45–54	h 4.8 (5, 5, 4.5)
P5	*Academia*: agricultural economist; expertise in value chains, food security and consumption	Italy	35–44	h 3.9 (2.3, 5, 4.5)
P6	*Industry*: sales manager for a refrigeration manufacturer; buys raw materials and sells products	Greece	35–44	h 3.8 (4.3, 4, 3)
P7	*Industry*: raw materials manager in a food company; recruits agriculturalists and keeps bees	Greece	18–34	l 0.2 (0, 0.5, 0)
P8	*Industry*: settlements coordinator in a mortgages company; verifying and approving mortgages *	Australia	35–44	h 3.7 (1, 5, 5)
P9	*Industry (Academia)*: researcher in agriculture; expertise in food chemistry and -microbiology	Greece	35–44	h 4.6 (3.7, 5, 5)
P10	*Academia (Industry)*: researcher in natural cosmetics; expertise in food science	Tunesia	18–34	h 4.3 (3, 5, 5)

* active in finance, no experience in agrifood.

Table 5.2: Some topics raised by the participants, ordered by their experience with predictive regression (P2 and P7 have low experience; P3 has medium experience; others have high experience).



	P7	P2	P3	P8	P6	P5	P10	P9	P1	P4
Understanding the visualisation										
Past fit and uncertainty are not understood	•	•								
Need for control										
More control over the prediction model						•			•	•
Need for comparisons										
Comparing countries is relevant	•			•				•	•	
Comparing products is relevant			•			•	•	•	•	
Comparing prediction models is relevant				•					•	
Need for tailored explanations										
Explaining the past data									•	•
Explaining the model's development process		•			•	•			•	•
Explaining the prediction model		•	•	•		•	•		•	•
Understanding the algorithmic level										
Visual components gradually improve mental model			•	•	•	•	•	•	•	•

of the price evolution" (P10). However, P2 and P7 did not understand the past fit line and P10 expected details when hovering over it.

Finally, while most participants seemed to intuitively understand the past uncertainty, they often lapsed into vague descriptions or were unsure how it was computed; e.g., *"it's the same like before: [. . .] the uncertainty factor"* (P3) or *"I think that you used your future model, whatever the model, and you tr[ied] to predict the past, I don't know"* (P6; *you* refers to the interviewer). Especially P2 and P7 could not get their head around the past uncertainty, with P2 questioning what others perhaps did not ask out loud: *"If you have the real numbers from the past, what's important about the uncertainty?"* Furthermore, P10 seemed to misinterpret the prediction intervals for showing accuracy: *"past uncertainty, it gives us like our model is most of the time, 85% accurate, let's say, in this point, and at the same point here it's 90%. I mean it gives us a better understanding of the model and if it's accurate or not"*.

In conclusion, it would be helpful to clarify the past fit and uncertainty components, especially for participants with low experience in predictive regression (see Table 5.2). To clarify the uncertainty, adapting the fans' tooltip could be a start because P6 pointed out that currently, some might confuse

the word ‘occasions’ with ‘iterations’ and therefore misinterpret the X%-fan as representing *“X out of 100 calculations.”*

Visual encoding of information: visually encoding uncertain price evolutions as a line graph with fans was clear yet limited.

All participants understood the visual encoding of price evolution as a line chart, and also the visual encoding of uncertainty as fans did not seem to cause confusion. Regarding the latter, P1 and P3 discussed the different shades explicitly: *“The more prices you get scattering around the line, the more, the deeper the shadow becomes [and vice versa]. So statistically, more prices are expected to be falling in a short distance above or below the line”.* (P1) and *“as it goes [from the prediction line] to the borders, [. . .] the possibility it goes down”* (P3).

Yet, the visual encoding has two limitations. First, when uncertainty components are enabled, simultaneously plotting multiple countries can be *“a little bit confusing”* (P2) or *“a little bit disturbing”* (P10) because of the many different colours and the overlapping graphical elements that hamper hovering specific fans. For example, when P8 plotted about 15 countries simultaneously, he said bluntly: *“Oof. [. . .] Yeah, I’m not really gonna get much out of that”.* Fortunately, participants realised that the trade-off between completeness and overplotting is their own responsibility: *“you cannot compare, I don’t know, 10 different commodities in 10 different countries, otherwise no one can understand what is shown in the graph”* (P5). Second, although participants understood that the Y-axis unit was not important for the study, they frequently mentioned that it should be clarified in real-life applications. For example, P6 joked: *“I mean, what is this 300? 300 cows or what?”*

Interacting with the visualisation: participants did not experience major filtering or hovering issues; zooming might be handy.

The filtering functionality was clear for all participants. Regarding the hovering functionality, getting details-on-demand through hovering seemed natural for both the line chart and the uncertainty fans. One minor remark here is that P5, P6, and P7 did not spontaneously hover over the fans when they first saw them, which suggests that a real-life fan chart might need to stress this possibility. Two participants found the highlighting of hovered uncertainty fans suboptimal. First, P8 regretted that he could not simultaneously highlight a fan and see price details (*“as soon as I move my mouse out, I lose it [the fan tooltip], so it’s very fiddly”*); and he proposed to allow pinning the fans. Second, P10 agreed that highlighted fans obscure other details and suggested altering their visual encoding from fans to lines that indicate standard deviations along with the corresponding probabilities.

In addition, P10's interactions in Scenario 7 demonstrated that a zooming feature could improve usability: P10 disabled the future uncertainty to reduce the Y-axis' length and thus artificially zoom in on the past fit lines to better see small-scale changes.

> **Workflow**: the current workflow for selecting products and countries was clear, but alternative workflows might be more efficient.

All participants understood the current workflow of first choosing a product and then selecting one or more countries. Yet, P5 and P9 proposed alternative workflows that could improve usability when focusing on a fixed set of countries. Tapping into the idea of focusing on a single country, P9 found it *"a bit annoying that anytime we are choosing a product [we need] to select again a country; [...] if you choose a product, you can play with the countries, but if you choose a country you cannot play with the products"*. Thus, to make the process of comparing different products for the same country less *"time-consuming"*, he would reverse the current selection order. Generalising this idea, P5 suggested a two-step selection workflow: an initial step to *"include all I want in the analysis– for example, different products for the same country or different countries for the same product"*, followed by visualising the selected information. Then, *"a sort of matrix with all the countries I have selected"* instead of dropdown lists would allow to quickly (de)select countries or products, which is, for example, convenient to remove overlap in the visualisation.

5.4.2 Usefulness and Needs

Participants raised two themes on usefulness (*Overall usefulness of the visualisation* and *Usefulness of the visual components*) and three themes on their needs (*Need for control*, *Need for comparisons*, and *Need for tailored explanations*).

> **Overall usefulness of the visualisation**: a visual DSS similar to ours was deemed useful for different tasks in agrifood or finance.

All participants agreed that visual DSSs similar to ours can be useful for different tasks in agrifood or finance. Generally speaking, P2 said that *"it's a very good tool for everyone in the food industry"* and P5 expected that *"a lot of people are looking for something similar"*.

More concretely, participants indicated that visualising predicted product prices can benefit industrial and academical agrifood parties. For agrifood companies, our visual DSS could be *"useful mainly in order to make future schedules"* (P9)

such that *"people who make decisions [and] who need insights in future price evolutions [...] can make contracts [with suppliers] for the coming years in order to avoid to pay too much"* instead of reacting to the market (P3). In addition, P2 saw a link with food fraud detection: *"the food price many times affects the food fraud cases [so] it helps companies to predict [the number of] food fraud cases"*. In agrifood research, P4 explained that researchers often study economical aspects such as demand and logistics, so he found our visualisation *"very interesting [...] to make some evaluation about the importance of some particular market and which is the prospective of that market"*.

Participants also saw more general applications for our visual DSS. For example, P10 stated that exporting companies would be interested in predicting demand in foreign countries, and P8 indicated that financial companies would be interested in predicting interest rates because *"this sort of helps you make better business decisions [... and] be better prepared"*. Thus, our visual DSS could be more useful when people can upload and visualise their own data. Furthermore, our visualisation is not bound to be a standalone tool: P1 *"would expect to see this dashboard attached in [a full analysis of the prediction model]; a text, showing, explaining how it works"* and P3, anticipating that the prediction model could consider climate change and geopolitics, saw the opportunity to extend our dashboard with additional visualisations of, for example, temperature and carbon emissions.

Usefulness of the visual components: how useful visual components were depended on the context, but uncertainty was a natural requirement for many.

Participants often mentioned that the usefulness of the visual components depends on the desired insight. For example, while P5 found all components *"very useful"* to analyse a single time series, he would probably hide the past fit and past uncertainty when comparing multiple time series: *"It depends in my opinion on what you want to visualise"*. In addition, P9 distinguished between obtaining precise values and drawing overall conclusions about the trend: *"You need [...] the future prediction to have an exact number [...] but just to make conclusions, you don't need it. You just need the [future] uncertainty and the fit"*. Last, P6 noted that he did not need an explicit dotted line to get a feeling about the general past trend. Given these considerations, the flexibility to enable and disable visual components in our visual DSS seems very useful.

Regarding the uncertainty components, most participants considered them a natural requirement because of the predictive context. For example, P1 said *"Whenever we need to predict something, there is always an uncertainty in our prediction. So it's more something that I would expect"*. and P8 agreed *"There are always going to be [macro level] factors that sort of change the*

prediction". Some participants even asked for future uncertainty representations right in Scenario 1: *"It could be interesting [...] to have the minimum and maximum value in that prediction period. A sort of standard value. [...] I expect [...] a sort of uncertain value [instead of] a precise value"*. (P4) and *"Maybe you should add some best cases and worse cases"* (P10). While discussing uncertainty, participants also touched upon a fundamental trade-off: *"It's like a double-shaped blade, you know. It gives you more liberty in choosing which kind of occasions you will be having, and at the same time, it gives you like not accurate results"*. (P10), and *"The thicker the lines [fans] become, the more useless the data because [...] everything is within specs, but you see you have a huge variation"* (P1). P4 added that, instead of multiple uncertainty levels, he only required a 1σ-interval. Overall, it thus seems essential to visualise the uncertainty in predictions, potentially allowing to modify the number of shown uncertainty levels.

Need for control: some participants requested additional control over the visualisation or the prediction model.

Some participants proposed additional features to explore the visualisation. Specifically, P5 suggested to allow filtering on specific time intervals; P5 and P10 proposed to allow changing the currency such that end users can better relate to the price evolutions; and P8 was looking for more in-depth pricing details such as the price per unit, retail price, and trading indicators such as the moving average convergence divergence.

In addition, some participants highly experienced with predictive regression voiced a need for more control over the prediction model (see Table 5.2). For example, P1 explained that he wants absolute control over prediction models: *"I use quite often the regression, the data analysis function in Excel. So I use the data in the way I want. I fit the models that I consider to fit best for the case. [...] The visualisation [...] would be quite helpful but based on what I have seen until now, I wouldn't [...] consider very much the prediction values. I would only use it for historical data acquisition"*. P5 also seemed to allude to this by stating that our visual DSS would be *"extremely useful"* if scientists and practitioners could download the available data and graphs for further analysis. Other requests for control were changing the predicted time span (P2, P4, P8) and the time frame used for training the prediction model (P1).

Need for comparisons: participants found it important to simultaneously compare countries, products, and prediction models.

Participants across all levels of experience with predictive regression stressed the relevance of comparing countries (see Table 5.2). For example, P9 said: *"Of*

course comparing different countries is really useful because we are talking about [...] a unite Europe [and] you might have incoming products from different countries. [... You] might have a purchaser from Italy and one from Germany, so you have both as an alternative to buy materials". Given this united European market, P8 added that he liked comparing prices with the European average.

Regarding the need to compare products, two ideas to extend our visual DSS arose. First, P3, P5, and P9 suggested to compare *similar* products (e.g., cereals, sweeteners, vegetal oils) in the same graph to understand potential relations between them. Such insights could, for example, be useful for farmers and regulatory bodies: *"the decision for farmers to produce rice instead of maize, or wheat instead of barley and so on, could be strongly conditioned by the provision [...], and regulatory bod[ies] for the market can provide specific support for specific farmers"*. (P5). Second, P10 suggested to simultaneously compare *different* products: *"For instance, if you want to make a muffin, you would have like flour, wheat, some milk, some eggs, flavour vanilla or chocolate. So you wanna keep each ingredient into consideration. [...] Maybe you can have like a [curve] for each ingredient [... and see the total] cost [for] the final product"*.

Last, participants experienced with predictive modelling would find comparing different prediction models useful to get an idea about how well they agree on their predictions and to, as P8 mentioned, follow the most frequent prediction, giving more weight to sophisticated models. Still, P1 emphasised: *"[I] would expect each model to be discussed: why does this model predict different values from another one and the reasoning behind that"*.

Need for tailored explanations: participants required tailored explanations about different aspects with different levels of detail.

Participants brought up four different aspects for which they needed explanations, and, interestingly, Table 5.2 shows that these participants had low to high experience with predictive modelling. First, P1 and P4 required a discussion of the past data and sudden peaks or troughs, backed by economical factors. Both P9 and P10, however, suspected that people active in industry would be most interested in explanations regarding the future, rather than the past. Second, participants wanted to know more about the provenance and accuracy of the raw price data, the model developers, the data processing, and the training of the prediction model. Third, P2 and P6 wanted to know how reliable the predictions were: *"The [end user] needs to feel that the model is predicting OK without knowing though what the model is doing. [...] You need somehow to explain to the end user what could be the prediction capability"*. (P6). Fourth, typically triggered by the steep predicted price increases in Scenarios 1–8, many participants requested explanations about the prediction model itself.

For example, P3 asked about the model's input factors: *"For me, it's very critical to understand what factors this model takes into account to predict such a high rise of the butter [price]."*, and P4 wanted *"a basic idea on how the prediction model works rather than going with something sort of blindly, [to see] evidence that this all works"*.

Furthermore, two participants had opposite views on the required level of detail in explanations. On the one hand, P1 requested full transparency of the prediction model: *"If it is a regression, I would be interest[ed] to see the equation that comes from the model. I would expect to see a discussion on the price variation, the reasoning"*. On the other hand, P6 vividly argued that he did not need this amount of detail: *" I don't believe you need to give it to a third party, to a user, when [they are] looking at data, the mathematics behind the model. [...] In my job, for example, one of the most important things is to know raw material prices [...] and I need to have a good prediction. Now how the prediction works? I really don't care"*.

The two observations above seemed to be part of a more general phenomenon: many participants alluded to tailoring explanations, i.e., adapting them to different contexts and to the people that need them. For example, P4 attributed his need for a description of the model to his *"research mind"*, but added that seeing uncertainty already filled part of that need, while economists would probably require more details: *"After see[ing ...] the statistical evaluation [uncertainty], in my opinion, my need [for a more detailed explanation] is lower because I of course consider the fact that perhaps they derived from some economical model that are at the basis of this evaluation. [...] Perhaps for economist[s ...] it would be more interesting to know something more about the model. But of course, this is not my topic so for me it's sufficient what I see in the graph"*. Similarly, when P1 asked for *"a very thorough discussion"* of the prediction model, he added: *"But this is me, OK. I'm an engineer, I'm quite experienced in mathematics and statistics and you understand, I know how it can work. I don't know if the same discussion was done with somebody who is not quite good in maths or in statistics, what his[/their] perspective would be."* Finally, while P5 found our visual DSS useful for educational purposes, he acknowledged that he would require more a detailed explanation when using it in high-stakes contexts: *"If I need to use it for a practical or a professional use, like the support for the country or the region, for a specific policy, and so on, I think I have to give them, to guarantee them about the quality of the data. And if I don't know exactly the model, what you have included and so on, and I couldn't replicate your analysis, it's quite impossible to use it as a standard or a benchmark"*.

5.4.3 Model Understanding

This section uncovers how the visual components and functionalities in our visual DSS impacted participants' understanding of the prediction model. Three themes, *Understanding the algorithmic level*, *Understanding the outcome level*, and *Understanding by comparing countries*, reveal that understanding manifested itself on an algorithmic and an outcome level.

> **Understanding the algorithmic level**: the visual components improved participants' understanding of the prediction model's technicalities, but only gradually.

In Scenarios 1 and 5, all participants indicated that simply plotting predictions does not invoke model understanding. For example, P5 stated: *"I have no idea which kind of variables you included in the model, if the model is based on different variables, I don't know, so the general international market or a policy decision, a local decision in France, or climate change or climate information. [...] and the technological evolution or [...] macroeconomic data"*. This lack of understanding was typically followed by a request for an explanation.

Yet, the stepwise introduction of extra visual components improved many participants' mental model of the prediction model, ranging from a better intuition to identifying the true modelling technique (see Table 5.2). To P3, P4, and P8, the future uncertainty suggested the model to be a statistical technique: *"It was more clear to me that we're not talking about, let's say, absolute values, but talking about the statistical model, so there you can see the possibility of the price evolution of the butter to be inside this space"* (P3). After enabling the past fit, P4 and P10 noticed that the past fit and future prediction formed a continuous curve, which gave them a better idea about how the prediction was constructed: *"[I] know in a better way the model [...] the evolution of the future is more clear [...] Of course, I don't know which is the mathematical model but I know that this is in a sort of curve, fit that you obtain, and so the model, I see the input from this evolution of the data"* (P4). Visualising the past uncertainty sometimes further reinforced understanding the link between past fit and future prediction: P4 noted that *"with this representation [...] the future prediction is completely integrated in the previously data"* and also P6, while unsure about how the past uncertainty was generated, got the feeling that the prediction was based on the trend line. After seeing the uncertainty and fit components, P1, P8, and P9 even strongly suspected that the prediction model was a regression. For example, P1 correctly identified the prediction as a third-degree polynomial, but he admitted that the visual components did not reveal the precise mathematical equation.

P6 explained how the *"step by step approach"* allowed him to *"understand parts of how the model works"* without revealing the technicalities: *"If you would only show me the first picture, no, I would not be able to tell you how the model works, but going to the future uncertainty and past uncertainty, and presenting also the trend line, then OK, you get a clearer picture of how the model probably works. But still, the details, it's not something that I think you can get with these simple steps"*. Furthermore, he added that none of the visual components was all-enlightening: *"Obviously it had to do with the whole sequence. [...] Step by step then you can get it. But it's not like you go like you know 'wow, wow, this is clear now'. [...] it is a gradual, let's say, picture"*. Interestingly, to improve the mental model faster, P4 suggested an alternative *"more logical"* order for enabling the visual components: he would first show the past data, past fit, and past uncertainty to explain *"that you have a statistical consideration"* and only then show the future prediction and future uncertainty to clarify that they are *"derived from the past fit"*.

Understanding the outcome level: the visual components allowed participants to interpret model outcomes and assess their accuracy.

Participants often commented that uncertainty did not explain the prediction's upward trend. For example, P5 said that *"uncertainty doesn't explain the prediction, it's just more inclusive"* and P1 added that he did not know whether he *"should expect that the price would increase or would decrease sharply"* after the prediction horizon.

However, the uncertainty and past fit components gave participants insights into model performance. In particular, P5 explained that the uncertainty revealed how well the model fits the data: narrow uncertainty meant a good fit; wide uncertainty meant a worse fit. The past uncertainty was also *"a sort of measure of the robustness of the model"* (P5), which gave a *"better understanding of the model and if it's accurate or not"* (P10) and indicated whether *"the past performance might repeat itself in the future, providing that the trend remains the same"* (P8). The past fit, then, allowed participants to detect outliers due to exceptional market events. For example, when P1 enabled the past fit, he said: *"[The] model has explained reasonably, reasonably, the variation of butter price throughout the decades. Could not predict the peak that occurred in 2008. Might have been an issue due to the financial crisis [...] We don't have this information but something has happened there that could not be predicted"*.

Finally, P6 proposed to assess model performance by comparing a country's past data with what the model predicts without that data: *"why you don't [...] compare what the model told us and what actually happened? Then you can evaluate also the effectiveness of your model"*.

Understanding by comparing countries: comparing countries had cons for understanding the algorithmic level, but pros for understanding the outcome level.

On the algorithmic level, the feature to compare countries sometimes led to misunderstanding the model's technicalities. This was illustrated by P3, who in Scenario 5 wrongly assumed that, to predict product prices in one given country, the prediction model also considered data from other countries: *"This model probably took into account what happened in the region, I mean in Europe, during this period of time. So that's probably why the price of the butter in France is going to rise so much. [...] now I can understand, let's say the reasoning behind this slope, why this slope is very steep [and] goes up"*.

On the outcome level, however, comparing countries allowed participants to better understand the model's performance. P4 and P10, for example, were especially interested in the model's consistency and expected that countries with similar price evolutions in the past would have similar price predictions. More importantly, in our experiment, showing data from France and the Netherlands allowed participants to compare the model's prediction for France with real data from the Netherlands. For P1, *"that was what actually convinced [him] that the model is quite unreliable"* because *"we can see that the actual data of Netherlands are far away from the prediction of the forecasted data for France"*. Similarly, P5, P6, and P8 emphasised that the divergence in price was accentuated by the fact that a large portion of the real data for the Netherlands did not lie inside the future uncertainty fans for France: *"the prediction buffer which should include all the data, more or less because it's 99% of the variation, doesn't include, doesn't encompass the real data [...] If we assume that [...] data of the Netherlands would be a reliable prediction [for France...] there is a big problem with the prediction model"* (P6).

5.4.4 Trust

Our results on trust consist of two parts. First, we present participants' quantitative trust evolution over the eight scenarios to spot differences and similarities. Next, we contextualise observed trends with the thematically analysed qualitative feedback.

Quantitative Results on Trust

Figure 5.4 shows the evolution of participants' reported trust scores over all scenarios. Overall, participants had very different trust evolutions. Yet, there

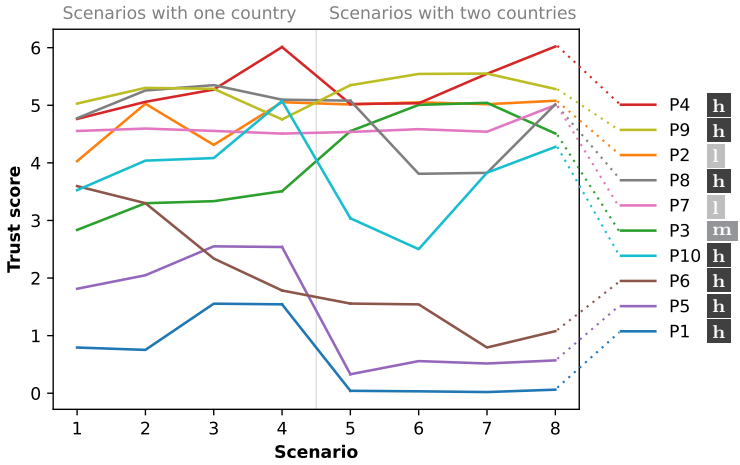


Figure 5.4: Participants' trust in the prediction model over eight scenarios. Scenarios 1–4 showed data for one country; Scenarios 5–8 showed data for two countries. Lines are slightly jittered for clarity. The legend includes the level of experience with predictive regression (l low, m medium, h high).

is a clear distinction between two groups: P1, P5, and P6 converged to low trust, whereas the other participants converged to at least rather trusting the prediction model. The level of experience with predictive regression did not explain this distinction because, for example, while P1 and P4 had the highest experience scores, they were both on different extremes of the trust scale. Another observation is that few participants reported dramatic changes in trust: only P6 and P10 have a difference of at least 2.5 between their minimal and maximal trust scores.

Qualitative Results on Trust

Four themes impacted trust in the prediction model. The first two themes, *Model performance* and *Model understanding*, were heavily impacted by expectation violation and expectation agreement: when participants encountered things that did not meet their expectations, their trust typically decreased, and vice versa. The other two themes, *Presence of uncertainty* and *Explanations*, tapped into what participants required for growing trust.

Model performance: seeing the model performance affected how participants assessed the model's trustworthiness; seeing model failures had a negative impact.

In Scenarios 1–4, participants assessed the prediction model based on the past fit and past uncertainty. The past fit did not decrease most participants' trust because it seemed to fit *"reasonably good the price variation, though in quite some [...] rough estimation"* (P1) and thus *"gives more robustness to the model"* (P5). Yet, for P6, the past fit highlighted that specific outliers were not foreseen by the model, which made him more unconfident: *"Why does not predict that it will have a peak and then go down again. [...] It does not persuade me. [...] I'm losing my confidence with a trend line"*.

Likewise, the past uncertainty led to mixed trust responses. On the one hand, some participants indicated aspects that increased their trust. For example, for P4, the option to do an *"evaluation of the data during the past"* increased his trust in *"the correctness of the model"*. P10 made a similar argument based on the fans showing the model's accuracy: *"I think this past uncertainty will add more credibility to our prediction model [...] I think I'm better trusting this model, [...] I can have a better understanding [...] of the prediction model as it goes over the years"*. Furthermore, P8 observed that most of the data points lay inside the uncertainty fans, but found it reassuring that some lay outside: *"[the price] falls out every now and again, which I mean, it does happen with everything. [...] I guess it increases my trust because if it was too perfect, you'd be like, you know, I mean, nothing in life is 100% certain, so why would this thing be?"* On the other hand, P6 actually became more hesitant when seeing a peak outside the 99%-fan: *"So there is a problem there, right? I know that it is only for a small period of time, like few months that the model fails over whatever I see here, like 25 years. But still, it fails. Is it acceptable? I don't know. I mean, if it was inside the band that I see here, maybe I would be happy"*.

In Scenarios 5–8, participants often assessed the prediction model by comparing the prediction for France with the real data or the past fit for the Netherlands. Many participants noticed a divergence between both: *"the butter price in France historically was closely linked to the butter price in Netherlands and we can see that the actual data of Netherlands are far away from the prediction of the forecasted data for France"* (P1). As Figure 5.4 shows, this resulted in a huge drop in trust for P1, P5, P6, P8, and P10 because they expected a prediction for France similar to the data for the Netherlands. For example, P1 said that he *"would not trust this model at all"* because it *"convinced me that the model is quite unreliable"*, and P6 motivated: *"I don't trust the model. You see, the real data was totally different than the prediction. [...] obviously, you*

prove that in a sense there are flaws in the model prediction".

Yet, not all participants experienced the divergence as an expectation violation. For example, P4 pointed out that the long-term performance seemed good and hypothesised that market events might have caused the divergence: *"in 2010 you have a differentiation. [...] the events that you have in Netherland are perhaps due to particular events that you had there, which I don't know, of course. [...] in the extrapolation, [...] the values are different, but the behaviour is very similar. [...] But I consider that at the end, five years later or 10 years later, also in Netherland you have the same price".* P8 make a similar remark in Scenario 6, restoring his trust afterwards: *"if you look at around 2016, the price prediction is way off. Way way off, but it sort of meets the further it goes along. So I think [...] if I was trying to make a price prediction like five years in the future or something, I trust it more, rather than, I would if it was one or two years in the future".* However, P5 called these observations of 'good' long-term performance a *"bias in the visualisation"* caused by the prediction for France coincidentally stopping at the real peaks of the Netherlands.

> **Model understanding**: participants' trust reactions differed depending on how they understood the prediction model on an outcome or algorithmic level.

Participants' model understanding on an outcome level influenced their trust. A typical example was Scenario 1, where the prediction line caused a lot of scepticism because it violated many participants' expectations for two reasons. First, participants could not understand its steep slope. For example, P10 joked *"it can't be like this: it goes like higher up in the sky. [chuckles]"* and P6 added: *"The trend of the previous ten years, no 20 years, does not imply that you're gonna have this rapid index increase".* Instead, some expected a price behaviour *"similar like the last 10 years, let's say"* (P3). Second, participants noticed that it did not have peaks or troughs like the past data: *"The thing that I'm worried about it that the curved line is like so decent, so perfect, so shaped".* (P10); and *"that peak that I see on October of 2007 and that trough that I see on March of 2009 is not what I see in the model prediction, comparing five years to five years".* (P6). However, most participants still reported a trust score above neutral because of mitigating considerations that agreed with their expectations. For example, P6 noted that in the last few plotted years, *"there has been an increasing rate which does not look too different toward what the prediction model has there".* Furthermore, due to *"the global inflation and the economic crisis etcetera, and a lot of pressure on the market places"* (P10) increasing prices seemed plausible: *"usually we have increase prices, not decreases [laughs], so that's why I'm more in the part that I'm trusting the prediction"* (P9).

Participants' trust was also affected by their model understanding on an algorithmic level. First, understanding decreased trust under expectation violation. For example, P6 understood that predictions were based on the past fit, but observed several unexpected things, which is why he insisted: *"I have a better understanding how the model works, but I don't trust it, I insist"*. Second, understanding increased trust under expectation agreement. For example, in Scenario 5, P3 gained trust because he built a (wrong) mental model that met his observations: *"this model probably took into account what happened [...] in Europe [...]. So that's probably why the price of the butter in France is going to rise so much. [...] I would say that I'm not suspicious anymore. [...] Because now I can understand, let's say the reasoning behind this slope"*. Furthermore, P9 reported high trust scores because he saw *"nothing strange. It's just what I was expecting to see. [...] it's just a regression [...] for me that I'm understanding how the models are working now, it looks normal"*. One comment here is that P9, upon seeing the diverging behaviour of France and the Netherlands in Scenario 5, also mentioned: *"it might, change my trust for the model as a model, OK, and how you incorporate the model in your data set but not for the prediction that we are generating for the future. Maybe a better model will give you better [results]"*. This suggests that P9 based his trust on *how* the prediction outcomes were computed, rather than whether regression was a suitable technique.

Presence of uncertainty: seeing that the prediction model accounted for uncertainty did not decrease participants' trust.

In Scenario 2, none of the participants indicated that their trust in the prediction model decreased because of the presence of future uncertainty. On the contrary, most participants' trust increased. P9, for example, explained why: *"the more descriptive the model becomes, and the more alternatives that it gives you, it makes you trust more. When you have just a line, you more or less, you cannot believe that things in real life are so accurate, right? [chuckles ...] I would say that future prediction without future uncertainty is not much trustful"*. While P1 and P3 agreed with this, they both stressed that the uncertainty did not increase their trust dramatically because it did not take away their need for an explanation: *"it's a model that takes some reasonable uncertainty, but still I cannot trust it because I don't know how it was developed"*. (P1); and *"I'm more, let's say, confident about this prediction model. But still, I want to know the reason why the butter has to go up"*. (P3).

For P4, the uncertainty overall generated more trust because it suggested the prediction model to be the product of scientific studies: *"I trust in a more–I suppose that behind this value you have some studies, some studies that come from your research for your ability"*. Furthermore, related to algorithmic model

understanding, P4 believed that the uncertainty suggested the prediction model to be of a statistical nature: *"I prefer the fact that the model works in a statistical way because with some consideration, I suppose this is more right in a model that works in the future. [...] I'm more and more trusting, trust about the correctness of the model"*.

Explanations: participants considered explanations about the development process and the prediction model requisites for building trust.

To trust the prediction model, participants mentioned that they needed an explanation about the development process and data provenance. For example, P1 said that *"in order to trust a prediction model, I need to know how it was developed"*. Furthermore, P4 and P5 alluded to the importance of who developed the prediction model. P4 referred to trusting the model developers' competence: *"when I approach information that come from an organisation or something like you, I suppose, my behaviour is to accept this evolution because I suppose that you have the competence to develop a model. [...] I have to believe in you with some [...] suspicious behaviour"*. In turn, P5 argued that a model stemming from an official institution might be more reliable: *"if such a prediction comes from an official body like FAO or World Bank or so on, could be more reliable, I can say. If come from a university [...] it's not an official body and it's more difficult to understand. So I just can imagine that [...] when a World Bank provide prediction, it's the fruit of the convergent opinion of different practitioners and scientists"*. Concerning the data provenance, P1 asked about the accuracy of the given historical data because *"in order to trust a prediction model, I need to know [...] what is the raw data [in]put"*.

Participants also considered an explanation about the prediction model itself key for building trust. For example, P5 did not trust the prediction in Scenario 1 because *"I have no idea how you provide this prediction, how you calculate it and the model behind. [...] there is no explanation of the model, and it's quite difficult to trust in the model without any description"*. P1 agreed: *"whenever I have a prediction model, I always try to find the physics and engineering behind that. If there is no physics explanation or engineering explanation, I'm quite sceptical"*.

5.5 Discussion

This section answers our research questions by discussing our quantitative and qualitative results. Then, based on our observations, it underlines the need for user-centred approaches in agrifood to increase the uptake of visual DSSs.

5.5.1 A User-Friendly and Useful Visual DSS

Our results show that participants were generally very positive about our prototypical visual DSS in terms of usability (RQ1): the visualisation, its interaction possibilities, and the general workflow were clear overall. In addition, participants imagined that a visual DSS similar to ours would be useful as support in several decision-making contexts, including food fraud detection, business scheduling, and market evaluation (RQ2). They also highly appreciated that our visual DSS fulfilled their need to compare countries and that visual components could be restricted to those relevant for desired insights. Thus, our prototype seems to be a user-friendly flexible basis for more advanced visual DSSs that extend our interface, and could be embedded in (dynamic) analytics reports.

Yet, we recognise two points of attention related to people's experience with predictive modelling. First, while many participants stressed the usefulness of uncertainty, our prototype could not remove all confusion around past uncertainty and past fit. Thus, especially for people who are less experienced with predictive modelling, it seems necessary to elaborate on the past fit and uncertainty components when used in a visual DSS. This could be realised with more detailed tooltips, a brief information screen, or—as suggested by Sacha et al. (2016)—a simple tutorial with some exemplar usage scenarios. Second, especially people with high predictive modelling experience could have a need for controlling and comparing different prediction models. To meet this need, visual DSS in agrifood could draw inspiration from visual analytics systems evaluated in other domains (Ali et al., 2019; Badam et al., 2016; Bögl et al., 2014).

5.5.2 Tailoring, Tailoring, Tailoring: Different End Users, Different Needs

Participants covered three important needs (RQ2): controlling the visualisation and prediction model; comparing countries, products and prediction models; and getting explanations about the past data, data processing, prediction reliability, and prediction model. Interestingly, other studies on predictive DSSs also revealed a need for comparison. For example, comparing cows' milk production allowed animal researchers to identify trends, clusters, and anomalies (Di Silvestro et al., 2014); and product demand analysts expressed the need to compare prediction performance for similar products (Sun et al., 2020).

Overall, participants' needs seemed heavily subject to their personal background and job activities. This shows the importance of tailoring visual DSSs and

explanations on at least three levels. First, *tailoring towards the application context*: the specific agrifood subdomain and the overall goal of the visual DSS determine which functionalities and visual components are useful. Second, *tailoring towards experience with predictive modelling*: for people with low experience, an intuitive understanding of the prediction model and little control over the prediction model might suffice, whereas people with high experience might require mathematical explanations and control over the prediction model. Third, *tailoring towards tasks*: different tasks and desired insights might require different visual explanations, similar to what Gutiérrez et al. (2019b) argued for.

5.5.3 Gradual Model Understanding through Visual Analysis

The visual components and comparison functionality in our visual DSS affected participants' model understanding on two levels (RQ3). On an *algorithmic level*, many participants gradually grew a better intuition of the model's technicalities. In XAI terms, the visual components thus served as explanations that fostered their mental model. On an *outcome level*, participants could better interpret predictions and assess their accuracy.

However, some participants created mental models that did not stroke with the real regression model. For example, they assumed that the model based its predictions on price evolutions in multiple countries or considered additional input variables such as climate and geopolitics. This suggests that complementary explanations are necessary to avoid wrong assumptions, bearing in mind that these explanations should balance soundness and completeness (Kulesza et al., 2013): simply adding more information does not necessarily spark useful mental models. Other participants' model understanding did not improve because they could not analyse the visualised information thoroughly, most likely due to low experience with predictive regression or time series analysis overall. To grow correct model understanding, such end users seem to require more guidance in the data analysis process; it is unclear whether the current exploratory nature of our visual DSS fits this need.

5.5.4 Trust Is Multi-Faceted and Evolves

Our results subscribe to the multi-faceted and evolving nature of people's trust in a prediction model (RQ4), similar to many previous studies (Hoff and Bashir, 2015; Holliday et al., 2016; Nourani et al., 2020; Ooge and Verbert, 2021). We identified four themes that influenced people's trust: the model's performance, understanding the model, uncertainty in the model's outcomes, and explanations

about the development process or the prediction model itself. The former two themes were strongly coloured by whether participants' expectations were violated or met; the negative impact of expectation violation is in line with findings from Kizilcec (2016). The latter two themes covered what participants deemed necessary to grow trust. The fact that participants required the presence of uncertainty for building trust reinforces the call for incorporating uncertainty in visual DSSs for agrifood.

We observed clear evidence of trust calibration (Sacha et al., 2016): participants' trust was based on a continuous trade-off between the aforementioned four themes. The direction in which their trust evolved then depended on which theme was most dominant. For example, most participants initially focused on requiring explanations. Some then evolved to distrusting the prediction model due to low performance, whereas others developed more trust due to observations that matched their model understanding. This explains the different trust evolutions in our quantitative measurements. An important note here is that the quantitative scores are hard to compare directly because participants typically have different calibrations for scoring. On an individual level, though, we found that most participants' trust scores did not change drastically over the eight scenarios. For participants with low experience in predictive modelling, this was most likely due to their inability to fully analyse the visualised information. Why these participants trusted the prediction model nevertheless is unclear. Potentially, factors such as good usability fostered their trust, or the participants reported what they conceived as desirable.

5.5.5 Fostering Appropriate Trust Through Usefulness and Meeting Needs

While our results presented four evaluation metrics and their corresponding themes separately, some themes are connected or partially overlap. Figure 5.5 summarises all themes together with their most relevant relations grounded in our qualitative data. The relations clearly link usefulness to trust, either directly, or indirectly via model understanding.

Two direct relations concern uncertainty and explanations. First, while uncertainty was considered a natural and useful requirement for bringing nuance to predictions, participants also considered it a requisite for building trust. There exist interesting parallels in other domains: for example, people tend to discount weather forecasts without uncertainty (Franconeri et al., 2021). Second, participants often stressed a need for explanations about the prediction model and its development process, adding that they could not build trust

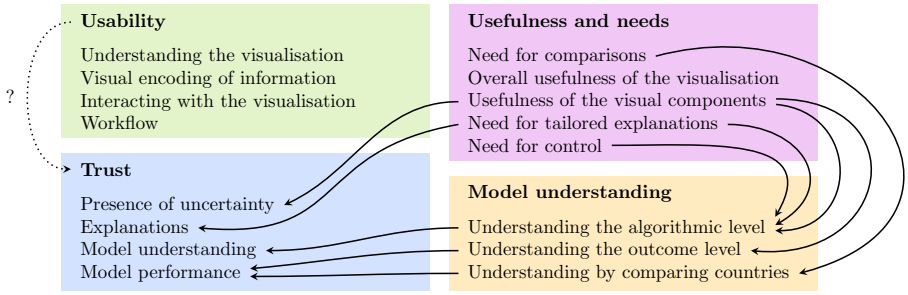


Figure 5.5: Summary of the themes on usability, usefulness and needs, model understanding, and trust. Some relations between themes are indicated with arrows; themes are reordered to avoid overlap.

without them. This illustrates the relevance of XAI research into the utility of explanations (Davis et al., 2020).

Two indirect relations link usefulness to trust through model understanding. First, the visual components in our DSS were deemed useful for understanding the model on an algorithmic level. Control over the prediction model and tailored explanations about the prediction model were expected to facilitate the same. In turn, observing things that agree with model understanding led to increased trust. This suggests that improving model transparency with tailored explanations, for example carefully designed visualisations, can foster appropriate trust, which is in line with common beliefs in the XAI community (Gunning and Aha, 2019). Second, the visual components and the functionality to compare countries in our DSS allowed participants to better understand model outcomes, which in turn revealed model performance. Seeing the prediction model's performance allows assessing its trustworthiness, which is essential for appropriate trust (Han and Schulz, 2020; Solhaug et al., 2007).

5.5.6 Taking a Step Back: Increasing Uptake of DSSs in Agrifood with User-Centred Approaches

Before concluding, we reflect upon the broader impact of our findings for agrifood. Central in our overall story was the lacking uptake of (visual) DSSs in agrifood. Rose et al. (2016) pointed out that trust is a key factor for increasing uptake. Quotes from our interviews such as *"I think that for a scientist I can use prediction data only if my trust on this data is full"* (P5) and *"you don't have the time to [...] explore if the model works or does not work. [...] I just want to believe what I have in front of me"* (P6) indeed seem to confirm that

people will not use applications they distrust. From this point of view, it seems reasonable that scholars and practitioners in agrifood and other domains often advocate for designing DSSs that increase trust.

However, simply designing for increasing trust is not always desirable and should not be the final goal because trust eventually manifests itself when applications prove to be reliable and useful over time (Davis et al., 2020). Our results, summarised in Figure 5.5, support this claim: the relations between usefulness and trust suggest that useful and tailored visual DSSs may eventually foster appropriate trust. Therefore, it seems recommended to apply user-centred approaches to design useful DSSs that meet end users' needs. In the long run, this can foster appropriate trust and in turn uptake. Furthermore, user-centred approaches have the additional asset of exposing people to new technologies (Parker and Sinclair, 2001), which can also stimulate trust (Rose et al., 2016). Thus, user-centred approaches seem vital for ameliorating the current low uptake of visual DSSs in agrifood.

5.5.7 Limitations and Transferability

Our research is subject to some limitations. Most importantly, our sample of 10 participants is most likely too small to achieve full data saturation in our qualitative results. Yet, it is encouraging that our trust themes largely correspond to those found in our pilot study (Ooge and Verbert, 2021). Larger studies could investigate whether more themes emerge concerning trust as well as the other evaluation metrics. To further validate our observed differences between people with different levels of experience in predictive regression, it would be particularly interesting to include more people with low or medium experience. Furthermore, future work can investigate the transferability of our results to other domains such as finance and healthcare, where predictive models play an important role too. Since our sample contained only one participant active in finance, we cannot draw strong conclusions on potential differences with agrifood yet. Finally, as good performance is a core factor for uptake of DSSs (Rose et al., 2016), real-life applications based on our prototypical visual DSS should include suitable models for forecasting time series, for example, exponential smoothing or LSTM (Brockwell and Davis, 2016; Hyndman and Athanasopoulos, 2018).

5.6 Conclusions

We presented a prototypical visual DSS for agrifood that incorporates price prediction, uncertainty and visual analytics techniques. An elaborate evaluation with 10 participants active in agrifood or finance revealed many insights concerning usability, usefulness and needs, model understanding, and trust. For example, participants were generally very positive about our prototype's usability and discussed needs regarding control, comparison, and explanations. Our results also show that usefulness and trust are related, either directly, or indirectly through model understanding. Moreover, we observed that participants' job activities and experience with predictive modelling influenced their perceptions and needs. Combining all these findings illustrates that user-centred approaches are vital for increasing the uptake of visual DSSs in agrifood.

Acknowledgements

This research was funded by Research Foundation-Flanders (FWO, grant G0A3319N), the Slovenian Research Agency (grant ARRS-N2-0101), and the European Commission (Horizon 2020, grant 780751). Thank you to all participants for their valuable feedback and openness. Thank you to Vivi Katifori for helping us with the recruitment of participants and setting up the interviews. Thank you to Oscar Alvarado for sending us references on HCI and thematic analysis. Thank you to Francisco Gutiérrez and Nyi Nyi Htun for initial brainstorms on our visual DSS. Thank you to Aditya Bhattacharya, Robin De Croon, Diego Rojo, Arno Vanneste and the two anonymous reviewers for providing helpful comments that improved this text.

The Human Side of Chapter 5

In-Person and Virtual Journeys

I started this project in January 2020, at the start of a period of frequent self-doubt and gloom. I guess stress, my yearly winter blues, and the social restrictions during the COVID lockdowns weighed heavier than expected. In addition, it didn't help that I couldn't turn my head around how to meaningfully assess people's model understanding, trust, and perceptions of uncertainty visualisation during an interview. One of the things that really helped me in this difficult period were long walks in the Egenhoven Forest, Jesuit Park, and Heverlee Forest. I also started photographing more, especially small things such as the blue beetle in the picture. Unless I was listening to Numberphile podcasts to absorb other researchers' life lessons, the comforting nature helped me think about a decent research plan. Fun fact: I had many breakthroughs in a deserted research site, but I'm still unsure whether I was allowed to be there.

Songs on repeat:

- *Indecision* and the rest of the *Nothing's Real* album by Shura
- *Comeback Kid* by Sharon Van Etten
- *Broken Sleep* and the rest of the *Myopia* album by Agnes Obel

Blue beetle in Egenhoven Forest – April 2020

Once I finished my research plan and prototype, I finally kicked off the interviews in July 2020. It was wonderful to talk via Skype with people located all around the world and their feedback gave me the perfect research materials. My virtual travels to Tunisia, Greece, Italy, Hong Kong, and Australia also ignited a desire to explore new parts of the world in person once the COVID situation allowed it (Belgium was in a kind of soft lockdown during the summer of 2020). This picture reminds me of the Droste effect you sometimes encounter when screen sharing during a video call.

Songs on repeat:

- *Goya! Soda!* by Christine and the Queens
- *Cookie Jar* by Doja Cat
- *Back of a Cab* and the rest of the *Make My Bed* album by King Princess
- *Pretty Girl* by Clairo

Bicycle parking next to the imec tower – May 2020

This gift from my sister symbolises an adventurous writing journey in August 2021. Back in October 2020, I attended the online TREX workshop on *TRust and EXpertise in Visual Analytics*. The work presented there gave me an extra boost because it aligned perfectly with the topics I covered in my interviews. In the following months, I transcribed, annotated, and coded my interviews. Yet, analysing them over and over again, I was struggling to mould participants' conflicting perceptions into a consistent story. In the meantime, a new edition of TREX was announced and the submission deadline happened to be the day before I planned my summer leave. A perfect target. However, being a talented procrastinator and doubting the value of my thematic analyses, I still hadn't started writing a paper the day before. Then, I decided to write a paper in 24 hours. The story had been in my head for months and once in the writing zone, I couldn't stop. Overnight, I realised I had to get my COVID vaccine in the morning in my parents' home town. And so it happened I was frantically writing the paper's discussion on the train there, finishing it while getting vaccinated. Yes, people asked questions. Even after rushing to my parents' house, I immediately locked myself in my room to do the finishing touches and submit the paper. Only when I allowed my family to come in, I realised I submitted too late due to time zone differences. Fortunately, the ever-sweet workshop organiser Mahsan Nourani saved the day and my paper: it got accepted (Ooge and Verbert, 2021).

Wooden T-Rex – August 2021

By April 2022, I had finished the thematic analysis of the interviews and a first version of the paper. Right on time for my first adventure abroad since the start of the COVID pandemic: the CHI 2022 conference in New Orleans (United States). Even though I attended the conference without a paper (that is, as a tourist), it was an incredible experience. After almost three years, I could finally connect in person with the human-computer interaction community, and I enjoyed networking and exchanging research ideas wholeheartedly. After the conference, my partner Yens came over and we had an amazing time exploring New Orleans. One day, we organised a second hand book store tour and discovered *Arcadian Books & Prints*, where thousands of books were crammed in a space the size of a large bedroom. The book maze reminded me of my thematic analysis: interesting stuff everywhere, yet it took time and effort to spot the patterns hidden in the chaos. Shortly after New Orleans, charged with new energy, I finished and submitted the paper. End of June 2022, Yens attended a conference in Belgrade (Serbia) and it was my turn to visit him. Coincidentally, I received the reviews during breakfast in the heart of the city. And thus, this project filled with journeys fittingly ended during one.

Songs on repeat:

- *Woman Is a Word* by Empress Of
- *I Don't Even Smoke Weed* and the rest of the *Us* album by Empress Of
- *Love Is A Drug* and the rest of the *I'm Your Empress Of* album by Empress Of

Book store in New Orleans – May 2022

Part II

Explainability Through Visualisation-Supported Justification

Chapter 6 presents a study on how visualisation-supported justification for recommended learning exercises affects teenagers' trust in an e-learning platform. This chapter was published as a conference paper (Ooge et al., 2022a):

> ***Ooge, J.****, *Kato, S.**, and Verbert, K. (2022). Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In* 27th International Conference on Intelligent User Interfaces*, IUI '22, pages 93–105, New York, NY, USA. Association for Computing Machinery*

This work is the outcome of the master's thesis by Shotallo Kato, which I guided intensively. As joint first authors, we contributed equally to defining the research plan, iterating over the visual explanation designs, and interpreting the results. Shotallo conducted all user studies, did the implementation, and did most of the data analysis, whereas I did most of the writing. Moreover, I presented the paper at the IUI 2022 conference. The methods, results, and text were discussed with Katrien Verbert.

Relevant to this part of the thesis is my contribution to the following conference paper (Bhattacharya et al., 2023), briefly described on Page 159:

> *Bhattacharya, A., **Ooge, J.**, Stiglic, G., and Verbert, K. (2023b). Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data- Centric Explanations and Combinations to Support What- If Explorations. In* Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 204–219, New York, NY, USA. Association for Computing Machinery*

As the second author, I conducted a focus group, helped iterate over the visual explanation dashboard and the research plan, and provided feedback on earlier versions of the paper.

Chapter 6

Explaining Recommendations in E-Learning

In the scope of explainable artificial intelligence, explanation techniques are heavily studied to increase trust in recommender systems. However, studies on explaining recommendations typically target adults in e-commerce or media contexts; e-learning has received less research attention. To address these limits, we investigated how explanations affect adolescents' initial trust in an e-learning platform that recommends mathematics exercises with collaborative filtering. In a randomized controlled experiment with 37 adolescents, we compared real explanations with placebo and no explanations. Our results show that real explanations significantly increased initial trust when trust was measured as a multidimensional construct of competence, benevolence, integrity, intention to return, and perceived transparency. Yet, this result did not hold when trust was measured one-dimensionally. Furthermore, not all adolescents attached equal importance to explanations and trust scores were high overall. These findings underline the need to tailor explanations and suggest that dynamically learned factors may be more important than explanations for building initial trust. To conclude, we thus reflect upon the need for explanations and recommendations in e-learning in low-stakes and high-stakes situations.

6.1 Introduction

People are increasingly relying on recommender systems that suggest relevant items, for example movies and music, tailored to their needs and interests.

However, people are often left in the dark when it comes to why something has been recommended. In the scope of *explainable artificial intelligence* (XAI), many researchers agree that accompanying recommendations with explanations is often desirable because it can, for example, increase appropriate trust in the recommender (Adadi and Berrada, 2018; Mohseni et al., 2021; Tintarev and Masthoff, 2011), which in turn can increase people's willingness to adopt technologies and their outcomes (Wang and Benbasat, 2005). Therefore, XAI and trust have become prominent research topics in human-computer interaction.

However, the degree to which results of previous research on explaining recommender systems can be generalized is limited because of three reasons. First, studies are mostly framed in application contexts like media recommending, e.g., (Berkovsky et al., 2017; Gedikli et al., 2014; Millecamp et al., 2019; Tintarev and Masthoff, 2012), and e-commerce recommending, e.g., (Panniello et al., 2016; Pu and Chen, 2006; Wang and Benbasat, 2005). Other contexts such as education are explored less (Barria-Pineda, 2020). Second, most study participants are university students or adults, resulting in scarce results for adolescents (ages 11–19 (Fitton et al., 2013)). Third, on a methodological level, most XAI research measures the effect of explanations by comparing recommender systems with and without explanations. However, this comparison could be unfair as recent studies suggest that the mere presence of *placebo explanations* (i.e., explanations without any meaningful content) can already increase someone's trust in an intelligent system (Eiband et al., 2019).

To address these limitations, we investigated how explanations affect adolescents' trust in an e-learning platform that recommends mathematics exercises, and added placebo explanations as an extra baseline. In particular, we had two research questions:

RQ1. Can explanations increase adolescents' initial trust in an e-learning platform that recommends exercises?

RQ2. How do placebo explanations influence adolescents' initial trust in such an e-learning platform?

Our research contribution is threefold. First, we show that explaining recommendations can significantly increase initial trust in an e-learning platform if trust is measured multidimensionally. However, when measuring trust one-dimensionally, the increase is not significant, which suggests that mainly dynamically learned factors grow initial trust. Second, by comparing our explanation interface with a placebo baseline, we reveal that adolescents have different needs for transparency, so tailoring explanations is essential. Third, we present unique data on how adolescents trust and interact with our e-learning

platform, which we share publicly in the spirit of open science[1]. In sum, we hope our work inspires other researchers to more often target adolescents and study the impact of tailored explanations in e-learning.

6.2 Background and Related Work

This section discusses some challenges of explaining artificial intelligence, and particularly recommender systems. Then, it zooms in on trust in automated systems and previous studies on the trust effects of explaining recommendations.

6.2.1 Explainable Artificial Intelligence

Ever since the resurgence of artificial intelligence, there has been a call for algorithmic transparency. Sophisticated algorithms are namely often 'black-boxes': it is unclear how they precisely process vast amounts of input data to obtain an output. Not explaining algorithms' outcomes may suffice for low-stakes applications such as movie recommendation but becomes unacceptable in high-stakes contexts such as healthcare and e-learning. *Explainable artificial intelligence* (XAI) is an umbrella term for techniques that try to explain the logic behind algorithmic decision-making, such that people can understand it, grow appropriate trust in the algorithm, and detect potential biases (Gunning and Aha, 2019). A substantial challenge is that XAI encompasses many intertwined topics including trust, fairness, bias, causality, accountability, privacy, and human reasoning (Abdul et al., 2018). As a consequence, it is hard to find all-embracing definitions for XAI and concepts like 'explainability', 'interpretability', 'understandability' and 'intelligibility' (Doshi-Velez and Kim, 2017; Gilpin et al., 2018; Lipton, 2018).

Because of its broadness, the XAI problem can be approached from different angles. Researchers in artificial intelligence follow an *algorithmic* approach: they develop model-specific and model-agnostic techniques to investigate the local and global behavior of machine learning models and their robustness against data perturbations (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020; Guidotti et al., 2019b). In contrast, researchers in human-computer interaction follow a *human-centered* approach: they often draw on the social sciences (Ehsan and Riedl, 2020; Miller, 2019) and let human reasoning processes inform XAI techniques (Wang et al., 2019a). In short, this led to the understanding that there is no such thing as a one-size-fits-all explanation. Instead, design requirements for explanations depend on the application context (Dhanorkar

[1] https://github.com/JeroenOoge/explaining-recommendations-elearning

et al., 2021; Vellido, 2020) and the target audience's goals and personal characteristics (Berkovsky et al., 2017; Millecamp et al., 2019; Mohseni et al., 2021); and explanations can be evaluated according to several metrics (Hoffman et al., 2019; Mohseni et al., 2021).

6.2.2 Explaining Recommendations

A lot of XAI research builds upon earlier research with recommender systems (Tintarev and Masthoff, 2007a). For example, Herlocker et al. (2000) compared several explanation designs for collaborative filtering recommenders to increase acceptance of recommendations. Today, explaining recommender systems is still a hot research topic, e.g., (Donkers et al., 2020; Jin et al., 2018; Kouki et al., 2019; Tsai and Brusilovsky, 2019b), generating lively reciprocity with the wider XAI domain.

In general, explanations for recommendations come in three representational forms (Nunes et al., 2017). First, *textual explanations* use natural-language phrases. Many commercial applications already employ these kinds of explanations, following patterns like "*People who liked X also liked Y*" for collaborative filtering recommenders, and "*You will like X because it has Y and Z*" for content-based recommenders. Second, *visual explanations* use (interactive) visualizations to efficiently convey a lot of information. For example, Herlocker et al. (2000) used histograms to show how neighboring users rated a recommended movie; Tsai and Brusilovsky (Tsai and Brusilovsky, 2019a) explained similarity-based recommenders amongst others with radar charts and Venn diagrams; and Bostandjiev et al. (2012) visualized a music recommending process with an interactive pathway chart. Third, *hybrid explanations* leverage both textual and visual information. For example, Gedikli et al. (2014) used tag clouds in which word size encodes relevance, and Szymanski et al. (2021) combined a partial dependence plot with text on how to interpret the visual information.

Designing explanations for recommendations brings challenges concerning *what* and *how* to explain (Eiband et al., 2018). Usually, the recommendation algorithm constrains the explanation type (Tintarev and Masthoff, 2011). For example, collaborative filtering recommendations cannot be explained by their inherent features. Furthermore, designing explanations involves making several trade-offs (Kulesza et al., 2013). Tintarev and Masthoff (Tintarev and Masthoff, 2007b, 2011) discussed this in detail and outlined seven goals for explanations which are not all simultaneously satisfiable: transparency, scrutability, effectiveness, persuasiveness, efficiency, satisfaction, and trust.

6.2.3 Trust in Automated Systems

Trusting automated systems has been found essential for adopting them (Pu and Chen, 2006; Wang and Benbasat, 2005). At the same time, trust research is somewhat controversial (Davis et al., 2020) because optimizing systems' designs to grow trust might lead to inappropriate trust, which can entail undesirable effects like misusing technology (Bussone et al., 2015; Merritt et al., 2013). In addition, trust is a complex topic. On the one hand, it has been defined in many different ways, depending on the field or context (Madsen and Gregor, 2000) and entailing different themes such as competence, benevolence, and reliance (Chopra and Wallace, 2003; Cramer et al., 2008; Grandison and Sloman, 2000; Lee and See, 2004; Muir, 1987; Wang and Benbasat, 2005). On the other hand, it has been recognized that trust is not static but evolves (Holliday et al., 2016; Nourani et al., 2020; Ooge and Verbert, 2021). Thus, measuring trust in automated systems is challenging and researchers have proposed explicit and implicit measuring techniques.

Explicit measuring techniques ask people about their trust perceptions in questionnaires or interviews. *One-dimensional* approaches measure trust with a single Likert-type question (Holliday et al., 2016; Millecamp et al., 2019; Nourani et al., 2020). Although this method is quick and easy, it is susceptible to people interpreting 'trust' differently. Therefore, *multidimensional* approaches use Likert scales to measure trust as an ensemble of multiple constructs. For example, McKnight et al. (2002) introduced the concept of *trusting beliefs* (Vidotto et al., 2012), consisting of the constructs *competence*, *benevolence*, and *integrity*. Later research added more constructs, including *perceived transparency* and *intention to return* (Berkovsky et al., 2017; Pu and Chen, 2007). Overall, while a multidimensional approach is more nuanced than its one-dimensional counterpart, it requires longer questionnaires and is therefore more time-consuming.

Implicit measuring techniques avoid the self-reporting bias in explicit measurements by measuring trust through an intermediary. Examples are: loyalty measured by the number of logins after sign-up (McNee et al., 2003; Tintarev and Masthoff, 2011), acceptance rate for recommendations (Cramer et al., 2008), time spent on a page, click-through rate, and page-exiting manner (Fox et al., 2005). In the context of explaining recommender systems, implicit measurements for trust have not yet been widely adopted, possibly because intermediaries like loyalty require long(er)-term studies.

6.2.4 Trust in Explained Recommendations

Previous research has shown that providing explanations for recommendations can increase the acceptance of recommendations (Cramer et al., 2008; Herlocker et al., 2000), and increase people's trust in the recommender system (Berkovsky et al., 2017; Pu and Chen, 2006). While previous studies typically focused on recommenders for movies or e-commerce, e.g., (Kunkel et al., 2019), research in an e-learning context is limited (Barredo Arrieta et al., 2020; Daher et al., 2017). This is unfortunate as Abdi et al. (2020) recently demonstrated the potential of a transparent educational recommender system: an Open Learner Model (Bull and Kay, 2010) improved understanding of and trust in recommendations for learning materials.

As trust is a relative measure, it must be compared to some baseline. Studies on the effects of explanations typically include a baseline with no explanations. However, a lesser applied baseline are *placebo explanations*. These 'pseudo explanations' are semantically insensible (Langer et al., 1978), i.e., they do not reveal any information about why something was recommended, for example "*This has been recommended to you because this is what the algorithm calculated.*" Surprisingly, Eiband et al. (2019) found that placebo explanations can invoke similar trust levels as real explanations. However, Nourani et al. (2019) found conflicting results outside the domain of recommender systems: placebo explanations lowered the perceived accuracy of an image recognition system.

6.2.5 Underexplored Research Areas

Our literature overview shows that XAI re-nourishes the interest in explaining recommender systems and how that affects trust in recommendations. However, we see two underexplored areas. First, research on trust and explaining recommender systems primarily focuses on university students or adults and often neglects adolescents. Second, while e-learning platforms increasingly adopt recommendation algorithms (Abdi et al., 2020; Dahl and Fykse, 2018; Klinkenberg et al., 2011; Manouselis et al., 2014; Verbert et al., 2012), they lack explanations for their recommendations. Our research addresses both shortcomings: we design hybrid explanations for an exercise recommender on an e-learning platform and investigate their effects on adolescents' *initial* trust (i.e., trust based on their first impressions of the platform).

6.3 Materials and Methods

This section presents our e-learning platform with explanations for recommended exercises and our overall study design. Our research was approved by the ethical committee of KU Leuven (reference number G-2021-3233-R2(MAR)).

6.3.1 E-learning Platform with an Exercise Recommender

For our study, we built upon an existing e-learning platform called Wiski (Ooge, 2019), which was developed in Drupal 7 and contains over 1000 multiple choice exercises on mathematics topics in the Belgian high school curriculum. To estimate the difficulty level of exercises for each student, we set up an *Elo rating system* (Elo, 1978) for students and exercises: if a student correctly solves an exercise, their Elo score rises and the exercise's Elo score drops, and vice versa.

We used the Elo rating in two ways. First, students could see the estimated difficulties while browsing exercises (see Figure 6.1d) to manually pick exercises suited for their level of mastery. Second, inspired by Dahl and Fykse (Dahl and Fykse, 2018), we recommended exercises with an algorithm implemented in Python 3.8.5. When students solved an exercise on a certain topic, they received three suggestions for follow-up exercises on the same topic. Broadly, our recommender system combines Elo ratings and collaborative filtering: it looks for candidate exercises based on a student's Elo rating and recommends those that the student is most likely to answer correctly. More specifically, to recommend exercises about topic T for student A, our algorithm follows three steps. First, the 7 exercises about topic T with an Elo score closest to the value $\text{Elo}_\text{A} + 50$ are selected as candidates. We added the constant 50 to promote recommendations that slightly exceed students' level of mastery (Wauters et al., 2012). Then, for each candidate exercise E, the algorithm estimates with nearest-neighbors how many attempts A may need to solve E: it looks for students who solved E, selects at most 40 of them close to A in terms of attempts for previously solved exercises (Pearson similarity), and takes a weighted average of their number of attempts for E. Finally, the three candidate exercises with the lowest average number of attempts are recommended in ascending order.

6.3.2 Explanations for Recommendations

To accompany the recommended exercises, we designed three explanation interfaces, following a user-centered design process. Specifically, we iteratively refined an initial design during three rounds of think-aloud studies with 16

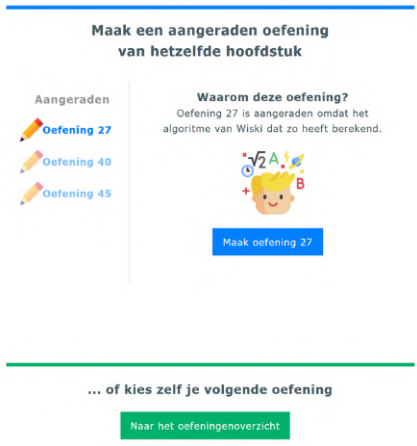
participants (1 teacher, 5 middle school students, 9 high school students, 1 university student). In these think-alouds, participants executed predefined tasks that tested the usability of our interfaces and answered additional questions related to usability, transparency, and explanations in general. We wrote down all relevant remarks and afterwards grouped them thematically to identify the most frequent issues. Based on the collected feedback, we dropped initial designs for transparency pages that explained collaborative filtering, and made the role of certain components in our explanation interfaces more explicit such that students could process them quicker. More details can be found in Kato's Master's thesis (Kato, 2021).

Figure 6.1 presents our three final explanation interfaces. The first interface (Figure 6.1a) contains a real explanation, consisting of three parts [*English translation in brackets*]: 1 a why-statement which indicates that the exercise was recommended based on both the student's level of mastery and the exercise's difficulty [*Why this exercise? Wiski thinks your current level matches that of this exercise!*]; 2 a justification-statement with the student's estimated number of tries needed to solve the exercise [*Wiski expects that you will need 1 or 2 attempts to answer exercise X correctly, based on your results and that of your peers*]; 3 a histogram of how many tries similar students required for the exercise, inspired by Herlocker et al. (2000) [*Number of attempts peers needed to solve exercise X correctly*]. To avoid students seeing (nearly) empty histograms at the experiment's cold start, we pre-populated the data set with mock data based on logging data from a past experiment on Wiski that used identical exercises (Ooge, 2019). The second interface (Figure 6.1b) contains the placebo explanation "*Exercise X is recommended because this is what Wiski's algorithm calculated*", which indeed conveys no information about how our recommendation algorithm works. Finally, the third interface (Figure 6.1c) simply states that the exercise was recommended, without further clarification.

6.3.3 Participant Recruitment

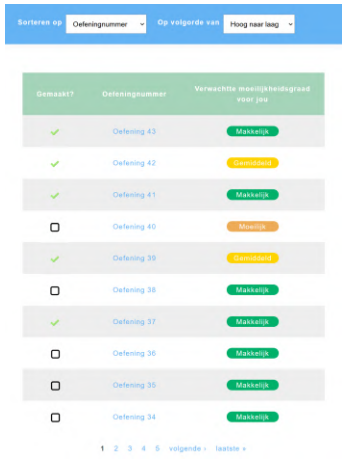
We contacted teachers of 18 high schools in Belgium (Flanders) and invited them and their students to participate in our research. Teachers and students received an information leaflet that described the research process, stressing that students could not be coerced into participating and would receive an equivalent substitute task if they did not wish to participate. Interested students then gave informed consent and students under the age of 16 also required signatures from their parents. In addition, we recruited extra participants through snowball sampling.

(a) A real explanation for the REAL group with 1 a why-statement, 2 justification-statement, and 3 histogram.



(b) A placebo explanation for the PLACEBO group with a why-statement that the exercise is recommended by an algorithm.

(c) No explanation for the NONE group, only a statement that the exercise is recommended.



(d) Exercise list: rows contain an indication of being solved, a link to the exercise, and a difficulty label (easy, average, hard).

Figure 6.1: The three explanation interfaces in our randomized controlled experiment **(a–c)**. In each interface, the top part (blue) shows real, placebo, or no explanations. The bottom part (green) allows students to return to the exercise overview **(d)**.

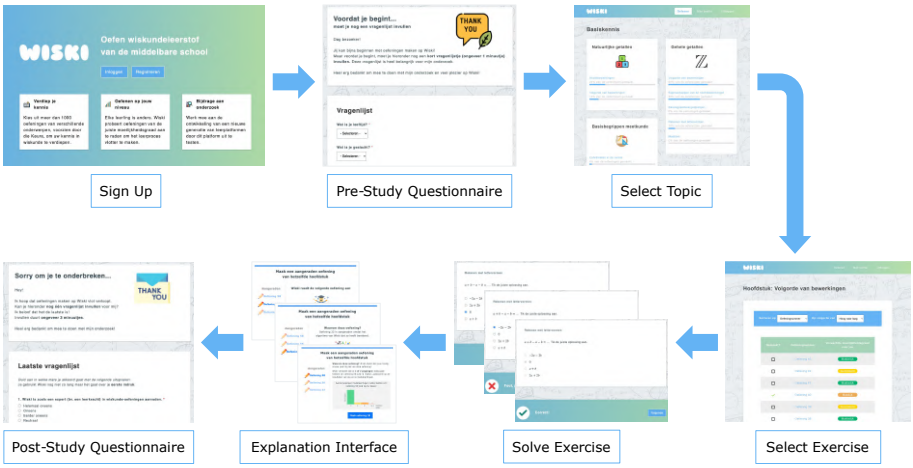


Figure 6.2: Flow chart of our study: sign up, pre-study questionnaire, solving exercises and interacting with an explanation interface five times, and post-study questionnaire.

6.3.4 Study Design

To assess the effects of our explanation interfaces on initial trust, we conducted a randomized controlled experiment (Glennerster and Takavarasha, 2013) with three research groups: *REAL*, *PLACEBO*, and *NONE*, corresponding to the explanation interfaces in Figure 6.1a to 6.1c, respectively. Following the steps in Figure 6.2, all participants (1) registered on our platform and were randomly assigned a research group; (2) answered a pre-study questionnaire with questions related to their demographics, experience with computers and e-learning platforms, mathematical background, and self-perceived mastery in mathematics; (3) solved five exercises and interacted with their research group's explanation interface after each exercise; (4) answered the post-study questionnaire in Table A.1 with questions on trust; and (5) optionally used the platform freely until the end of the study. Thus, participants' experience on our platform only differed in the explanation interface shown after solving exercises. In the background, we also logged whether participants selected recommended exercises.

We decided to let participants answer the post-study questionnaire after five exercises because (a) they then all interacted with an explanation interface equally often, and (b) they often participated during a mathematics period at school and needed to finish in under an hour. The post-study questionnaire itself contained nineteen 7-point Likert-type questions divided into seven groups (see

Table A.1). We measured trusting beliefs, consisting of *Competence* (Q1–Q5), *Benevolence* (Q6–Q8), and *Integrity* (Q9–Q11) with a validated questionnaire by Wang and Benbasat (Wang and Benbasat, 2005). To fit the original questions in the scope of Wiski, we translated them to Dutch and made them easier to understand for adolescents by simplifying some vocabulary. The average of the scores for trusting beliefs, *Intention to return* (Q13–Q14), and *Perceived transparency* (Q15) yielded a multidimensional trust score. In contrast, *Trust* (Q12) assessed one-dimensional trust by explicitly asking about trust in Wiski's recommendations. Finally, *General questions* (Q16–Q19) collected extra information about how participants perceived explanations. Furthermore, after each question group, we added a text field in which participants could motivate their Likert-type responses. In the end, we thematically analyzed these written qualitative data to gain further insights into participants' rationale for picking a specific quantitative score. Measuring trust through the above-mentioned constructs aligns with how other recommender systems are evaluated in the literature (Berkovsky et al., 2017; Chen, 2008; Cramer et al., 2008; Gedikli et al., 2014; Wang and Benbasat, 2005).

6.3.5 Statistical Analysis

We analyzed our data with Pingouin 0.3.11 (Vallat, 2018) in Python 3.8.5. We used non-parametric statistics to avoid normality assumptions, similar to other studies involving Likert-type data, e.g., (Abdi et al., 2020; Cramer et al., 2008). More specifically, we tested for significant differences between research groups with Mann-Whitney U and used Kendall's τ to test for correlations. To interpret the former as a test for difference in medians, we assumed equal data distributions in our research groups.

6.4 Results

In total, 37 students (ages 13–18, 13 male, 24 female) participated in our research: 3 students were from 9th grade, 18 from 10th grade, 8 from 11th grade, and 8 from 12th grade. Figure 6.7 shows their distribution over the three research groups: 12 in REAL, 12 in PLACEBO, and 13 in NONE. Figures 6.3 and 6.4 plot their responses to the post-study questionnaire.

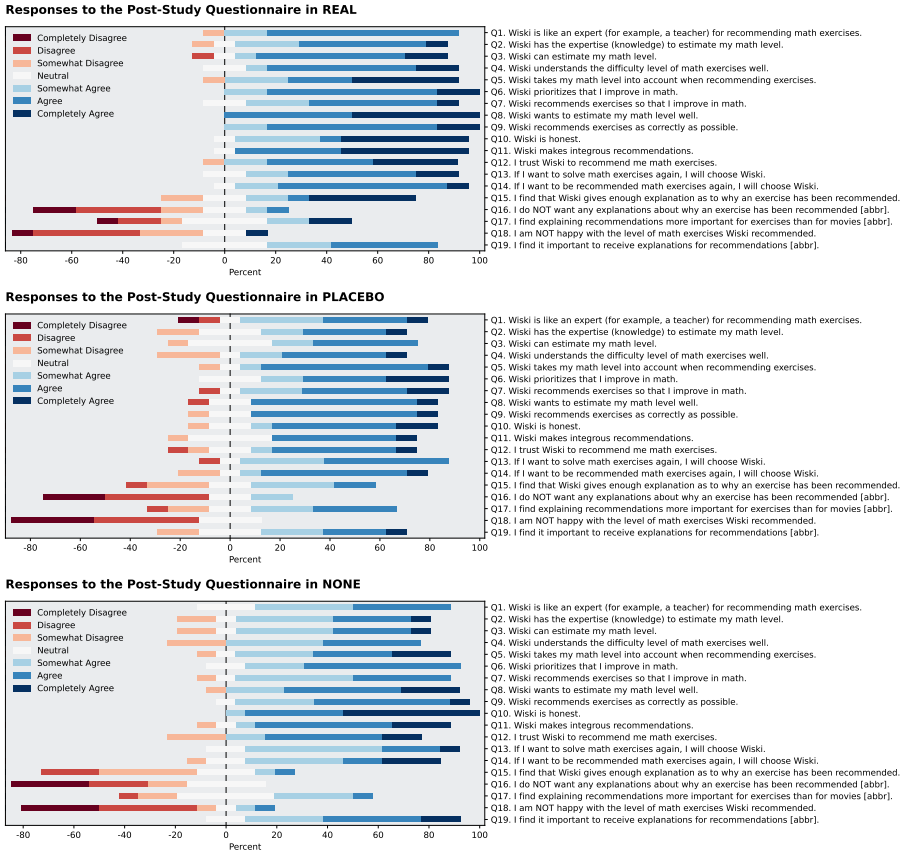


Figure 6.3: Diverging bar charts of the responses to the post-study questionnaire in Table A.1 for each research group.

6.4.1 Effects of Real Explanations

Table 6.1a and 6.1b depict the outcomes of one-sided Mann-Whitney U tests, comparing REAL to NONE, and REAL to PLACEBO. Median competence, trusting beliefs, perceived transparency, and multidimensional trust were significantly higher in REAL ($p < 0.05$). However, there was no significant increase in integrity, one-dimensional trust or intention to return. For benevolence, there was only a significant increase ($p < 0.05$) when comparing REAL to NONE.

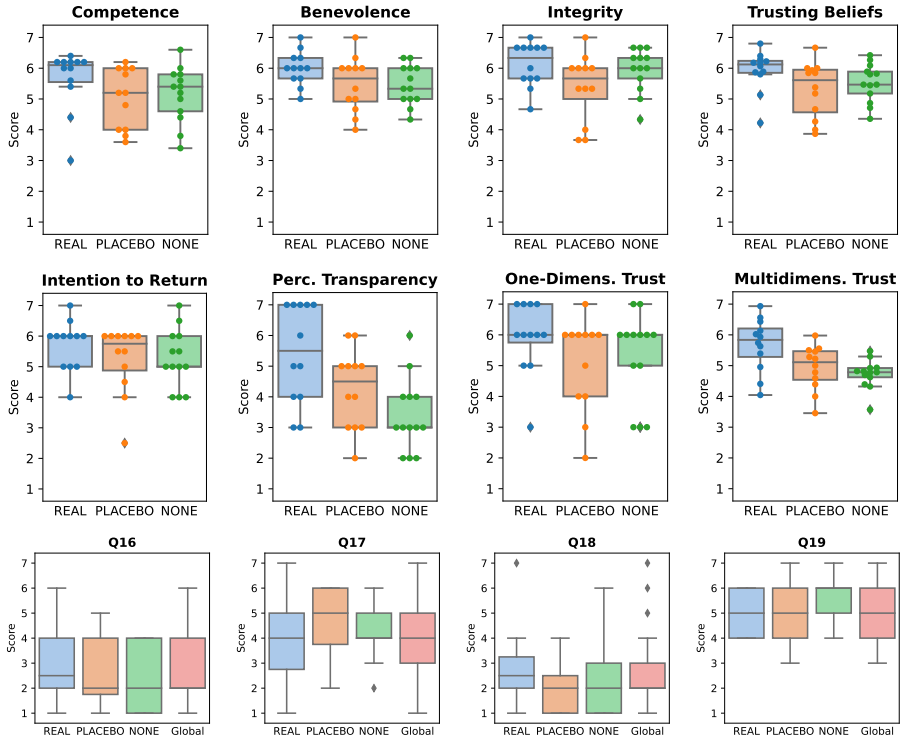


Figure 6.4: Box plots of the responses to the post-study questionnaire in Table A.1 for each research group.

The qualitative responses[2] on Q15 showed that perceived transparency was somewhat controversial in REAL. Some participants were positive about the explanations: "*I found the explanation that Wiski gave correct and satisfactory.*" Other participants did not seem to be satisfied with the explanations and may have wanted a different type of explanation: "*Doesn't it just state how many tries Wiski thinks I would need to find the correct answer. It doesn't explain specifically.*" Finally, there was also evidence that some participants did not require explanations: "*I didn't really read the explanation...*"

6.4.2 Effects of Placebo Explanations

Two-sided Mann-Whitney U tests did not reveal any significant difference ($p < 0.05$) between PLACEBO and NONE: the smallest p-values were 0.099

[2]We translated the original Dutch responses to English as literally as possible.

Table 6.1: Results of one-sided Mann-Whitney U tests comparing the research groups. The common language effect size is the probability that a random value from the first group is greater than a random value from the second group.

(a) REAL vs. NONE

	p-value	U-value	CLES
Competence	0.030*	113.0	0.724
Benevolence	0.030*	112.5	0.721
Integrity	0.261	90.0	0.577
Trusting beliefs	0.048*	109.0	0.699
Intention to return	0.109	100.5	0.644
Perceived transparency	0.002**	130.5	0.837
One-dimensional trust	0.137	97.5	0.625
Multidimensional trust	0.002**	131.0	0.840

*$p < 0.05$, **$p < 0.01$, CLES = common language effect size

(b) REAL vs. PLACEBO

	p-value	U-value	CLES
Competence	0.023*	106.5	0.740
Benevolence	0.074	97.0	0.674
Integrity	0.054	100.0	0.694
Trusting beliefs	0.026*	106.0	0.736
Intention to return	0.139	90.0	0.625
Perceived transparency	0.041*	102.0	0.708
One-dimensional trust	0.071	96.5	0.670
Multidimensional trust	0.013*	111.0	0.771

*$p < 0.05$, CLES = common language effect size

(perceived transparency) and 0.143 (integrity); all other values were above 0.327. Still, it is interesting that in our sample PLACEBO got the lowest median for competence and integrity (see Figure 6.4).

As in REAL, the qualitative responses concerning perceived transparency (Q15) showed very different sentiments in PLACEBO. On the one hand, some participants did not perceive the placebo explanations as real explanations, as seen in responses like "*Wiski just says calculated by the algorithm of . . .*" and "*It would be nice for an extensive explanation as to why it is better to solve this exercise.*" On the other hand, several participants found the explanation satisfactory, stating: "*Wiski says that the algorithm recommends the next exercise thus I trust the algorithm*" and "*I don't think that there needs to be more explanation as to why an exercise has been recommended.*"

6.4.3 Effects of No Explanations

The qualitative responses on Q15 were quite consistent within NONE: close to all participants who gave a meaningful response indicated that they did not see an explanation or missed it. For example, one participant stated: "*I find it unfortunate that [Wiski] does not say why a certain exercise was recommended. It is nice to know why this exercise fits you, but there should also not be too much information as then it would not be fun to read.*" Yet, surprisingly, two participants seemed to believe they *did* receive explanations: "*If you want to solve a new exercise, it is useful that you know why this exercise is recommended, the website does this well*" and "*Yes I find that there is enough explanation.*" Finally, one participant formed a particular mental model of our recommender system: they believed the recommendations depended on the self-reported mastery level of mathematics in the pre-study questionnaire.

6.4.4 Correlations

Figure 6.5 shows the correlations between the various trust constructs and one-dimensional trust: competence ($\tau = 0.69$) and integrity ($\tau = 0.71$) are correlated the most, whereas perceived transparency ($\tau = 0.17$) the least. In fact, perceived transparency has little to no correlation with any of the trust constructs. Figure 6.6 shows how all trust scores and questions Q16–Q19 are correlated. Especially notable is the moderate correlation between satisfaction with the level of recommended exercises (Q18) and most trust scores. We also found that one-dimensional trust is moderately correlated with trusting beliefs ($\tau = 0.68$) and multidimensional trust ($\tau = 0.52$). The latter two constructs are in their turn correlated too ($\tau = 0.56$).

6.4.5 Recommendation Clicks

Recall from Section 6.3.1 that, after participants solved an exercise about topic T, our explanation interfaces recommended three exercises to solve next. Participants could either accept one of these recommendations or ignore them and return to the exercise overview for topic T (Figure 6.1d) to select a next exercise themselves. Figure 6.8 shows that participants mostly decided to solve the first recommended exercise, followed by returning to the exercise overview. In addition, one-sided Mann-Whitney U tests revealed that the NONE group accepted significantly less recommendations than both REAL ($p = 0.007$, $U = 67$, CLES $= 0.827$) and PLACEBO ($p = 0.039$, $U = 72$, CLES $= 0.727$).

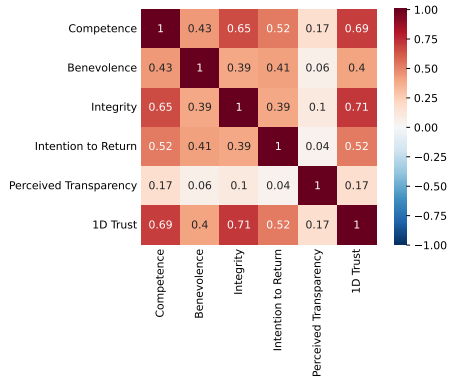


Figure 6.5: Kendall's τ correlations between trust constructs and one-dimensional trust.

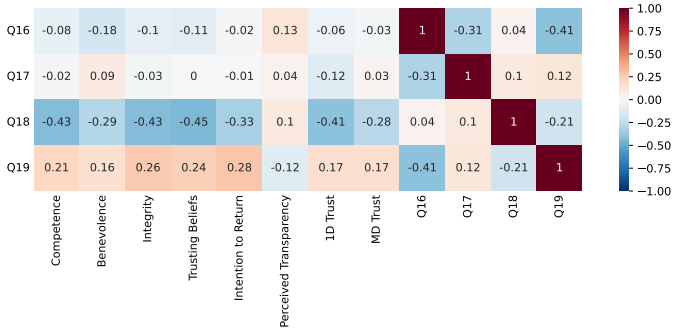


Figure 6.6: Kendall's τ correlations between trust constructs and questions on the need for explanations (Q16–Q19).

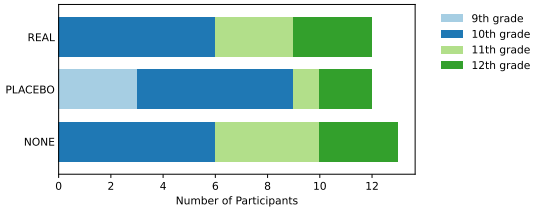


Figure 6.7: Distribution of the 37 participating students over the three research groups.

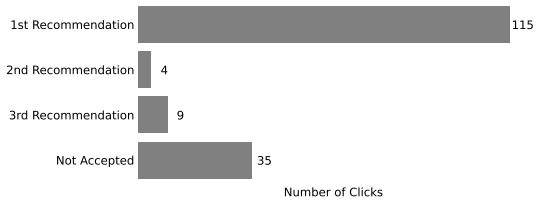


Figure 6.8: Distribution of how often each option in the explanation interface was clicked.

6.5 Discussion

This section answers our research questions by discussing how adding real, placebo, or no explanations to our e-learning platform affected adolescents' initial trust in our platform. Then, based on the observations, it underlines the need for tailoring explanations, and reflects upon the broader scope of explanations and recommendations in e-learning.

6.5.1 Explanations Increase Multidimensional Initial Trust...

Previous work has shown that well-designed explanation interfaces can increase adults' trust in recommendations (Eiband et al., 2019; Pu and Chen, 2007; Zhang and Chen, 2020). RQ1 asks whether the same holds for adolescents in an e-learning context. Two parts of our results suggest a confirmatory answer if trust is defined as an average of trusting beliefs, intention to return, and perceived transparency.

First, Table 6.1a shows that adding explanations significantly increased two out of three trust constructs: trusting beliefs and perceived transparency. The third construct, intention to return, was not significantly affected, which conflicts with the findings from Pu and Chen (Pu and Chen, 2007): they reported that higher competence perception results in higher intention to return. One possible reason for this conflict might be that Pu and Chen's explanations assisted in buying expensive products, which seems more precarious than solving recommended exercises on an e-learning platform.

Second, participants with real explanations accepted significantly more recommended exercises than participants with placebo or no explanations. Building upon the observation by Cramer et al. (2008) that acceptance of recommendations is correlated to trust, this further suggests that trust was higher for adolescents who saw real explanations.

6.5.2 . . . But Not One-Dimensional Initial Trust

However, if trust is measured one-dimensionally with a single Likert-type question, there was *no* significant increase in trust compared to using placebo or no explanations. This shows that RQ1 cannot be answered in a univocal way, and puts our findings for increased trusting beliefs and multidimensional trust into perspective. First, our results seem to imply that multidimensional trust measurements are more nuanced than their one-dimensional counterpart, which matches with the well-known statement that trust is multi-faceted and cannot be fully captured by a single question (Hoff and Bashir, 2015; Ooge and Verbert, 2021). Second, as most participants across the three research groups reported relatively high one-dimensional trust (see Figure 6.4), the explanations may not have been the most important factor for trusting the e-learning platform. Instead, participants may have built initial trust mainly because of dynamically learned factors (Hoff and Bashir, 2015) such as the perceived accuracy of the recommender system, the exercises' overall quality, or the platform's appearance. This is further backed by the correlations in Figures 6.5 and 6.6: whereas one-dimensional trust is barely correlated to perceived transparency and need for explanations (Q16, Q17, Q19), it *is* correlated to integrity, competence, and being satisfied with the exercises' level (Q18). Thus, explanations for recommendations seem to increase competence, which in turn increases initial trust. This further justifies the presence of competence in many definitions of trust (Grandison and Sloman, 2000; Muir, 1987; Wang, 2014).

6.5.3 Placebo Explanations Are a Useful Baseline

RQ2 is concerned with how placebo explanations influence adolescents' initial trust in our e-learning platform. We found no significant differences in initial trust when using placebo explanations over no explanations. This differs from results by Eiband et al. (2019), who found that placebo explanations *do* increase trust compared to no explanations. Reasons for the differing results could be the low sample size in both their and our study, the different study context, or the different methods for measuring trust. On a methodological level, Eiband et al. (2019) suggest using placebo explanations as a placeholder when insufficient information is available for real explanations. Based on our results, however, we would discourage this as it may undermine the platform's perceived transparency, competence, and integrity (see Figure 6.4 and Table 6.1b; the p-value for integrity is only slightly larger than 0.05).

However, when studying the impact of explanations, we do see several advantages for using placebo explanations as a baseline. For example, they allow to collect information about how critical participants stand towards explanations and

how attentive they are. In our study, we find it rather encouraging that most adolescents noticed that our placebo explanations were meaningless. Furthermore, combining placebo explanations and qualitative responses allows to gain insights into how much transparency participants actually need. In our study, some adolescents required a more detailed explanation while others did not require much or any transparency. This underlines the importance of research on tailoring explanations based on transparency needs.

6.5.4 Tailoring Explanations Remains Important

Our qualitative data show that not all adolescents perceived the utility and transparency of our explanation interfaces in the same way. Some adolescents even had their own perception of what a good explanation is and sought explanations that go beyond our focus on exercises' difficulty level and estimated number of attempts. To accommodate different transparency needs, it seems essential to tailor explanations to the audience that sees them.

On the one hand, the think-aloud studies in our user-centered design process gave us some insights into *what* parts of our real explanation interface may be tailored. First, middle school students (7th and 8th grade) typically found it harder to understand the histogram in our explanation, which suggests that this particular age group might require additional clarification for the histogram or an entirely different (visual) explanation. Second, some participants valued explicit wordings in the interface as it allowed them to process the given information quicker and better, while others considered this as rather redundant.

On the other hand, we can only speculate on *how* to concretize the tailoring process. One possibility is to give adolescents *direct* control over the explanations' type or detail level, or over whether they see any explanations at all. In practice, this could be done by iteratively querying students who are exposed to explanations and then modifying those explanations based on their indicated needs. A potential drawback is that incomplete or no explanations can negatively impact adolescents' mental model of the recommender system, as illustrated by the participant in our NONE group who believed that the exercise recommendation depended on their self-reported mastery in mathematics. Another possibility to tailor explanations is to *indirectly* customize them according to personal characteristics (Berkovsky et al., 2017; Millecamp et al., 2019). There is, however, an ethical challenge here as underage adolescents cannot or should not always pass delicate personality information without parental consent.

6.5.5 Taking a Step Back: Recommendations and Explanations in E-Learning

To conclude, we briefly reflect upon the premise of recommending exercises and explaining the underlying algorithm in e-learning. Do recommendations always need explanations? Should e-learning platforms always recommend exercises? We distinguish between situations in which little or much is at stake.

In *low-stakes* situations, accepting unsuitable recommendations does not have severe repercussions, so quickly accepting whichever recommendation seems reasonable. In our short-term experiment, students understood that accepting recommendations involved little risk, which may explain why they most often selected the first recommended exercise (all participants were aware of *three* recommendations in our think-aloud studies, so we assume this holds for our final study). In addition, some teachers instructed students to drill a specific topic, so it is plausible that some students were more interested in solving as many exercises as possible rather than carefully choosing their next exercise. In such 'drilling' situations, recommending only one exercise (the best fit) at a time might be sufficient, and full-fledged explanations might be excessive. However, in our experiment, students who were left in the dark as to why an exercise was recommended were more eager to select one themselves in the exercises overview. Perhaps this was the case because they perceived the displayed difficulty levels (see Figure 6.1d) as a kind of explanation. Thus, even in low-stakes contexts, it seems desirable to provide some minimal information about the (recommended) exercises.

In *high-stakes* situations, it becomes more important to investigate the benefit of recommendations, and there, we hypothesize that explanations become more important too. When students have limited time to prepare for an exam, for example, it seems plausible that they seek a justification for why they should spend time solving a recommended exercise. Regarding recommendation, we have three remarks: (1) in a school context, teachers are in the perfect position to judge which topics are best suited for a particular student, so it is interesting to study how they can steer recommendations based on their domain knowledge; (2) we believe it remains important to give students the freedom to select exercises themselves, for example to follow teachers' instructions; (3) contrary to our basic recommender system with one overall Elo score for each student, more sophisticated algorithms, e.g., (Abdi et al., 2019), could work with topic-specific Elo scores and process students' and teachers' feedback on the Elo scores to converge towards reasonable ratings more quickly.

6.5.6 Limitations and Future Work

Our research has limitations that affect the generalizability of our results. First, with only 37 participants divided over three research groups, our sample is relatively small. In addition, although we specifically focused on adolescents, the age range of 13–18 is still relatively large, especially given the turbulent stage of life that it spans. Thus, our results should be interpreted cautiously. Second, since Elo scores of students and exercises become more accurate as more students solve exercises, the accuracy of recommendations and explanations might have changed during the experiment. However, as participants were equally satisfied with the level of recommended exercises (Q18, see Figure 6.4), this should not have biased the results significantly. Third, some participants communicated that the exercises on our platform are rather basic. If solving an exercise takes an insignificant amount of time, the importance of picking a suitable recommendation becomes smaller. Future studies could thus be conducted with more challenging exercises to investigate whether our results hold. Fourth, although the post-study questions for trusting beliefs were based on those by Wang and Benbasat (Wang and Benbasat, 2005), we modified and translated them to match them to an e-learning context and adolescents. Future work can validate our questionnaire. Fifth, our short-term study could only assess initial trust, whereas trust evolves (Holliday et al., 2016; Nourani et al., 2020; Ooge and Verbert, 2021). Long-term studies could measure trust implicitly through loyalty (McNee et al., 2003; Tintarev and Masthoff, 2011). Overall, our methods and our valuable data on how adolescents trust and interact with a recommender system can be used as starting points for future research.

6.6 Conclusion

This paper tackled the complex topic of trust in an e-learning platform that explains why it recommends certain exercises. Specifically, we investigated how real and placebo explanations affect initial trust. Contrary to the vast majority of other human-computer interaction research on this topic, we focused on adolescents as the target audience.

Our randomized controlled experiment with 37 high school students showed that our explanation interface increases adolescents' initial trust when trust is measured as a multidimensional construct of trusting beliefs, intention to return, and perceived transparency. However, this effect did not hold when we considered measurements of a single Likert-type question on trust. This two-sided result seems to imply that one question cannot capture the multi-faceted nature of trust and that dynamically learned factors such as perceived accuracy

of the recommendation algorithm and the website's appearance may be the leading cause for gaining initial trust in our e-learning platform. Furthermore, compared to using no explanations, we found that placebo explanations did not offer any significant trust differences quantitatively. However, the divisive qualitative responses revealed that tailoring explanations based on transparency needs remains essential. Finally, we reflected upon whether explanations and recommendations are always desirable in e-learning, distinguishing between low- and high-stakes situations.

In sum, while our study has some limitations, our results do seem to indicate that explaining recommendations on an e-learning platform is an asset for high school students. Therefore, accompanying recommendations with explanations should be considered when designing e-learning applications similar to ours for adolescents. We also advise researchers who study the impact of tailored explanations to include placebo baselines in their studies: they may give more insights into how much transparency people actually need, compared to no-explanation baselines alone.

Acknowledgements

We are very grateful to all involved adolescents for participating in our studies, their parents for giving parental consent, and their mathematics teachers for inviting us into their (virtual) classroom. This work was supported by the Research Foundation–Flanders (FWO, grant G0A3319N) and the imec.icon project AIDA financed by Flanders Innovation & Entrepreneurship (grant HB.2020.2373).

The Human Side of Chapter 6

Collaboration and Community

Shotallo (Sho) was one of the students whose master's thesis I guided in the academic year 2020–2021. Coincidentally, when we first met in July 2020, I had just finished fine-tuning my conditionally accepted paper for CHI PLAY 2020 (Ooge et al., 2020), which was an outcome of my own master's thesis. Even cooler was Sho continued working on the e-learning platform I developed as a student. It was a tough year to conduct the research, especially because of the COVID context which, for example, made it hard to recruit participants since staff members in schools were already overworked. When I later saw the stairs ornament in the picture, I imagined it was Sho battling against the challenges he faced and I hoped my role had been similar to that of the supportive armrest of the stairs. Our collaboration taught me the value of truly and deeply working together in academia, building on each other's ideas without holding back.

Songs on repeat:

- *Blue Coloured Mountain* by Szymon
- *Breathing* by Oscar and the Wolf
- *Tough On Myself* and the rest of the *Cheap Queen (Deluxe)* album by King Princess

Stairs in the University Library of KU Leuven – September 2021

In July 2021, I met up with Sho and Kenan (another great student whose master's thesis I guided) to celebrate their successful theses. It was the first time we saw each other in person, but we connected as well as during our virtual meetings and we had a wonderful evening together. I was so proud of them! I was touched when they both thanked me with a gift. Sho gave me a beautifully enamelled plate, which is still in my living room. Every time I put my keys on it, I think about our collaboration and paper.

Brown enamelled plate – October 2023

In the period where Sho and I were writing the paper, I shot this picture because it reminded me of the fully virtual collaboration we had. The video calls we needed to rely on for months during COVID lockdowns still felt surreal to me, as if what happened through that medium took place in a parallel world. It was fun to return to that world with "virtual Sho" in January 2022 when the excellent reviews came in and we could celebrate our paper acceptance over Teams.

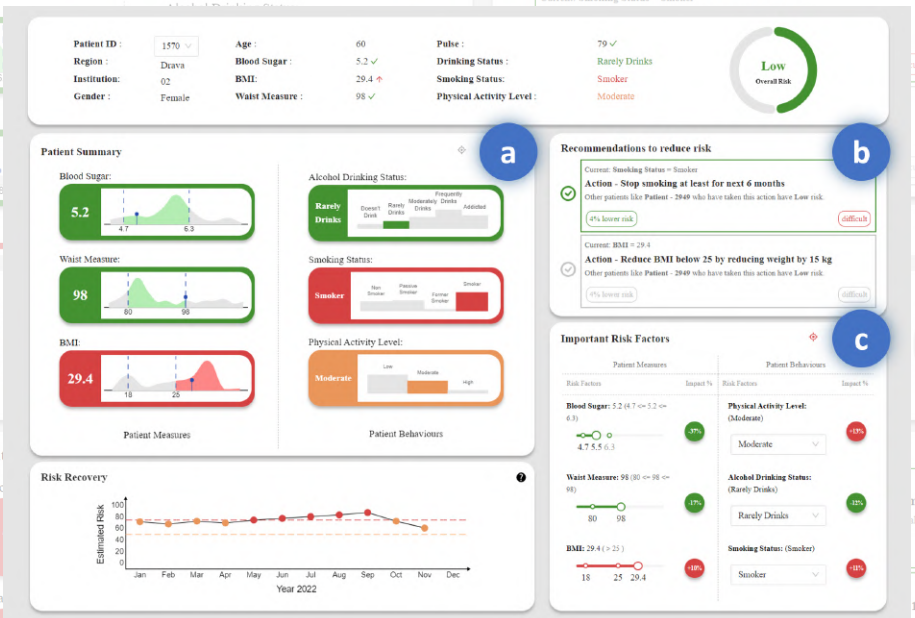
M Leuven museum – September 2021

When the IUI 2022 conference couldn't be held in person in Helsinki (Finland), I was pretty discouraged. It was the third conference in a row where I couldn't present in person and I was afraid the responses to our paper would be lukewarm because of that. However, I actually loved the virtual conference in March 2022: I felt welcomed by the community, and the supportive feedback and the many interested people were heartwarming. It was partly because of this energy that I managed to also present a poster and doctoral consortium besides the paper. To me, the photograph symbolises how a community can flourish under ominous circumstances. Note the little rain drops (sweat? tears?) on the flowers.

Songs on repeat:

- *Washing Machine Heart* by Mitski
- *Summer depression* by girl in red
- *Never Knew Love Like This Before (Single Version)* by Stephanie Mills

Gardens of the Royal Greenhouses in Laeken – May 2021



Relevant to this part of the thesis is a visual analytics dashboard by Aditya Bhattacharya (Bhattacharya et al., 2023), to which I contributed. The general idea is to assist healthcare professionals such as nurses and physicians with monitoring patients' risk of diabetes onset and recommending measures to minimise that risk (see Page 246c). Several (visualisation-supported) explanations help gain insights in the underlying prediction model:

- (a) *Data-centric explanations* visualise patients' health data and compare them to all other patients.
- (b) *Counterfactual explanations* suggest feasible actions that patients can take to reduce their predicted risk.
- (c) *Feature importance explanations* show which factors had the largest influence on the predicted risk score.

Part III

Explainability Through Visualisation-Supported Control

Chapter 7 presents a study on how control over an e-learning recommender system and visualising its impact affects adolescents' trust in an e-learning platform. This chapter was published as a conference paper (Ooge et al., 2023):

> ***Ooge, J.**, Dereu, L., and Verbert, K. (2023). Steering Recommendations and Visualising Its Impact: Effects on Adolescents' Trust in E-Learning Platforms. In* Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 156–170, New York, NY, USA. Association for Computing Machinery*

This work is the outcome of the master's thesis by Leen Dereu, which I guided intensively. As the first author, I mainly defined the research plan, helped iterate over the control mechanism and visual designs, and assisted with data analysis and interpreting the results. Leen conducted all user studies and did the implementation, whereas I did most of the writing. Moreover, I presented the paper at the IUI 2023 conference. The methods, results, and text were discussed with Katrien Verbert.

Chapter 8 presents a study on how control, *what-if* explanations, and motivational feedback affect adolescents' learning, motivation, and trust in an e-learning platform. At the time of writing, an improved version of this chapter was under review for the CHI 2024 conference.

> ***Ooge, J.***, Szymanski, M.*, Vanneste, A., and Verbert, K. (2024). Steer, See Impact, Solve: How Learner Control and Visual Explanations Impact Learning, Motivation, and Trust. Submitted to* CHI 2024.

As joint first authors, Maxwell Szymanski and I designed the enhanced control mechanism together, conducted think-aloud studies, and planned and conducted the final experiment. Arno Vanneste designed prototypes 1 and 2 with my intensive input, and we evaluated them in think-aloud studies together. I conducted the focus groups together with Bram Faems. Moreover, I implemented the final prototype and recommendation system, analysed all data, and wrote the paper. The methods, results, and text were discussed with Katrien Verbert.

Chapter 7

Steering Recommendations and Visualising Its Impact

Researchers have widely acknowledged the potential of control mechanisms with which end-users of recommender systems can better tailor recommendations. However, few e-learning environments so far incorporate such mechanisms, for example for steering recommended exercises. In addition, studies with adolescents in this context are rare. To address these limitations, we designed a control mechanism and a visualisation of the control's impact through an iterative design process with adolescents and teachers. Then, we investigated how these functionalities affect adolescents' trust in an e-learning platform that recommends maths exercises. A randomised controlled experiment with 76 middle school and high school adolescents showed that visualising the impact of exercised control significantly increases trust. Furthermore, having control over their mastery level seemed to inspire adolescents to reasonably challenge themselves and reflect upon the underlying recommendation algorithm. Finally, a significant increase in perceived transparency suggested that visualising steering actions can indirectly explain why recommendations are suitable, which opens interesting research tracks for the broader field of explainable AI.

7.1 Introduction

Recommender systems have long been actively studied to help reduce information overload in contexts where people are searching for relevant content. To better anticipate people's changing preferences and needs, researchers have increasingly

acknowledged the importance of control mechanisms with which people can actively steer recommendations (Jannach et al., 2017). Studies have shown that being able to control recommendations can increase satisfaction with, perceived understanding of, and trust in a recommender system, which can in turn increase acceptance of recommendations (Knijnenburg et al., 2012a). At the same time, too much control can overwhelm people and incur high cognitive loads (Andjelkovic et al., 2016; Bollen et al., 2010).

However, most research on controlling recommender systems is limited because of two reasons. First, studied target audiences typically consist of adults, whereas in practice younger audiences such as adolescents (ages 12–19 (Fitton et al., 2013)) are just as much, if not more, exposed to recommendation algorithms. Second, recommender systems are most often studied within contexts such as multimedia, e-commerce, and other services, and it is unclear whether findings therein always transfer to other application domains. In a high-stakes domain such as education, for example, it is crucial to properly understand the effects of control mechanisms, especially now that e-learning platforms are increasingly recommending learning content to personalise learning. Thus, it is important to design control mechanisms fit for an educational context; reflect on how much control students, teachers, and other parties should get; and find suitable ways to communicate the impact of steering.

To address these limitations, we conducted a study on how adolescents trust an e-learning platform when they can steer recommended exercises and see their control's effects. Our research questions were as follows:

RQ1. How does the ability to control recommended exercises affect students' trust in an e-learning platform?

RQ2. How is students' trust in an e-learning platform affected when they see a visual representation of their impact when controlling recommended exercises?

Our research contribution is threefold. First, we present a control mechanism and a visualisation of its impact, which have been found useful and usable by adolescents in a user-centred design process. Second, we discovered that a control mechanism does not necessarily change trust, neither when measured directly, nor when measured as a construct of competence, benevolence, integrity, intention to return, and perceived transparency. We also found, however, that a control mechanism can stimulate adolescents to reflect more upon their mastery level and the underlying recommendation system. Third, we show that visualising the control's impact can increase trust and perceived understanding

of recommendations. Additionally, we share our dataset[1] on how adolescents trust our platform and interact with our control mechanism, allowing further exploration and direct comparison in future research. In sum, our contributions highlight the potential of control mechanisms and related visualisations for adolescents in e-learning.

7.2 Background and Related Work

This section first discusses existing research on user control in recommender systems and then briefly highlights the overlap with explainable AI research, focusing on trust. Next, it zooms in on educational recommenders and relevant pedagogical background.

7.2.1 Control over Recommendations

In real-world settings, the accuracy of recommendation algorithms is subject to people's changing preferences: preference information known to the system can become outdated, leading to inaccurate recommendations (Amatriain et al., 2009). To ameliorate this problem, many control mechanisms have been developed to actively involve people in recommendation processes (Jannach et al., 2017). For example, during preference elicitation, people can exercise control through preference forms (Hijikata et al., 2012) or conversational dialogues (Göker and Thompson, 2000). In addition, after being shown recommendation results, people can steer these results through critiquing (Chen and Pu, 2012; Luo et al., 2020; Petrescu et al., 2021), dynamical filtering and re-sorting (Bostandjiev et al., 2012; O'Donovan et al., 2008), interactive (visual) explanations (He et al., 2016; Schaffer et al., 2015; Tsai and Brusilovsky, 2019b, 2021), or changing the recommendation algorithms itself (Ekstrand et al., 2015).

Yet, how much control and which control mechanisms a recommender system should incorporate depends on the context, application, and end-user (Cramer et al., 2008; Jameson and Schwarzkopf, 2002; Jin et al., 2020). Therefore, researchers have been studying the effects of providing control to end-users from different human-centred perspectives (Konstan and Riedl, 2012; Xiao and Benbasat, 2007), including perceived variety of recommendations, personal characteristics, trust, and understanding of the recommendation system. Specifically, Knijnenburg et al. (2012b) found that control can increase perceived variety of recommendations. Furthermore, preference for control methods in recommender systems depends on personal characteristics such as personality

[1] https://github.com/JeroenOoge/steering-recommendations-elearning

traits, need for cognition, and mood (Jin et al., 2020; Knijnenburg et al., 2011; Millecamp et al., 2018). Regarding trust in recommendations, control is highly valued for achieving personal goals but can also raise distrust about whether the control is just an illusion (Harambam et al., 2019). Finally, control mechanisms can increase overall system satisfaction and improve understanding of the recommendation process (Knijnenburg et al., 2012a).

7.2.2 Explainable AI and Trust

The challenge to make recommendation algorithms more transparent fits in the wider field of *explainable AI* (XAI). Essentially, XAI is an umbrella term for techniques that explain the outcomes of AI models, such that a specific audience can better understand and appropriately trust them (Barredo Arrieta et al., 2020; Guidotti et al., 2019b; Gunning and Aha, 2019; Hind, 2019). Research on these techniques brings together many concepts of interest, including fairness, privacy, bias, human reasoning, accountability, and ethics (Abdul et al., 2018).

One frequently studied concept in XAI is *trust* in automated systems (Lee and See, 2004). Some work approaches trust from an algorithmic perspective, for example by considering it equivalent to reputation in recommender systems (O'Donovan and Smyth, 2005). However, XAI more often approaches trust from a human-centred perspective. Definitions for human-AI trust are heavily debated, but most agree that trust is an attitude in a situation of vulnerability and positive expectations (Vereschak et al., 2021). Thus, from this angle, trust is a human belief that can be wrongly calibrated to the objective trustworthiness of an automated system (Han and Schulz, 2020). Besides, trust building and calibration is influenced by how a system behaves: people's trust typically fluctuates until they feel sufficiently familiar with the system (Holliday et al., 2016; Nourani et al., 2020; Yu et al., 2017a).

Given the lack of well-accepted definitions, researchers measure human-AI trust in many ways. For example, some researchers consider trust as a *one-dimensional*, i.e. monolithic, concept and typically measure it with a single Likert-type question. While some studies, e.g., (Holliday et al., 2016; Nourani et al., 2020), apply this strategy because it is quick, they are limited since a single question cannot measure a complex concept such as trust (Hoff and Bashir, 2015). Alternatively, other researchers consider trust as a *multidimensional* ensemble of several constructs which they typically measure with multiple Likert-type questions. For example, McKnight et al. (2002) introduced *trusting beliefs* as a composition of competence, benevolence, and integrity; and Ooge et al. (2022a) measured trust as the average of trusting beliefs, intention to return, and perceived transparency.

7.2.3 Educational Recommender Systems

Recommendation techniques are increasingly being integrated in digital learning environments (Khanal et al., 2020; Zhai et al., 2021). However, educational recommender systems differ from their general-purpose counterparts: they intend to facilitate achieving learning goals, are subject to a pedagogical context, and consider end-users' educational role or mastery level instead of personal characteristics (Garcia-Martinez and Hamou-Lhadj, 2013; Manouselis et al., 2013). In general, educational recommender systems can support learning in several ways (Drachsler et al., 2015). For example, they can recommend courses (Aher and Lobo, 2013; Farzan and Brusilovsky, 2011), suggest additional learning resources (Tang and McCalla, 2005), and support teachers to improve their courses or monitor their teaching resources (Gallego et al., 2013; García et al., 2009).

In the spirit of XAI for education (Khosravi et al., 2022), educational recommender systems are often requested to allow steering and to justify their recommendations. Steering could occur, for example, in the form of explicitly asking learners for feedback on exercises' difficulty after completing them (Michlík and Bieliková, 2010). Furthermore, recommendations tailored to learners' mastery level can be justified by showing how the system estimates that mastery level (Kay and Kummerfeld, 2019). In the context of open learner models (Bull and Kay, 2010; Conati et al., 2018; Hooshyar et al., 2020), Mabbott and Bull (Mabbott and Bull, 2006) found that learners felt less comfortable having full control over a learner model, compared to only making suggestions; and Abdi et al. (2020) found that an open learner model increases understanding of recommendations.

7.2.4 Estimating Mastery and Exercise Difficulty

From a pedagogical perspective, students' mastery level can be assessed based on several frameworks. One famous framework is Bloom's revised taxonomy (Krathwohl, 2002), which consists of two dimensions: a knowledge dimension with four levels (factual, conceptual, procedural, and metacognitive knowledge) and a cognitive process dimension with six levels (remember, understand, apply, analyse, evaluate, and create). Another framework is the Dreyfus model (Dreyfus, 2004), which proposes five skill acquisition levels: novice, advanced beginner, competent, proficient, and expert.

From a computer science perspective, different techniques can simultaneously estimate learners' mastery level and exercises' difficulty based on how learners perform while solving exercises (Torkamaan and Ziegler, 2022; Wauters et al.,

2012). Specialised models such as item response theory (Kadengye et al., 2015) or knowledge tracing (Guo et al., 2021), however, need to be calibrated on large item sets with known difficulties (Pelánek, 2016; Wauters et al., 2012). A classic alternative that circumvents this disadvantage is the *Elo rating system* (Pelánek, 2016), which was originally introduced by Arpad Elo (Elo, 1978) for rating chess players. Translated to education, the Elo rating system assigns dynamic ratings to both learners and exercises: the higher a learner's rating, the higher their mastery level; and the higher an exercise's rating, the more difficult it is. Furthermore, Elo ratings are of interval scale and their range can be chosen arbitrarily. Each time a learner l answers an exercise e, the Elo ratings of l and e are updated as follows:

$$\begin{aligned} \mathrm{Elo}(l) &= \mathrm{Elo}(l) + k \cdot (X_{le} - P(X_{le} = 1)) \\ \text{and} \quad \mathrm{Elo}(e) &= \mathrm{Elo}(e) - k \cdot (X_{le} - P(X_{le} = 1)), \end{aligned} \tag{7.1}$$

where k is a fixed learning-rate parameter that determines how strongly the attempt influences the Elo rating, $X_{le} \in \{0, 1\}$ reflects whether l answered e correctly, and

$$P(X_{le} = 1) = 1/\big(1 + \exp(\mathrm{Elo}(e) - \mathrm{Elo}(l))\big) \tag{7.2}$$

is the modelled probability for a correct answer. In words, whenever someone correctly solves an exercise, their Elo rating increases and the exercise's Elo rating decreases, proportional to how unexpected that correct answer was; vice versa for incorrect answers. Besides its intuitive functioning, the Elo rating system has the asset that it can be extended to multivariate settings (Abdi et al., 2019), adapted to consider how quickly students solve questions (Klinkenberg et al., 2011), and combined with other techniques such as collaborative filtering (Dahl and Fykse, 2018; Ooge et al., 2022a).

7.3 Materials and Methods

This section presents our e-learning platform and design decisions inspired by a pilot study with teachers and an iterative design process with students. Next, it describes our main study design, which was approved by the ethical committee of KU Leuven (reference number: G-2022-4917).

7.3.1 E-Learning Platform with Personalised Exercises and a Control Mechanism

We built upon *Wiski*, an existing e-learning platform for middle and high school students (Ooge, 2019). Essentially, Wiski's core functionality is solving multiple-choice questions about maths topics in the Belgian school curriculum. Through an iterative design process with students and teachers, we extended this core with three functionalities: (a) composing exercise series recommended for students' mastery level, (b) giving students partial control over their estimated mastery level, and (c) visualising the impact of that control. Think-aloud studies in which adolescents executed predefined tasks on a low-fidelity version of our e-learning platform ensured that these new functionalities were deemed useful and usable.

Personalised exercise series Brief semi-structured interviews (Leech, 2002) with 4 high school teachers learned us that teachers appreciated the idea of an e-learning platform that recommends exercises tailored to students' mastery level. In addition, to give students sufficient time to adapt to new difficulty levels, teachers advised recommending exercise *series* instead of individual exercises. We therefore decided to let our platform estimate students' mastery level and exercises' difficulty with an Elo rating system and then use those estimates to recommend exercise series. Specifically, whenever a student l would select a topic to practise, they would start a series consisting of two exercises, e_1 and e_2, chosen such that $P(X_{le_1} = 1)$ and $P(X_{le_2} = 1)$ were closest to 0.7; a value yielding reasonably challenging exercises (Klinkenberg et al., 2011). Probabilities were estimated with a variant of (7.2), which originates from a chess context:

$$P(X_{le} = 1) = 1/\big(1 + 10^{(\mathrm{Elo}(e) - \mathrm{Elo}(l))/400}\big).$$

To set up our Elo rating system, students could initialise their Elo rating with the slider in Figure 7.1, which indicated five thresholds inspired by the Dreyfus model (Dreyfus, 2004). In the background, the slider's range corresponded to the interval $[1000, 2000]$, which roughly corresponds to typical Elo scores for novice (1000) and expert (2000) chess players. Furthermore, exercises' initial Elo ratings were set by teachers who participated in our main study. Concretely, teachers used the thresholds in Figure 7.1 to estimate the difficulty of all exercises belonging to the subjects they wanted to cover in class. In case multiple teachers were interested in the same subjects, we only asked one of them to set the initial ratings, distributing the workload evenly. Finally, we set the hyperparameter k in (7.1) to 160 to allow for relatively large Elo changes.

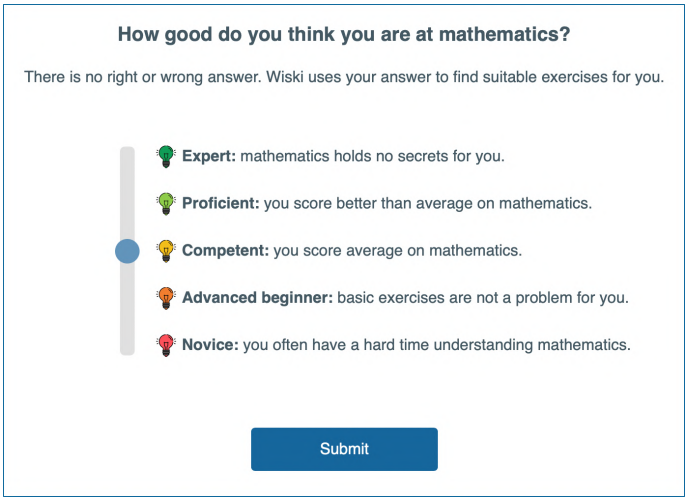


Figure 7.1: Students initialised their maths mastery level with a continuous slider that indicated five thresholds: novice, advanced beginner, competent, proficient, and expert.

Control mechanism and impact visualisation Through two rounds of think-aloud studies with 11 adolescents (2 middle school, 9 high school), we iteratively designed a control mechanism and a visualisation of the exercised control's impact. First, the control mechanism in Figure 7.2 allowed students to modify their mastery level and thus steer the difficulty of subsequent recommendations. Specifically, after finishing an exercise series, students could indicate whether they wanted easier or harder exercises. In the background, this would lower or raise their Elo rating up to 10%, respectively. The think-alouds learned us that the slider provided intuitive and sufficient control. In addition, adolescents preferred to reflect in terms of their mastery level and were sometimes confused by a preliminary design that also allowed them to steer exercises' Elo ratings with a similar slider. Second, the visualisation of the control's impact in Figure 7.3 contained three parts: a fixed explanation; a description of how mastery level changed due to solving an exercise series and subsequent steering; and a line chart of the latter information. The think-alouds learned us that adolescents preferred the line chart over an animated bar chart. More details on our iterative designs can be found in Dereu's master's thesis (Dereu, 2022).

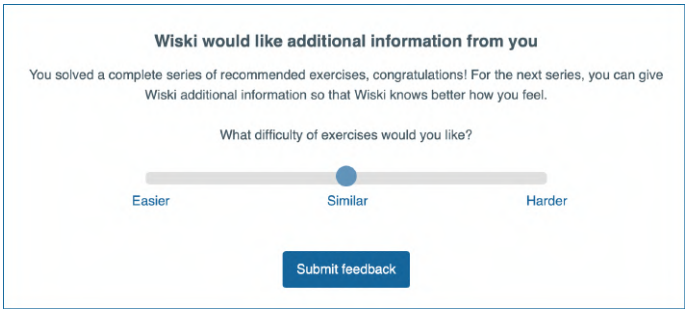


Figure 7.2: After each exercise series, students could steer subsequent recommendations with a 20-step slider: lowering their mastery level yielded easier series, and vice versa.

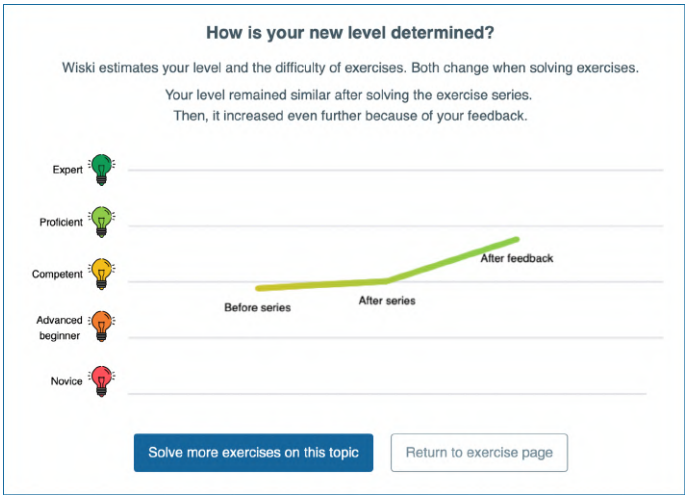


Figure 7.3: Visualisation of students' steering impact after an exercise series. The top describes the evolution of students' mastery level; the bottom visualises it.

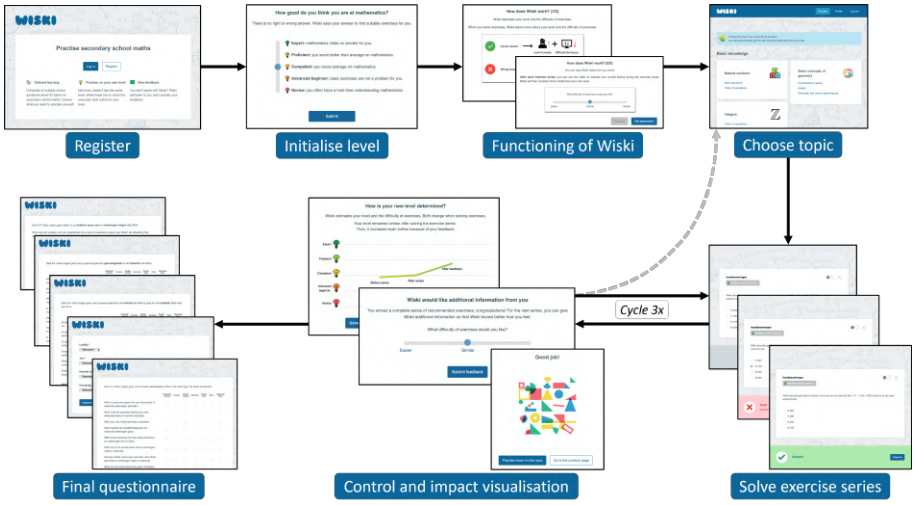


Figure 7.4: Flow of our study: registering, picking an initial mastery level, reading a global explanation on the functioning of Wiski, solving three series (i.e., six exercises) potentially followed by steering one's mastery level and seeing its impact, and finally filling out a questionnaire.

7.3.2 Study Design

To answer our research questions, we conducted a randomised controlled experiment with three groups: in NONE, participants did not have any control over recommended exercises; in CONTROL, participants could steer their mastery level with the slider in Figure 7.2; and in CONTROL+IMPACT, participants additionally saw the visualisation of their control's impact in Figure 7.3. The flow of our study is depicted in Figure 7.4. First, participants registered on our Wiski platform and were randomly assigned to one of the three research groups. Then, they initialised their maths mastery level with the slider in Figure 7.1 and saw one or two of the screens in Figure 7.5 which globally explained Wiski's recommendation algorithm. Next, participants chose a maths topic on the practice page and solved three series, each consisting of two exercises. We chose this relatively low number of series to ensure participants could finish the study in under 50 minutes. After each series, participants could adjust their mastery level and see its impact, depending on their research group. Finally, participants filled out a questionnaire and could continue to freely use the platform. Thus, participants' experience with Wiski only differed in whether or not they could control their mastery level and see a visualisation of their control's impact. In the background, we logged all Elo rating changes.

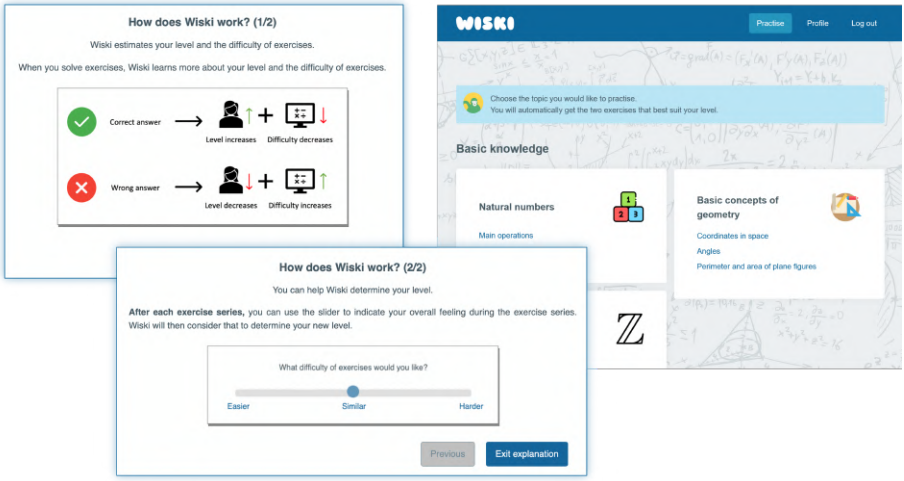


Figure 7.5: Wiski explained in two ways how it personalises exercise series. **(l)** After registration, all participants saw a global explanation; participants in CONTROL and CONTROL+IMPACT saw an additional screen. **(r)** The practice page for picking maths topics explained recommendations: *"You will automatically get the two exercises that best suit your level."*

Our final questionnaire contained the 31 Likert-type questions in Table A.2, scored on a 7-point range. The first part captured trust. Similar to Ooge et al. (2022a), we measured trust both with a single question and as the average of trusting beliefs, intention to return, and transparency. Slightly different is that, for more reliable scores, we measured transparency with three questions from the ResQue questionnaire (Pu and Chen, 2010) instead of one. The second part of our questionnaire, also based on ResQue (Pu and Chen, 2010; Pu et al., 2011), captured three control aspects: overall control, preference elicitation, and preference revision.

Our questionnaire also contained open text fields that encouraged participants to elaborate on their Likert-type responses. Furthermore, we explicitly asked participants whether they trusted our platform for recommending maths exercises, whether they (would have) liked controlling the desired difficulty level of exercises, and whether they (would have) liked seeing the impact of that control. Only the open question on trust was mandatory and the latter two questions included screenshots similar to Figures 7.2 and 7.3.

7.3.3 Participant Recruitment

We contacted 30 secondary school teachers in Belgium (Flanders) via email and LinkedIn, inviting them and their students to participate in our research during school hours. We asked teachers to not coerce students into participating and to prepare exercises on paper should some students refuse to participate. Four teachers accepted our invitation: they passed through a brochure to students and their respective parents, which communicated our study goals, data management, and Covid-19 precautions. Interested students gave informed consent and students under 16 required parental consent. Ultimately, all 76 invited students (ages 12–17) participated in the study. We excluded 5 participants from the analysis due to incomplete questionnaires, ending up with 22 participants in NONE, 25 in CONTROL, and 24 in CONTROL+IMPACT.

7.3.4 Data Analysis

We analysed the collected quantitative data in R 4.2.1. To compare the three research groups in terms of trust and control perceptions, measured as an *average* of several Likert-type questions, we first conducted one-way ANOVA tests after checking the requirements: independence was guaranteed by the randomised set-up, assuming a normal distribution was plausible given the central limit theorem, and equal variances were verified with F-tests. Then, constructs that differed in at least two groups ($p < 0.10$) were compared in more detail with unpaired t-tests, which assume the same as ANOVA. To compare trust measured with a *single* Likert-type question, we applied Mann-Whitney U tests to avoid normality assumptions. All t-tests and Mann-Whitney U tests used $p < 0.05$ as threshold for significance and were one-sided with alternative hypothesis that groups with more functionalities score higher.

To get further insights into differences between research groups, we thematically analysed (Braun and Clarke, 2012) the qualitative feedback stemming from the open questions in our questionnaire. For presentation here, we translated the original Dutch responses to English, only correcting grammar and spelling.

7.4 Results

Figure 7.6 shows the number of participants per grade, equally distributed over the three research groups. To get a detailed understanding of how participants filled out the Likert-type questions, Figure 7.7 depicts the distribution of responses in each research group. In turn, Figure 7.8 gives a more aggregated

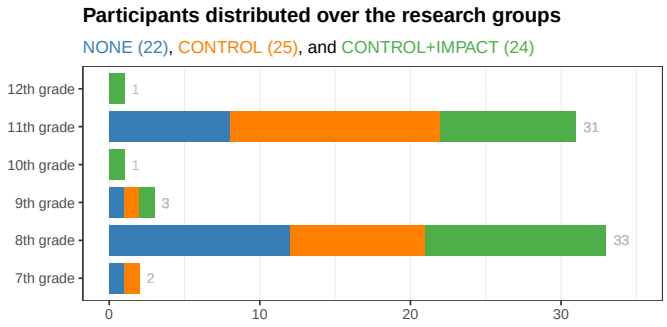


Figure 7.6: Participants distributed over the research groups per grade: most were in 8th and 11th grade.

view of participants' responses per research group. Recall that multidimensional trust is the average of trusting beliefs, intention to return, and transparency. Overall, the median scores of all measured constructs lay between neutral and rather agree. ANOVA tests found that competence, integrity, intention to return, control, and preference elicitation did not differ significantly in the three research groups ($p > 0.20$).

7.4.1 Effects Without Control or Seeing Its Impact

Qualitative responses confirmed that most participants in NONE trusted the platform overall. Over one third of the participants seemed to have based their trust on the platform's design and utility: they found that *"the website looked professional,"* was *"good for practising for tests,"* was *"a good way to practise maths to improve,"* and seemed to contain exercises that *"fit well to the subject matter."* Furthermore, two participants believed the platform was developed by teachers or experts. Another third of the participants commented on whether exercises had a suitable difficulty level. In case they found exercises well-tailored, participants appeared trusting, for example, *"The website looks [...] trustworthy. I also have the feeling that the exercises are of a good level."* Conversely, a few participants appeared distrusting or hesitant because they *"often got the same questions they had already answered correctly before,"* which gave them the feeling their mastery level stagnated and they could memorise answers. Finally, four participants alluded to potentially different trust perceptions in the long term: *"I have not been able to practise and use the site enough, so I cannot give a good final assessment either (at the moment)."*

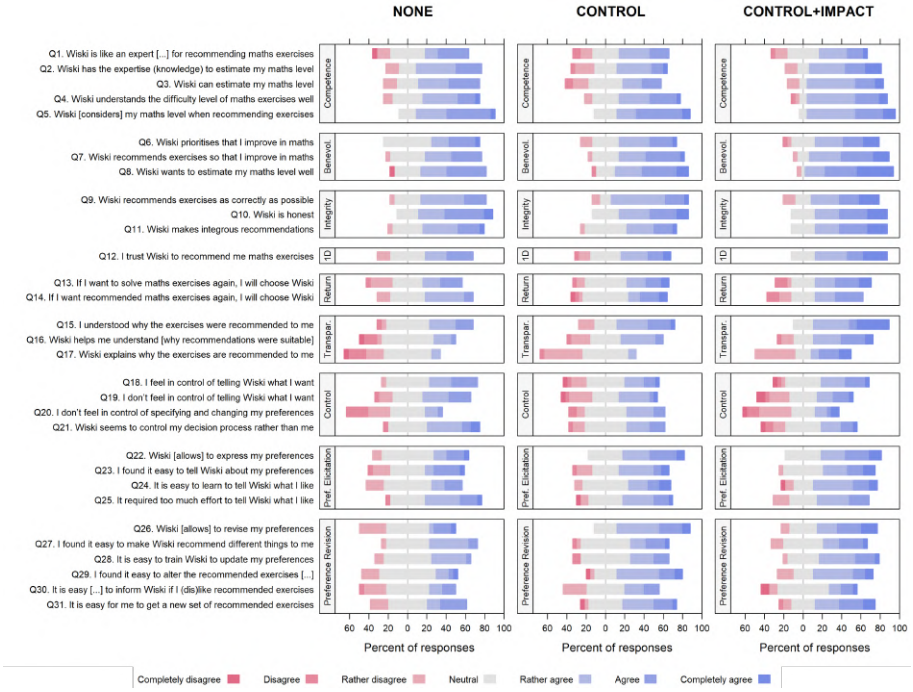


Figure 7.7: Diverging bar charts (Heiberger and Robbins, 2014) of responses to the questionnaire in Table A.2 after reverse-scoring, comparing the three research groups. Questions have been abbreviated for brevity and have been grouped per construct for clarity.

Thirteen participants in NONE commented on obtaining control over recommended exercises. Apart from one indifferent individual, all of them were in favour of extra control. Only three, however, clarified why: *"This allows you to give a bit of direction to what exercises you want yourself. Also, if you perform a bit less well, you still get some more difficult exercises to see what they entail."*

7.4.2 Effects of Controlling Recommendations

The first column in Table 7.1 shows that one-sided tests did not reveal statistical differences between NONE and CONTROL ($p < 0.05$). Thus, our sample did not provide evidence against equal means for any measured construct. Only transparency and preference revision were borderline non-significant.

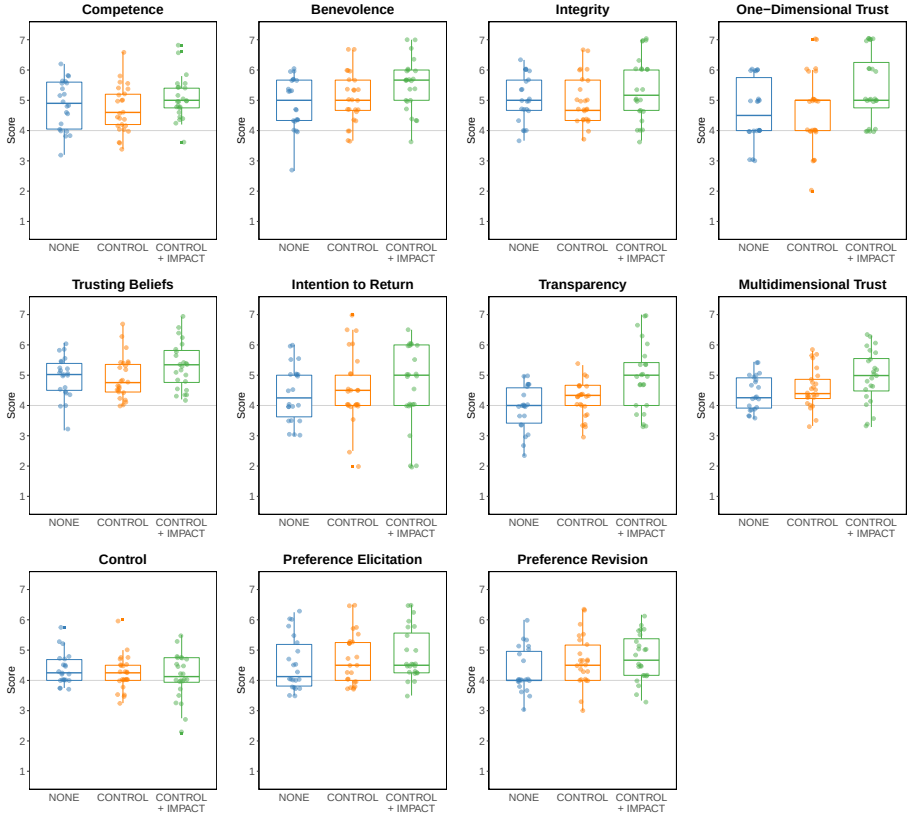


Figure 7.8: Box plots of the responses to the questionnaire in Table A.2 for each research group. For visual clarity, the overlaying dot plots are slightly jittered horizontally and vertically.

The qualitative responses on trust showed that two thirds of the participants in CONTROL seemed trusting and mostly supported that perception by the platform's ability to tailor exercises: *"It seems reliable at first sight and it also asks good questions adapted to your maths level"*; *"It can assess your level and provide further exercises to raise your level"*; and *"[I trust Wiski] if you can enter your own level."* Furthermore, similar to the responses in NONE, some participants referred to the platform's *"professional"* design and utility to *"learn something new."* In addition, two participants mentioned repeatedly occurring exercises but did not seem troubled by that: *"Wiski knows when I have some difficulties with exercises and when I don't. That's why difficult exercises are recommended again."*

Table 7.1: Comparing the research groups with t-tests (Mann-Whitney U test for one-dimensional trust). Cells contain the effect sizes (second group mean minus first group mean).

	NONE **vs.** CONTROL	NONE **vs.** CONTROL+IMPACT	CONTROL **vs.** CONTROL+IMPACT
Benevolence	0.16 ($p = 0.263$)	0.61 ($p = 0.011$)	0.45 ($p = 0.035$)
Trusting beliefs	-0.01 ($p = 0.529$)	0.38 ($p = 0.042$)	0.40 ($p = 0.030$)
Transparency	0.29 ($p = 0.068$)	1.04 ($p = 0.000$)**	0.74 ($p = 0.002$)*
One-dimens. trust	0.00 ($p = 0.504$)	0.78 ($p = 0.017$)	0.78 ($p = 0.020$)
Multidimens. trust	0.15 ($p = 0.207$)	0.55 ($p = 0.009$)*	0.40 ($p = 0.039$)
Preference revision	0.33 ($p = 0.080$)	0.43 ($p = 0.030$)	0.10 ($p = 0.325$)

*$p < 0.01$, **$p < 0.001$, non-significant results ($p \geq 0.5$) are greyed out

There were, however, also mixed trusting sentiments: while six participants did see benefits in our platform for casual practice, they hesitated to blindly adopt it in the long term for two reasons. First, some were bothered by the algorithmic nature of recommendations: *"It's a programme and not a teacher so I don't quite trust it"* and *"[It's] just an AI [...]. Wiski is good but I'd rather seek advice from a physical person."* Two quotes might explain this sentiment: *"It remains a computer system that can always be flawed"* and *"It only has a limited view of my maths skills."* Second, practice in the context of preparing tests or exams might require the presence of a teacher: *"Sometimes teachers have their own way of asking questions and this may not always match the exercises offered by Wiski."*

Furthermore, all respondents in CONTROL and CONTROL+IMPACT were very positive about the feature to control recommendations. The ability to modify the difficulty level of recommended exercises was especially appreciated to not *"get stuck"* when *"you find the exercises too difficult or too easy"* and when *"you want to try something harder but also go for something easy once in a while."* Yet, one participant noted that while *"the slider is nice to make small adjustments, it's not convenient to specifically choose a new level because [they] wanted to go up 1 level in difficulty and went up 2 levels,"* alluding to the five mastery levels depicted in Figures 7.1 and 7.3. Someone else agreed that it was indeed *"difficult to find the perfect level."* Finally, one participant admitted they were *"not sure whether [Wiski] understood [they] wanted slightly more difficult exercises"* when using the slider.

7.4.3 Effects of Visualising the Impact of Control

The second and third columns in Table 7.1 show the results of comparing NONE to CONTROL+IMPACT, and CONTROL to CONTROL+IMPACT, respectively. Both one-dimensional trust and multidimensional trust increased significantly ($p < 0.05$). The latter relates to an increase in two of its components: trusting beliefs and transparency. First, trusting beliefs increased due to higher perceived benevolence. Second, participants perceived the platform as significantly more transparent, with the average score in CONTROL+IMPACT lying 1 point higher than in NONE. Regarding control, however, only preference revision was deemed significantly higher in CONTROL+IMPACT, compared to NONE.

In CONTROL+IMPACT, most qualitative responses regarding trust were positive. Similar to CONTROL, two thirds of the respondents focused on how well exercises were tailored. Most of these participants trusted the platform and highlighted that exercises were well-tailored: *"I think Wiski does give exercises at my level. It's nice that when you get a lot of exercises right, you get more difficult exercises to challenge yourself. You notice that they get harder, therefore I trust the recommended exercises"* and *"[It's] handy that this platform can estimate your level, the exercises recommended by Wiski are therefore well fit."* Yet, three participants were rather distrustful because exercises seemed ill-tailored or repetitive to them: *"I have now made some exercises and have not yet found the level that suits me. So I am more inclined to make exercises in my textbook because I know we should be able to achieve that level."* Other participants seemed to prefer consulting a teacher or using Wiski only for supplementary exercises: *"I think Wiski is well-made and does its best to help but I don't think it can really determine my maths level."* Finally, two participants touched upon long-term trust: *"It's hard to say whether I fully trust it after just a few exercises."*

Few participants in CONTROL+IMPACT commented on the feature to see their control's impact, yet those who did found it useful to see their evolution and current level. In contrast, in NONE and CONTROL together, most participants commented on whether they would have liked a screen similar to Figure 7.3; all but one would. Many comments tapped into seeing and understanding one's current mastery level: *"This can be useful in several ways to see why you are at a certain level"* and *"Then you can see how well some exercises go."* One participant wrote: *"That's pretty handy to see how bad you are at maths."* Another frequent related theme was the possibility to see one's evolution: *"That would be useful because then you know how you are progressing."* Finally, one participant brought up motivation, stating *"I think this could also be motivating."*

7.4.4 Correlations

Figure 7.9 shows the relations between all measured trust-related and control-related constructs. Regarding the trust-related constructs, we found that competence, benevolence and integrity were moderately correlated to one another and were equally correlated with one-dimensional trust (all around $r = 0.60$). Furthermore, intention to return turned out to be most correlated to competence ($r = 0.57$). Regarding the control-related constructs, preference elicitation and preference revision correlated strongly ($r = 0.68$), but were barely correlated to control. In fact, control had little to no linear relationship with any of the constructs. Finally, the most correlated pair of trust-related and control-related constructs consisted of transparency and preference revision ($r = 0.52$), which is still relatively low as one construct explains only 25% of the variance in the other.

7.4.5 Elo Ratings

Figure 7.10 shows how participants' Elo ratings evolved during the experiment. In all research groups, the ratings gradually increased and finally participants in NONE had a lower average increase (58) than participants in CONTROL (101) and CONTROL+IMPACT (135). Yet, the trends and Figure 7.11 also show that participants in CONTROL and CONTROL+IMPACT most often increased their mastery level further after completing an exercise series. Ignoring Elo changes due to control, the average Elo increased with 60 in CONTROL and 98 in CONTROL+IMPACT. According to one-sided t-tests, however, these average Elo growths were not significantly larger than in NONE ($p = 0.132$). Figure 7.11 furthermore shows that participants used the control mechanism reasonably: most dots are in the top right quadrant, indicating that participants often further increased their mastery level after a successful exercise series; the left quadrants show that participants rarely boosted their mastery level after an unsuccessful exercise series and instead kept or downgraded it.

7.5 Discussion

This section interprets our results and answers our research questions. Based on our findings, we reflect upon implications for the explainable AI field and real-world e-learning platforms.

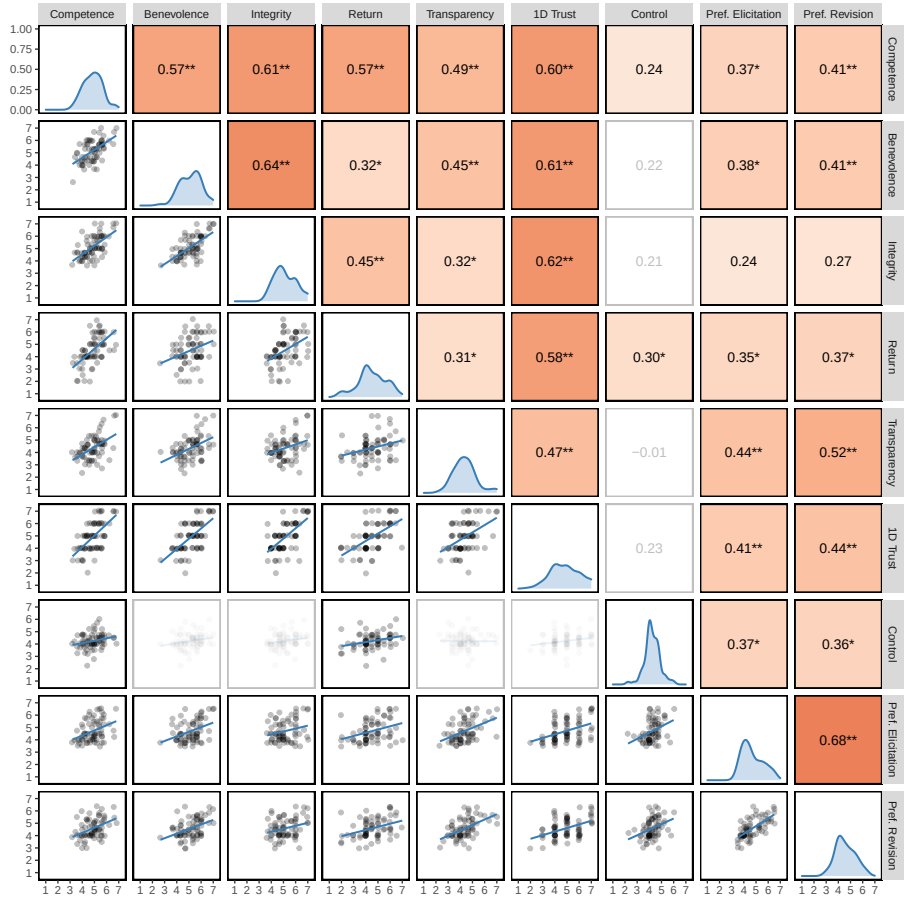


Figure 7.9: Relations between trust-related and control-related constructs. Lower triangle: dot plots with fitted regression lines. Diagonal: density plot of constructs. Upper triangle: correlations colour-coded by value (*$p < 0.01$, **$p < 0.001$). Non-significant relations ($p \geq 0.05$) are greyed out.

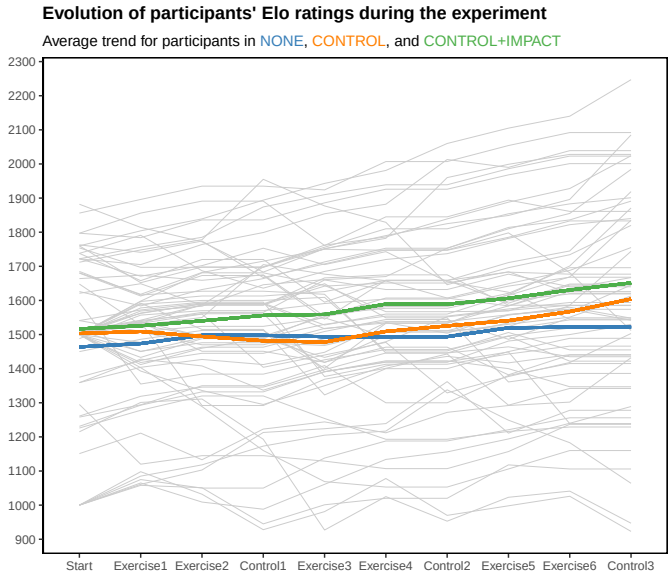


Figure 7.10: Evolution of participants' Elo ratings during the experiment and the average evolution per research group.

7.5.1 Sanity Check for Responses About Control

Before interpreting our results, we take a closer look at the findings regarding control. First, the quantitative results showed no higher sense of control in research groups with control, compared to the baseline without control. This unexpected result could be due to the measurement instrument rather than an actual lack of control: Figure 7.7 shows quite polarised responses on Q18–Q21, indicating that the questions may have been interpreted differently because they were too broad. In addition, the qualitative data confirmed that participants in CONTROL and CONTROL+IMPACT were very aware of the control mechanism and Figure 7.11 shows that they often used it. Second, preference elicitation was perceived equally amongst the three research groups. This was expected as participants could only indicate initial preferences by setting their initial mastery level and choosing maths topics. Third, also as expected, preference revision increased (almost) significantly when the control mechanism was added, but not when the control's impact was visualised. These observations support the sanity of our results.

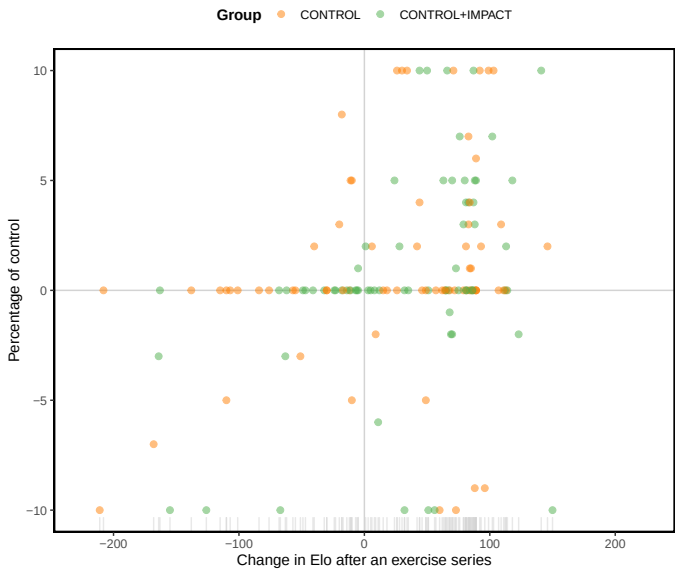


Figure 7.11: Elo changes after an exercise series compared with the control percentage chosen via the slider in Figure 7.2 (Easier = −10% and Harder = 10%).

7.5.2 Control Does Not Affect Trust but Stimulates Self-Reflection

RQ1 was concerned with how a control mechanism affects trust in our e-learning platform. Table 7.1 contains no evidence for significant effects on any of the measured trust components; only perceived transparency was borderline. The fact that one-dimensional trust did not differ significantly in NONE and CONTROL suggests participants did not consider the control mechanism a major factor for calibrating their trust.

However, the qualitative responses on trust interestingly revealed more self-reflection. Specifically, while participants in NONE most often described the platform's utility and design while discussing trust, participants who could control recommendations spontaneously referred twice as much to whether exercises were tailored to their personal mastery level. Some participants even reflected on the recommendation algorithm itself, questioning whether it was as competent as teachers. Thus, our qualitative findings suggest that control mechanisms similar to ours foster awareness of an underlying manipulable

algorithm. Growing such awareness seems very valuable in a world that becomes permeated by applications relying on algorithmic decision-making, so future experiments could investigate whether and why this effect holds in larger samples. One plausible explanation could be that controlling mechanisms are uncommon in current e-learning platforms and therefore caught participants' attention.

7.5.3 Seeing the Impact of Control Grows Trust

RQ2 asked how visualising the impact of control influences adolescents' trust in our e-learning platform. Our results showed a significant increase in one-dimensional trust, which suggests that the visualisation played a big role in growing trust. Multidimensional trust also increased significantly, partly due to a higher perceived benevolence. This could be explained by the following observation: Figures 7.10 and 7.11 show that most exercise series led to an increase in Elo rating, which implies that participants in CONTROL+IMPACT mostly saw increasing mastery evolutions. Thus, it seems plausible that participants who saw the visualisation considered our platform as more benevolent than participants who did not.

7.5.4 Visualising the Impact of Control is a Kind of Explanation

The most heavily changed trusting component was transparency: participants who saw their control's impact visualised considered the recommendations as more transparent. This suggests that participants experienced the visualisation as a sort of explanation. However, at first sight, it is not entirely clear *what* part of the recommendation process this explanation clarified for them. Comparing the responses for Q15–Q17 in Figure 7.7, we observe that roughly half of the responses for Q17 were negative, whereas most responses for Q15 and Q16 were positive. This seems to imply that participants did not view the visualisation of their control's impact as a *direct* explanation for why they received specific recommendations; which they indeed should not have. Rather, the visualisation arguably acted as an *indirect* explanation: participants felt they had a better understanding of why recommendations were suitable for them because they could repeatedly see how the e-learning platform estimated and modified their mastery level. Overall, visualising the impact of control seems to have reinforced participants' mental model (Johnson-Laird, 1983; Kulesza et al., 2012) of the recommendation system by gradually clarifying the behaviour of a crucial component of the recommender, namely iterative estimation of learners' mastery level.

7.5.5 Implications for Explainable AI Research

Our findings potentially have interesting research implications for the broader field of explainable AI. First, visual explanations intended for a lay audience may not need to explain complete algorithms in detail. Instead, explaining crucial components could suffice when complemented with a global reasoning rationale of the algorithm. In our case, this global reasoning rationale was provided as a simple sentence on the practice page (*"You will automatically get the two exercises that best suit your level"* in Figure 7.5). Another provoking idea is that control by itself could increase transparency. This turned out to be the case in our sample, although the increase was borderline non-significant. We hypothesise that on our platform the combination of exercising control and seeing the difficulty level of subsequently recommended exercises acted as a kind of *model inspection* (Guidotti et al., 2019b). In other words, participants could steer the recommendation algorithm and then see the impact on outcomes of the recommendation algorithm. If future research could confirm our hypothesis, this might be one of the earliest examples of effective model inspection with adolescents. Third, the qualitative responses regarding motivation open up research tracks on whether (visual) explanations can inherently motivate students to, for example, practise more, challenge themselves more, or – being hopeful – even appreciate maths more as a whole.

7.5.6 Taking a Step Back: Technology-Enhanced Learning and Control

Before we conclude, we briefly reflect upon control in e-learning. How much control should students get and does that imply taking control away from teachers? Should students always see their control's impact with respect to their mastery level?

Overall, students received our platform's control mechanism enthusiastically and seemed to have used it reasonably. The faster increase in Elo rating for students with control also suggests that the control mechanism allowed them to more quickly converge towards exercises with difficulty levels that best suited them. Moreover, the control mechanism and its accompanying impact visualisation seemed to have prompted students to think more consciously about which difficulty levels they could handle and how their mastery level changed. This is an important metacognitive skill, which is crucial in self-regulated learning (Zimmerman, 1990). For these reasons, we believe giving control to students can be an asset for e-learning platforms.

Yet, we see at least two nuances. First, giving too much control to students can be disadvantageous when it causes discomfort because of the responsibility it entails (Mabbott and Bull, 2006). In addition, students could abuse control over their mastery level in evaluative contexts: artificially decreasing their mastery level could allow them to obtain higher success rates when solving exercises, and artificially increasing it could trick inattentive teachers into overestimating their abilities. Thus, it is important to balance the amount of control with factors such as pedagogical responsibility and the use context and to not overly rely on Elo ratings for evaluation purposes. Second, providing students with control does not make teachers redundant. In our study, participants highlighted the still valuable role of teachers: providing extra feedback on students' progress, and verifying that exercises on the e-learning system are aligned with the curriculum and their usual style of interrogating. Furthermore, by monitoring or adapting students' mastery levels, teachers could additionally guide students who under- or overestimate themselves because of the Dunning-Kruger effect (Dunning, 2011).

Our visualisation of how control affected mastery level, and thus recommended exercises, was well-received too. However, some comments regarding motivation made us realise the potentially demotivating effects of frequently showing downward evolutions in students' mastery levels. Therefore, we argue that visualisations related to mastery level should be shown sufficiently infrequent to avoid potential negative motivational effects, yet frequently enough to allow intervention in case of learning issues. Such interventions could be facilitated by e-learning platforms in the form of alerts that inform students when it seems advisable they ask teachers for additional support. In line with teachers' desires in our pilot study, those alerts could also be shown to teachers so they can intervene, similar to existing work in the learning analytics community (Akçapınar et al., 2019; Denden et al., 2019).

7.5.7 Limitations and Future Work

Our research has several limitations which restrict how well our findings generalise. First, our sample was relatively small so some findings may not hold in larger studies and we could not investigate differences between age groups. Although we controlled for multiple testing by only conducting t-tests when ANOVA indicated a group-wise difference, false positive differences could remain. Second, since our study was not focused on developing a highly accurate recommender system, we generated recommendations with a simple Elo rating system. More sophisticated algorithms such as multivariate Elo-based models (Abdi et al., 2019) or knowledge tracing (Guo et al., 2021) could be considered for platforms deployed in the real world, especially because

competence is rather highly correlated to intention to return (see Figure 7.9). Third, the mechanism to steer recommendations was quite simple and only affected recommended exercises indirectly by altering mastery levels. Future studies with adolescents in e-learning could further study more advanced control mechanisms that affect recommendations directly, for example steering through interactive visualisations. Fourth, as our study was conducted in a class context, it is possible that some students noticed that their peers were shown a different version of our platform. Although we did not observe copying during the study, we are wary of adolescents' resourcefulness to copy and the bias it may have entailed. Fifth, as some participants indicated, our study was restricted to capturing trust while participants were arguably still familiarising themselves with the recommender and control mechanism. In this *learning phase* (Yu et al., 2017a), trust perceptions can change briskly, for example due to encountering unexpected behaviour such as repeated recommendations (Holliday et al., 2016; Nourani et al., 2020; Yu et al., 2017a). Thus, as briefly using our platform might have hampered reliable long-term trust assessment, our results should be interpreted cautiously. Sixth, our results regarding transparency relied on self-reported understanding. Future research could complement transparency measurements with testing effective understanding, for example through adjusted tasks. Overall, we hope our suggestions help to pursue research into providing adolescents with control over recommendations in e-learning.

7.6 Conclusion

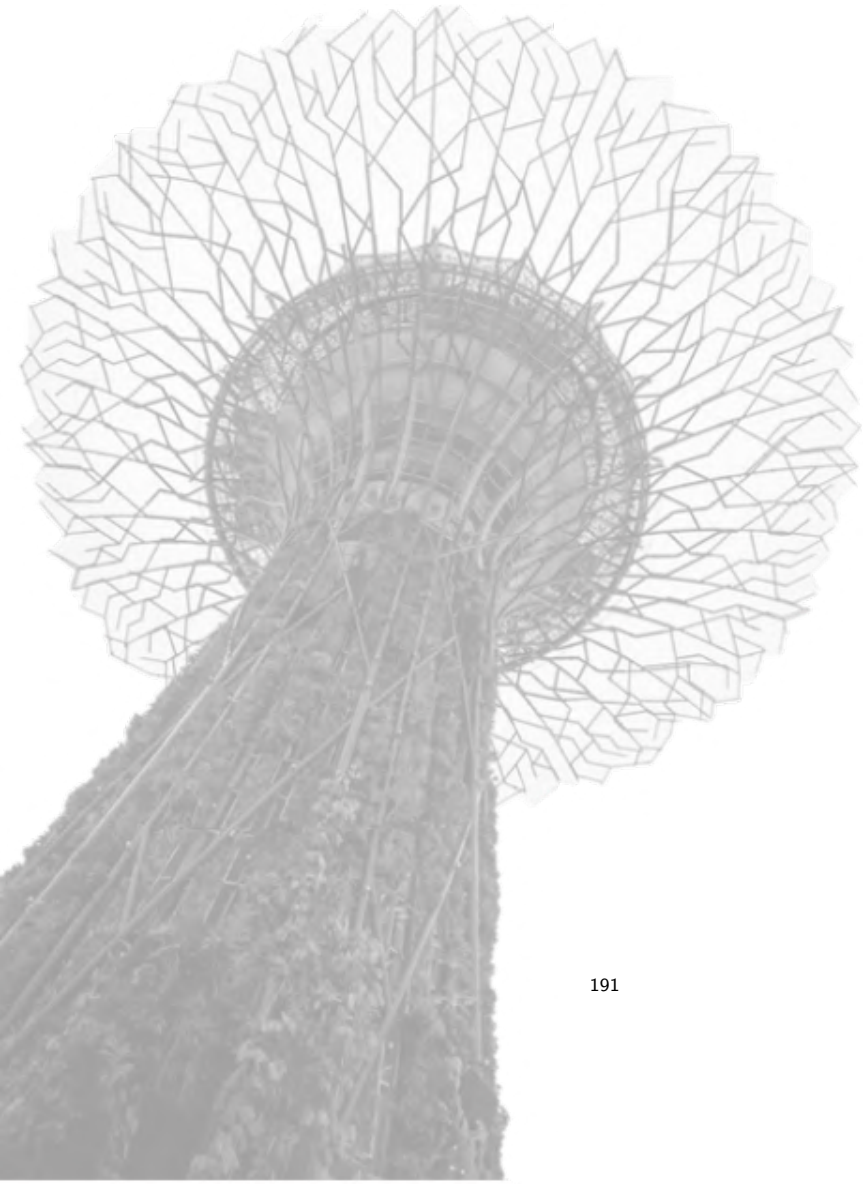
Our research explored how a control mechanism for steering recommended exercises and a visualisation of the control's impact influence adolescents' trust in an e-learning platform. We measured trust both with a single Likert-type question and as a multidimensional construct of trusting beliefs, intention to return, and perceived transparency. In addition, we collected qualitative feedback to further contextualise students' responses. Our randomised controlled experiment with 76 middle and high school students showed that our control mechanism did not significantly change any trusting perception. However, adolescents appreciated the feature and seemed to reflect more upon their mastery level and the recommendation system, which is highly favourable in the context of self-regulated learning. Furthermore, visualising the control's impact did increase trust and perceived understanding, which suggests several implications for the broader field of explainable AI. In sum, even though our study had limitations, we hope our methods, designs, and findings inspire other researchers to further explore the link between control mechanisms, explainable AI, and motivational techniques, especially in e-learning and targeting adolescents.

Acknowledgements

We thank all students who participated in our study, their parents for consenting, and their teachers for inviting us into their classrooms. This work was supported by Research Foundation–Flanders (FWO, grant G0A3319N), Flanders Innovation & Entrepreneurship (imec.icon grant HB.2020.2373), and KU Leuven (grant C14/21/072).

The Human Side of Chapter 7

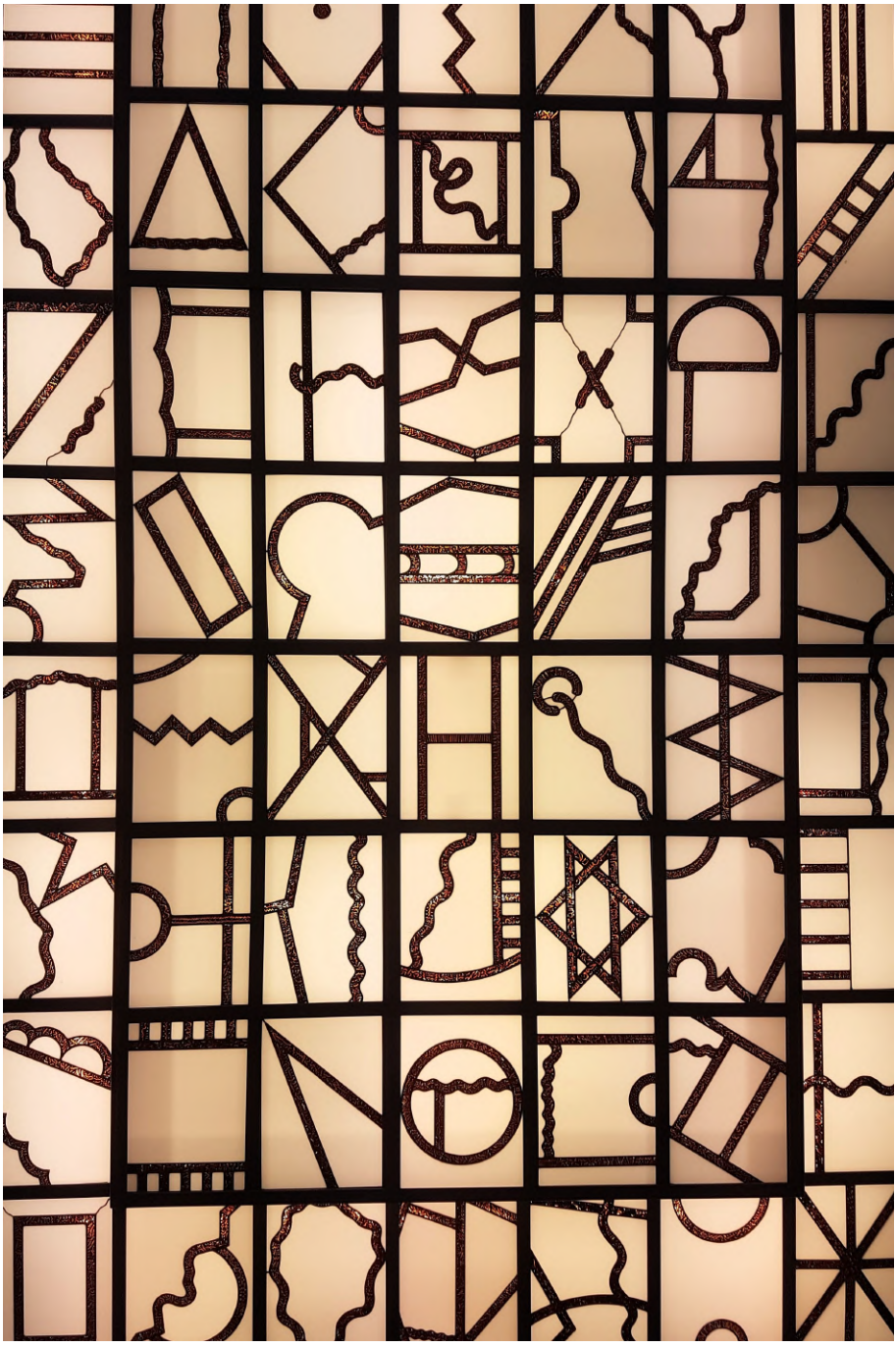
Consistent Work Ultimately Pays Off



Leen was one of the students whose master's thesis I guided in the academic year 2021–2022. We started collaborating right after my partner and I moved to a new apartment, which made it feel as if the research plan and the apartment's interior took shape simultaneously. Although I'm still terrible at it, Leen taught me the value of consistent and planned work. Each time we met, she had processed all my feedback and made significant progress, which was a pleasure because it allowed us to continuously polish the work. The photograph reminds me of how many small efforts can together form an impressive end result. From August to October 2023, I extended and refined Leen's work into the paper in Chapter 7. It was the first time I finished a paper without stressing too much about it.

Songs on repeat:

- *Spellbound* by Siouxsie and the Banshees
- *Keep Running* by Tei Shi
- *Almost There* by Anika Noni Rose



Ceiling of the Stadsmus museum in Hasselt – July 2022

The paper got accepted with excellent reviews and thanks to the Gary Marsden Travel Award from SIGCHI, I could travel to the IUI 2023 conference in Sydney. Then came another surprise: Li Chen and Yucheng Jin invited me to make a stopover in Hong Kong for a two-week research stay at their lab in Hong Kong Baptist University. Their hospitality and the university campus were amazing, and everyone I met was extremely helpful and kind. Besides collaborating on new research and giving a seminar, I learnt a lot from living in Hong Kong and exploring the area. One weekend, my friend Xianglin Zhao toured me around Hong Kong Island and made this unique picture with his drone.

Central district and Victoria harbour in Hong Kong – March 2023

IUI 2023 in Sydney (Australia): after 3.5 years of PhD, I could finally present a paper at an in-person conference. Such an exhilarating feeling! I remember I couldn't suppress a little cheer when I stepped on the stage to present the work Leen and I did. Similar to CHI 2022 in New Orleans, the in-person experience gave me a huge mental boost, especially when bumping into people with research topics similar to mine or people I'd met before virtually, and going for group dinners. Even though I truly enjoyed the conference and the city, I was also panicking about a postdoc proposal that needed to be submitted. On D-Day, I was moved when my friend Clara Bove especially returned from a trip into the city to support me. After the submission and last conference event, we continued discovering the stunning city and had so much fun that we forgot to take a picture together. Only later did I notice that this photograph of the reflected striking blue sky vaguely shows our silhouettes.

Songs on repeat:

- *Flowers (Demo)* by Miley Cyrus
- *Handstand* by Miley Cyrus
- *Violet Chemistry* and the rest of the *Endless Summer Vacation* album by Miley Cyrus

Darling Harbour in Sydney – March 2023

The three weeks in Hong Kong and Sydney had been wonderful yet busy and draining. On my way back to Belgium, I therefore stayed a couple of days in Singapore to catch my breath a little. Ironically, I was often gasping for breath because the country was so beautiful. I just couldn't get enough of the parks, museums, architecture, historical sites, diverse population, and perfect weather. Especially fascinating was how seamlessly nature and the bustling city fused, as captured by the iconic "super trees" in Gardens by the Bay, which were spectacularly lit at night. Reflecting upon my latest travels, I realised how important I find international collaboration and learning from other cultures, and to this day, I am incredibly grateful for my experiences in Australasia.

Gardens by the Bay in Singapore – March 2023

Chapter 8

Steer, See Impact, Solve: How Learner Control and Visual Explanations Impact Learning, Motivation, and Trust

Technology-enhanced learning and AI are on the rise, leading to questions about how learners can better understand the AI models that guide their learning process on e-learning platforms and how they can steer those models. Such functionalities are not only important for calibrating trust, but also for learning goals such as metacognition, motivation, and enjoyment. Fortunately, the learning research community has a long tradition in making automated learning systems more transparent and controllable with open learner models and learner control mechanisms, aligning with the modern stream of explainable AI for learning. Yet, published research in this area typically lacks needs studies, does not focus on controlling the selection of learning materials, and does not include interactive visualisations that show how the AI models underlying learning systems behave. Our work addresses these underexplored topics. Through an elaborate human-centred design process and a randomised controlled experiment with 170 adolescents in school, we designed and evaluated a control mechanism that enables learners to steer the difficulty of automatically composed exercise series, complemented with visual *what-if* explanations and motivational feedback. Our design process with pedagogical experts and adolescents uncovered many insights concerning XAI and control for education, for example that *why*

explanations might be more beneficial for teachers than young learners. Next, our experimental results suggest that control, visual *what-if* explanations, and motivational feedback did not have a strong short-term impact on motivation, metacognition, enjoyment, learning performance, or trust. However, they persuaded adolescents to steering towards more difficult exercises. Overall, these findings spark ideas for long-term studies on interactive visual explanations and control for adolescents and learners in general.

8.1 Introduction

In recent years, education is increasingly embracing technology-enhanced learning for personalised learning (Verbert et al., 2012) and learning is shifting away from traditional classrooms to e-learning environments (Salau et al., 2022). These evolutions make large-scale data collection possible, which in turn allows further adoption of artificial intelligence (AI). The histories of AI and education are deeply intertwined (Doroudi, 2022), leading to many examples of how AI can recommend learning materials (Drachsler et al., 2015; Khanal et al., 2020; Salau et al., 2022), automatically assess learners' mastery level (Galici et al., 2023; Klinkenberg et al., 2011; Torkamaan and Ziegler, 2022), and create educational content (Bitew et al., 2022; Khosravi et al., 2023; Kurdi et al., 2020; Ni et al., 2022). Similar to other domains, calls for *explainable AI* (XAI) and control mechanisms are emerging in education now (Khosravi et al., 2022). Interestingly, education has a long tradition in both aspects. First, to provide transparency, education has long been studying open learner models, which show learners what the system knows about them (Bodily et al., 2018b; Bull, 2020; Bull and Kay, 2007; Bull and McKay, 2004; Rahdari et al., 2020). Second, to foster metacognitive skills (Zimmerman, 1990) such as self-knowledge and reflection, learners have been given control over all learning aspects, including their learner model, the way learning materials are being selected and presented, and learning materials' difficulty (Brusilovsky, 2023; Bull and Pain, 1995; Kay, 2001; Mabbott and Bull, 2006; Papoušek and Pelánek, 2017).

However, existing research typically does not include needs studies of end users (Bodily et al., 2018a), potentially because it is challenging to meaningfully involve non-technical stakeholders throughout the design process of educational systems (Holstein et al., 2019). Given the rise of AI-supported educational systems, it is important to map end users' explainability and control needs. Furthermore, there is a lack of research on control mechanisms for selecting learning materials in collaboration with AI models (Brusilovsky, 2023). One possible reason is that learners might not always be ready to exercise control over learning materials to insufficient knowledge, especially

when they are young (Brusilovsky, 2023). Yet, learner control has generally been considered motivating and enjoyable (Clark and Mayer, 2011; Long and Aleven, 2017), so it seems beneficial to further explore mechanisms for shared learner control. Finally, current explanations are typically static, which is why even the wider XAI community requests for studies about whether interactive explanations have different effects (Abdul et al., 2018).

To address these challenges, we conducted an extensive iterative design process with end users in the scope of an e-learning platform, focusing on *why* and *what-if* explanations, and a mechanism for learner control over the difficulty of recommended exercises. Then, we investigated how our designs affected adolescents' learning attitudes and perceptions of an e-learning platform. Our general research questions were as follows:

RQ1. How can visual *why* and *what-if* explanations meet learners' and teachers' explainability needs on e-learning platforms and how should these explanations be designed?

RQ2. How can learners share control over the difficulty of recommended exercises and how should such a control mechanism be designed?

RQ3. How do visual *what-if* explanations and shared control over recommended exercises affect adolescents' motivation, metacognition, enjoyment, and trust in e-learning platforms?

Our research contributes in two ways. First, our extensive design findings reflect broader needs for learner control and explanations for recommendations, both for adolescents and other educational stakeholders such as teachers. In addition, it seems particularly relevant for educational XAI to investigate how it can help motivate learners. Second, our randomised controlled experiment shows that control mechanisms combined with visual explanations and feedback do not necessarily have strong short-term effects regarding motivation, metacognition, enjoyment, and trust. However, visual explanations and feedback do affect how adolescents interact with the control mechanism. We hope our methods and findings inspire longer-term follow-up studies that explore the interplay between learner control and transparency in education, especially when focused on metacognition and motivation.

8.2 Background and Related Work

This section provides some background and previous research findings about visual explanations and control mechanisms for AI models, and human-centred

concepts such as metacognition, motivation, and trust in AI systems.

8.2.1 Visual Explanations for AI Models

Researchers who focus on algorithmic XAI have developed many explanation techniques to provide insights into how black-box AI models get to their outcomes and behave (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020; Du et al., 2019; Guidotti et al., 2019b; Montavon et al., 2018; Stiglic et al., 2020; Vilone and Longo, 2020; Zhang and Chen, 2020). These explanations often capture a lot of information, which is why visualisations are often applied for effective communication. Examples include visualising feature importances (Bertrand et al., 2023; Lundberg and Lee, 2017), interactive sensitivity analysis (Hohman et al., 2019a; Szymanski et al., 2021), *why* explanations about recommendation processes (Bostandjiev et al., 2012), and example-based explanations (Cai et al., 2019). For education in specific, Ooge et al. (2022a) justify recommended exercises by visualising information about the collaborative filtering step, Barria-Pineda (2020); Barria-Pineda and Brusilovsky (2019); Barria-Pineda et al. (2018) justify recommended exercises by showing how likely learners are to solve them correctly, and Abdi et al. (2020) complement recommendations with a visual open learner model.

8.2.2 Control Over AI Models

The rise of AI-supported systems has raised questions about how control over decision-making should be distributed among AI systems and the people using them: should AI systems be given full control, or should there be human-AI collaboration (van Berkel et al., 2021)? This lead to more questions such as when people should be able to exert control and how they can do so. Furthermore, control and transparency are two sides of the same coin (Storms et al., 2022): having control over an AI system can grow better understanding of how it behaves and makes decisions, and, conversely, seeing how an AI model comes to its outcomes might evoke a higher need for correcting or steering it. This shows how user control is tightly linked to *explainable AI*, which is why it has been extensively studied in the general setting of recommender systems (Jannach et al., 2017).

In the context of learning, it is customary to talk about *learner control* (Brusilovsky, 2023; Kay, 2001). Different learner control mechanisms have been proposed, including directly changing the learner model, persuading the system to change it, or negotiating about the model contents (Bull and Pain, 1995; Mabbott and Bull, 2006). These mechanisms yielded mixed results. For example,

on the one hand, Long and Aleven (2016, 2017) found that learner who had access to both an open learner model and control mechanism performed better. Furthermore, Ooge et al. (2023) found that combining learner control with a visualisation of its effect improved adolescents' trust in an e-learning platform. On the other hand, Papoušek and Pelánek (2017) suggested that giving learners direct control over question difficulty is not beneficial overall, even though it can lead to higher engagement for learners who prefer easy questions. In addition, Jansen et al. (2016) found no beneficial effects for maths practice, improvement of maths skills, or self-belief concerning maths when learners could adapting the success rate of exercises.

8.2.3 Metacognition, Motivation, and Trust

Explanations for AI models and control mechanisms can affect many human-centred concepts. In education, researchers have worked on giving control to learners in the light of self-regulated learning, which aims to actively involve learners in their learning process in terms of metacognition and motivation (Zimmerman, 1990). Furthermore, the XAI community has acknowledged the important role of trust in human-AI interaction.

Metacognition refers to the awareness and understanding of one's own cognitive processes, including knowledge of one's strengths and weaknesses, monitoring of learning progress, and the ability to regulate and adapt learning strategies accordingly. By developing metacognitive skills, students can become more effective learners and achieve better learning outcomes, especially in combination with high learning motivation (Bahri and Corebima, 2015). The latter refers to the internal drive and desire to engage in learning activities (Lin et al., 2017).

According to organismic integration theory (Deci and Ryan, 2012a,b), *motivation* is no dichotomy of motivated and amotivated, but rather a continuum spanning intrinsic motivation, integrated regulation, identified regulation, introjected regulation, external regulation, and amotivation (Pelletier et al., 2013). Where intrinsically motivated people wish to complete specific tasks simply for the pleasure of it, amotivated people have no motivation at all. Motivation is crucial in education as it might affect performance (Filgona et al., 2020). To boost academic motivation and competence, social scientists have proposed the concept of *wise feedback* (Cohen et al., 1999; Yeager et al., 2014), which refers to a feedback approach where teachers convey high standards for students' performance together with belief in the students' potential to reach those standards (Yeager et al., 2017). This is related to stimulating self-efficacy, which is the belief learners hold about their capabilities (Bandura, 1995; Margolis and

Mccabe, 2006). Learners with high self-efficacy are willing to put more effort into learning and have higher persistence when facing difficulties (Khine, 2013).

Trust in AI systems is a slippery concept because many definitions have been proposed but none of those is widely accepted (Jacovi et al., 2021; Madsen and Gregor, 2000; Vereschak et al., 2021). Moreover, trust is challenging to measure because it evolves while using a system (Holliday et al., 2016; Nourani et al., 2020; Ooge and Verbert, 2021) and is affected by domain expertise (Nourani et al., 2020; Ooge and Verbert, 2021, 2022), visualised information and uncertainty (Mayr et al., 2019; Sacha et al., 2016), model accuracy (Papenmeier et al., 2022; Yin et al., 2019), level of transparency (Kizilcec, 2016), and many other factors (Hoff and Bashir, 2015). In addition, XAI researchers agree that simply growing trust is not always desirable: more important is *appropriate* trust (Gunning and Aha, 2019) and trust calibration (Han and Schulz, 2020; Solhaug et al., 2007), which implies that people should also distrust ill-performing systems. Some researchers even argue that XAI should abandon studying trust and rather focus on utility (Davis et al., 2020).

8.3 Methods and Materials

This section presents our e-learning platform and design decisions inspired by a pilot study with teachers and an iterative design process with students. Next, it describes our main study design, which was approved by the ethical committee of KU Leuven (reference number: G-2022-5810-R2(MIN)).

8.3.1 E-Learning Platform With Learner Control

We built upon *Wiski*, an existing e-learning platform for middle and high school students to practise maths (Ooge, 2019). *Wiski* contains thousands of multiple-choice questions on maths topics in the Belgian school curriculum and previous work has extended the platform with recommender systems along with visual explanations (Ooge et al., 2022a) and a learner control mechanism (Ooge et al., 2023). In our new extension of the platform, we worked towards combining both. Concretely, we refined one of the earlier exercise recommender systems to better personalise exercise series, implemented a new learner control mechanism, and incorporated *what-if* explanations and wise feedback. This yielded a new overall workflow for learners using the platform, as shown in Figure 8.1. The following paragraphs describe our three main adaptations incorporated in this workflow in detail.

Personalised exercise series Similar to previous work (Ooge et al., 2023, 2022a), our platform automatically generates exercise series tailored to learners' mastery level by following a two-phase process.

In the first phase, an Elo rating system (Elo, 1978; Pelánek, 2016) simultaneously estimates exercises' difficulty and learners' mastery level. In a nutshell, this estimation is based on iteratively updating ratings for both learners and exercises based on how learners solve those exercises. For example, when a learner correctly solves an exercise, their Elo rating increases and the exercise's Elo rating decreases; the change size relies on the difference between the initial ratings. The opposite happens when a learner answers an exercise incorrectly. Over time, these ratings converge to learners' 'true' mastery level and exercises' 'true' difficulty. Meeting the call for fine-grained mastery and difficulty assessment in earlier work (Ooge et al., 2023, 2022a), our platform implemented a multivariate Elo rating system (Abdi et al., 2019), meaning that learners had different ratings for different topics instead of one global rating. This acknowledges, for example, that learners can excel in solving equations yet struggle with geometry. Technical details about our implementation of the Elo ratings can be found in Appendix A.2.

In the second phase, a recommender system uses the Elo ratings to determine which exercises best match a learner's mastery level. Concretely, whenever learner L wants to practise a specific topic, the system lists all exercises on that topic apart from the last 20 that L solved. Then, for each exercise in the list, the system computes the probability that L would solve it correctly according to formula (A.1) in Appendix A.2. Finally, the three exercises with probabilities closest to a difficulty hyperparameter D are recommended. By default, our recommender sets $D = 0.5$, aiming for exercises that are neither too easy nor too hard and thus keep the learner in the so-called zone of proximal development (Murray and Arroyo, 2002).

Learner control Previous work (Ooge et al., 2023) has shown that high school students appreciate learner control. At the start of each exercise series, our platform therefore allows learners to steer the difficulty of exercise series with a slider, as depicted in Figure 8.1 (2). In the background, this slider changes the difficulty hyperparameter D in our recommender system from 0 (very easy) to 1 (very hard) with steps of 0.1. Note that this mechanism gives learners control over how difficult the exercise series will be, but not over its exact composition. Moreover, contrary to the control mechanism in (Ooge et al., 2023), our approach does not allow learners to change their mastery level directly. Instead, it allows learners to give a kind of signal to the platform that it over- or underestimates their mastery level.

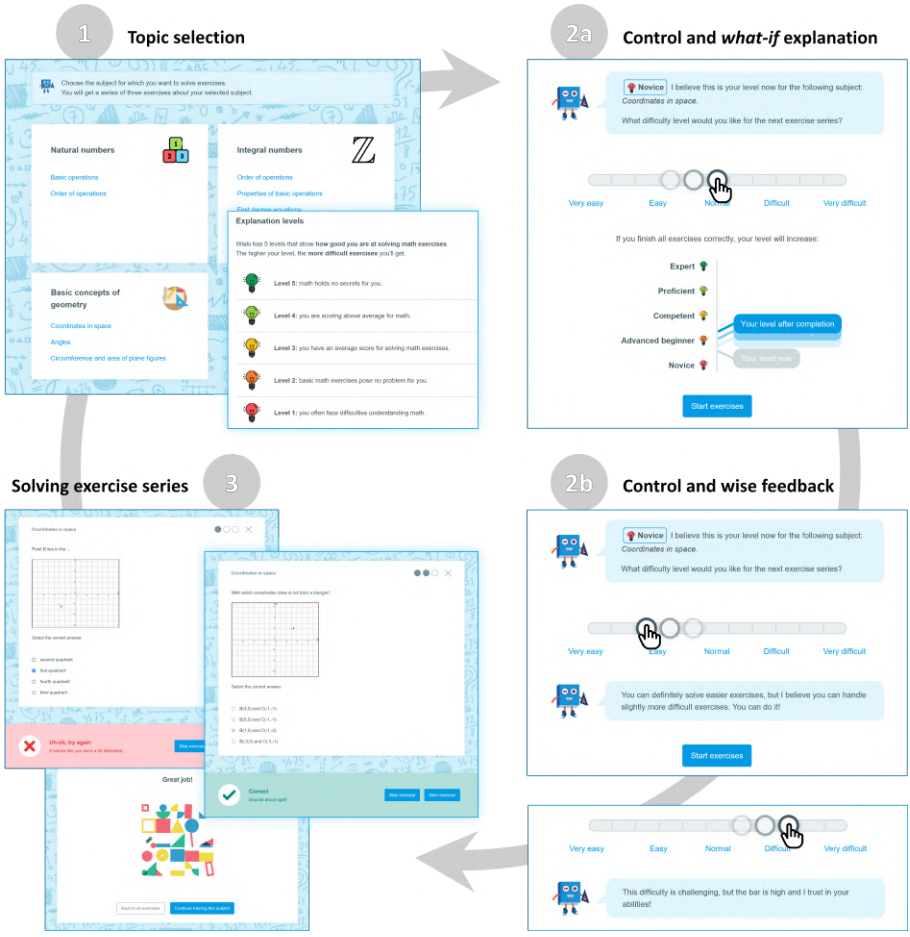


Figure 8.1: The overall workflow on our e-learning platform contains 3 steps: 1 learners choose a maths topic they want to practice and initially inform themselves of the platform's mastery level system; the platform automatically composes an exercise series of 3 exercises that fits the learners' mastery level, but learners can choose to steer the series' difficulty with a slider, additionally seeing 2a a *what-if* explanation, or alternatively 2b wise feedback; 3 learners complete the resulting exercise series and get immediate feedback on their answers' correctness.

***What-if* explanation and wise feedback** Communicating the impact of learner control is essential to foster trust in e-learning platforms that recommend exercises (Ooge et al., 2023). We therefore accompany our control mechanism with the visual *what-if* explanation shown in Figure 8.1 2a. This visualisation explains the potential impact of a recommended exercise series: it shows learners' current mastery level (grey label) together with the level they would obtain if they correctly solve all three exercises in the recommendation (blue label). The *what-if* explanation is fully linked to the control mechanism: when learners change the slider, the blue label changes position accordingly; higher chosen difficulties lead to a bigger increase and vice versa.

To motivate learners and stimulate them to reasonably challenge themselves, we also implemented a form of wise feedback linked to the control mechanism. Figure 8.1 2b shows how the platform's mascotte provides this wise feedback whenever learners choose a specific difficulty level with the slider. Specifically, splitting the $[0, 1]$ interval of difficulty levels into five equidistant subintervals, there are three feedback variants for each subinterval and one random variant is shown at a time. Table A.4 lists all variants.

8.3.2 User Studies to Inform Design

Considering the sensitive nature of our adolescent target group and educational setting, we meticulously designed the learner control mechanism, *what-if* explanation, and wise feedback. Adopting an iterative design approach, we engaged involved adolescents, teachers, and pedagogical experts. Concretely, we conducted two think-aloud studies with adolescents and two focus groups with teachers and pedagogical experts. Section 8.4 provides full details about the procedure and takeaways.

During our think-aloud studies, participants were asked to complete several predetermined tasks in a prototype and articulate their actions (Abras et al., 2004). The studies lasted 15–30 minutes depending on how much feedback participants gave, and participants were rewarded with a €15 voucher. We recorded the conversation for later analysis after written informed consent.

In our focus groups (Hennink, 2014), we presented our prototypical explanation interfaces to participants and asked them to discuss how they could be used on an e-learning platform, whether they met students' and teachers' needs, and how they may affect motivation, trust, and metacognition. The focus groups lasted 2 hours and were recorded after written informed consent.

8.3.3 In-Class Experiment

We conducted a randomised controlled experiment in a real class setting with four groups, which each saw a different version of the control interface we designed. Specifically, in NONE, participants did not have any control over the exercises' difficulty; in CONTROL, participants could steer exercises' difficulty with the slider shown in the upper half of Figure 8.1 2a; in WHAT-IF, participants additionally saw the *what-if* explanation underneath the slider as shown in Figure 8.1 2a; and in FEEDBACK, participants additionally saw the wise feedback underneath the slider as shown in Figure 8.1 2b. At the start of the experiment, participants reported self-estimated overall maths level and motivation for learning (Pelletier et al., 2013) in a pre-study questionnaire. Then, they could practice freely chosen topics for 15-30 minutes, following the flow in Figure 8.1. In the background, we logged all interactions with the slider, changes in Elo ratings, and performance. Finally, participants reported their trust in the platform, metacognition, enjoyment, and motivation in a post-study questionnaire.

Participant Recruitment Collaborating with a school in Belgium (Flanders) in the scope of a larger context on AI for education, we invited adolescents from grades 7 and 8 for participation. All interested students gave informed consent and required parental consent. Eventually, 170 students participated in the study. To ensure participants had sufficient experience with the control screen, we excluded from analysis those who saw the control screen less than 3 times or attempted less than 9 exercises; leaving 163 students. Furthermore, for the pre-study questionnaire, we only analysed the data of 159 participants who completed it attentively, i.e., needing more than 3s per question on average and having at least two different answers. We took similar measures for the post-study questionnaire, ending up with data of 120 students for trust, 107 for metacognition, and 101 for enjoyment and motivation.

Data Analysis To ensure our measurements are valid, we first tested internal validity (ω reliability (Dunn et al., 2014) with bias-corrected and accelerated bootstrap, 1000 replications). Then, we conducted exploratory factor analyses and refactored measured concepts when necessary (Knijnenburg and Willemsen, 2015); checking for multinormality with Mardia's test, factorability with the Kaiser-Meyer-Oblin test and Barlett's test of sphericity, number of factors with scree plots and parallel analysis. This process ultimately allowed us to test our hypotheses regarding the measured constructs, which are summarised in Table 8.1. We used one-sided t-tests; normality of the measurements was reasonable given the sample size and the central limit theorem.

Table 8.1: Overview of our hypotheses H1–H7 with corresponding measurement instruments and findings.

Hypotheses	Measurement Instruments	Result
Hypotheses regarding human perceptions		
H1. *Motivation*: Shared control over exercise selection leads to higher motivation. *What-if* explanations and wise feedback further increase it.	Self-constructed scale.	Not confirmed (all $p > 0.29$).
H2. *Trust*: Shared control over exercise selection alone does not increase trust in the platform, but together with *what-if* explanations and wise feedback it does.	Existing scales (Ooge et al., 2022a; Wang and Benbasat, 2005) measure one-dimensional trust, competence, benevolence, and intention to return.	Not confirmed (all $p > 0.07$).
H3. *Metacognition*: Shared control over exercise selection leads to higher metacognition. *What-if* explanations and wise feedback further increase it.	Self-constructed scale based on (Kay and Kummerfeld, 2019), time spent on control screen.	Not confirmed based on scale (all $p > 0.25$); partly confirmed for *what-if* based on time.
H4. *Enjoyment*: Shared control over exercise selection alone does not increase enjoyment, but together with *what-if* explanations and wise feedback it does.	Existing scale for endurability (O'Brien and Toms, 2010).	Not confirmed (all $p > 0.04$ before correction for multiple testing).
H5. *Learning performance*: Shared control over exercises selection leads to higher learning performance. *What-if* explanations and wise feedback do not further increase it.	Answers to exercises and Elo rankings of exercises. For more accurate estimates, we corrected performance as described in Appendix A.4.	Partly confirmed (all $p > 0.42$).
Hypotheses regarding interactions with the control slider		
H6. *What-if* explanations and wise feedback lead to more exploration, i.e., slider interactions.	Proportion of slider interactions.	Confirmed
H7. *What-if* explanations and wise feedback lead to higher chosen slider values.	Chosen slider values.	Confirmed

8.4 Design Process

This section describes our extensive user-centred design process for the interfaces in Figure 8.1, which contain a slider for learner control, a visual *what-if* explanation, and wise feedback. We present four major iterations together with takeaways relevant for future research on learner control mechanisms for e-learning platforms.

8.4.1 Prototype 1 and Informal Feedback

Based on our experience with recommender systems in education (Ooge et al., 2023, 2022a) and the general call for XAI in education (Conati et al., 2018; Khosravi et al., 2022), we set out to study which visual explanations can help justify recommended series of exercises for adolescents. Figure 8.2 shows our initial Figma prototype. The left side presents a series of recommended exercises together with details about their topic and difficulty, and previous attempts. The right side contains three parts: (1) a split bar chart showing the *importance* of the learner's current mastery level with respect to the recommended exercises; (2) a *global* explanation about the recommended exercises in terms of their similarity to other exercises; and (3) a *what-if* explanation that showed the change in mastery for all topics in case the learner solved the series correctly.

During meetings of a larger project on AI for education, we collected informal feedback from several high school teachers and edtech industrials on our prototype and elicited needs regarding e-learning platforms. Overall, teachers appreciated the idea of personalising the learning process with an e-learning platform, corresponding to previous research (Ooge et al., 2023). However, they found our three explanation types overwhelming and suggested an emphasis on explanations to boost students' motivation.

8.4.2 Prototype 2 and Think-Aloud Studies

Based on the teachers' feedback, we decided to simplify our design. Figure 8.3 shows how we particularly reorganised the right side. First, we merged our initial first two explanations into one overall *why* explanation for the recommendation: the topics of recommended exercises were depicted as a tree. Each branch showed all exercises of the corresponding topic scattered over the branch in rising difficulty, and also the learner's mastery level for that topic. Highlights showed how the recommended exercises were those laying closest to the learner's mastery, and hovering also revealed learners' previous attempts. Furthermore,

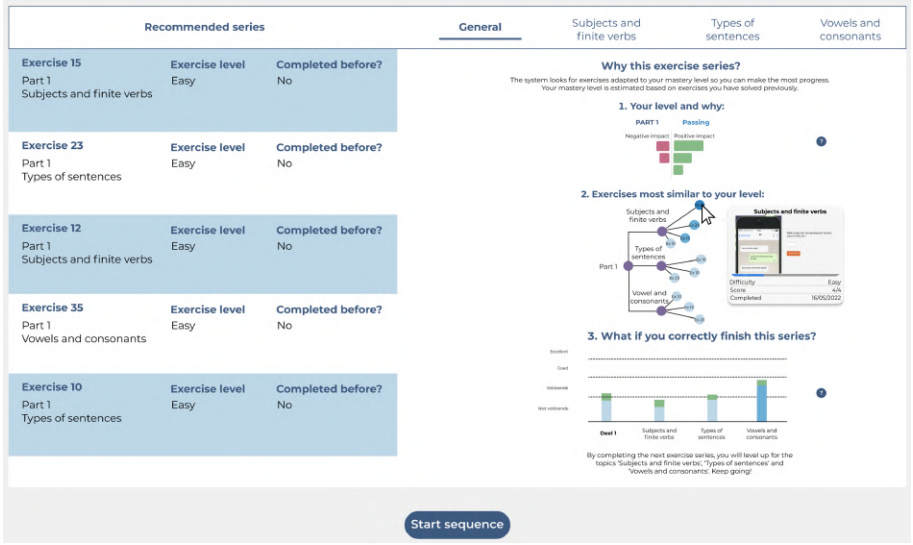


Figure 8.2: Screenshots of our first prototype. Left: a series of 5 recommended exercises. Right: justification for the recommendations with feature importance, a global explanation, and a *what-if* explanation.

we made the *what-if* explanation more prominent and visualised the change in mastery level on an axis identical to the tree branches.

We then conducted think-aloud studies with 6 students (**P1** and **P2** in 8th grade, **P3** in 9th grade, **P4** and **P5** in 10th grade, and **P6** in 11th grade) to test our prototype's overall usability, verify whether teenagers could understand our explanations, and elicit needs regarding personalised learning through an e-learning platform. Table 8.2 summarises our findings. Overall, participants did not experience major usability issues apart from the *why* explanation. The many dots in that visual explanation were rather intimidating and most participants needed some time or oral clarification to grasp their meaning. This taught us three important lessons.

Takeaways about design On a design level, we learnt that we could further clarify the descriptions above the visualisations by merging the legend in, similar to how we styled the mastery label consistently with the visualisation. Still, many participants simply did not read the description before we asked them to. As suggested by **P6**, a brief tutorial or animation explaining the visualisation could be a better alternative or useful addition. More importantly, even though only **P1** mentioned it explicitly, we found that participants did not really link

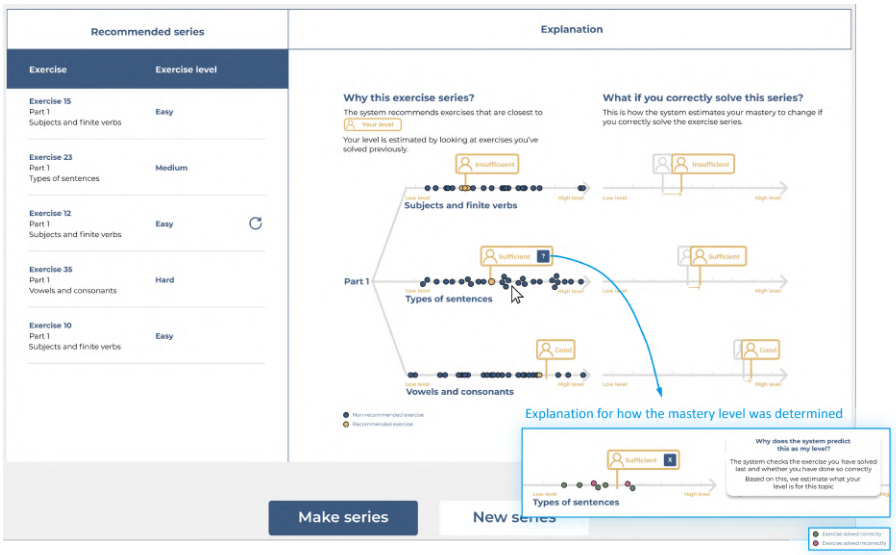


Figure 8.3: Screenshots of our second prototype with redesigned *why* and *what-if* explanations. Hovering scatter plots adds extra vertical jitter, and clicking an info button in the mastery level labels reveals previous performance.

the explanations to the recommendations. This shows that explanations should be well-integrated with whatever they are explaining, and that simply presenting them together is no guarantee for success.

Takeaways about XAI On an XAI level, we learnt that our explanations could only slightly increase participants' understanding of the recommendation algorithm. Moreover, participants typically created a particular mental model by combining elements in the interface and prior expectations. For example, **P1** wrongfully believed the dots in the *why* explanation represented other learners as they were shown together with their own mastery level, and therefore assumed the recommender applied collaborative filtering. This illustrates that cognitively overdemanding explanations might lead to mental models that seem sensible but are wrong. Interestingly, **P4** and **P6** mentioned they found the *what-if* explanation motivating as it gave them a kind of goal to work towards.

Takeaways about control Finally, even though participants saw many advantages in personalised recommendations, they also indicated that they wanted to keep some degree of control over which exercises to solve. For

example, P1 and P5 proposed hand-picking exercises from a list or table which indicates each exercise's difficulty level, and P5 even suggested to re-purpose the scatter plot in the *why* explanation as a way to select exercises.

8.4.3 Prototype 2 and Focus Groups

We also wanted to capture how educational stakeholders other than students received our prototype in Figure 8.3. Therefore, we conducted two focus groups G1 and G2 with respectively 4 (P1–P4) and 3 (P5–P7) pedagogical experts whose details are listed in Table 8.3. Table 8.4 summarises the themes that arose during the discussions. Overall, participants raised comments about the explanations which largely aligned with those of P1–P6 in the think-alouds, but also gave interesting insights into how teachers could benefit from explanations, how they could be improved according to pedagogical practices, and how our recommendation-driven e-learning platform could further meet needs in education. The following paragraphs sum up the main lessons for students and teachers, respectively.

Table 8.4: Findings of our two focus groups with pedagogical experts, discussing the prototype in Figure 8.3. The second column indicates in which focus groups the themes arose. Ticked comments have been addressed or supported in the next iteration, described in Section 8.4.4

Comment	Group	
***Why* explanation for students**		
Could be superfluous if students need to solve the exercises anyway	G1, G2	✓
Visually clean, but too complex and too detailed for students	G1, G2	✓
Showing all exercises could give students the impression they need to solve them all	G1	✓
Only show exercises that fit students' mastery level or are useful to reach a goal (e.g., unlock levels with exercises)	G1	✓
Use different colours for red and green as students tend to interpret them as a personal judgement	G1	
Could be useful for students to choose exercises themselves	G2	
***Why* explanation for teachers**		
Useful for teachers to monitor students	G1, G2	
Suitable for teachers to understand how and why exercises are being recommended	G1	
Useful for teachers to see distribution of exercises' difficulties and potentially identify problematic exercises	G1	
Could support dialogue between teachers and parents (e.g., explain how platform diversifies, show students' track record)	G1	
Construct learning paths by indicating at which mastery level students can switch to another topic	G2	

Continued on next page

Table 8.4 – Continued from previous page

Comment	Group	
***What-if* explanation**		
Highlight progress even when level stays identical	G1, G2	✓
To increase motivation, visualise goals and expectations (e.g., thresholds between mastery levels, path with steps to a goal)	G1, G2	✓
Most interesting part because it shows potential progress	G1	✓
General comments about explanations		
Use more supportive words for "(in)sufficient"	G1	✓
Relation between recommendations and explanations is unclear	G2	✓
Orient axes vertically to better represent the idea of "climbing up"	G2	✓
Both explanations are linked: harder exercises correspond to a higher mastery level	G2	✓
Good to avoid that students are brainlessly solving exercises; explanations can persuade them to practice attentively	G2	✓
Unclear whether students will really look at the explanations	G2	
Potentially only show explanations on demand	G2	
Recommender system and control		
Allow students to choose non-recommended exercises if they assume their mastery level differs from the system's estimate	G1	✓
Students might lack sufficient self-direction to choose the right topics and exercises	G2	
Restrict to practising one topic at a time	G2	✓
Recommend rehearsing theory when students' mastery level is too low or when they make X similar mistakes	G2	
Teachers could assign initial difficulties to exercises and mastery levels to students. Alternatively, let students do it themselves (i.e., foster meta-cognition) or automatically (e.g., based on previous learning data or a pre-test)	G2	
General comments about the platform		
Beware to not only make strong students stronger; also support less motivated, less literate, and cognitively weaker students	G1	
Let students contextualise their performance with analytics (e.g., performance, number of solved exercises, time spent, etc.)	G1	
Analytics can foster dialogue with parents (e.g., discuss when students perform best), even though this can evoke confrontations	G1	
Potentially let students compare themselves with others	G1	
Context can strongly influence students' performance (e.g., home vs classroom)	G1	
Tailor visual design to a younger audience without making it childish (e.g., use an avatar and more graphics)	G2	✓
The platform should inform students when they achieved a goal and not force them to continue practicing mastered topics	G2	
The platform should update exercises' difficulty in a data-driven way	G2	✓
Spaced practice based on mastery level or student motivation; students could themselves decide when to switch topics	G2	

Takeaways for students Participants in G2 appreciated that the explanations could stimulate metacognition, but at the same time questioned whether students would actually pay attention to them. In line with what P1–P6 reported in the think-alouds, participants deemed the *why* explanation too complex for

Table 8.2: Main findings of our think-aloud studies with the prototype in Figure 8.3, ordered by overall theme and frequency. The second column shows which participants raised the comment. Ticked comments have been addressed or supported in the next iteration, described in Section 8.4.4.

Comment	Participants	
Recommended series		
Understands which exercises are being recommended	P1, P2, P3, P4, P5, P6	✓
Expects to see own mastery level or confuses it with exercise difficulty	P1, P2	✓
***Why* explanation**		
Does not (immediately) understand the scatter plot (e.g., position and colouring of the dots are unclear)	P1, P2, P3, P4, P5, P6	✓
Own mastery level for all topics is clear (e.g., because of the consistent lay-out in the text above the visualisation)	P1, P2, P3, P4, P5, P6	✓
Does not find (or would not have found) extra clarification on how mastery level was determined	P1, P2, P3, P5, P6	
Legend clarifies that dots represent exercises	P2, P3, P4, P5	
Did not read text above the visualisation (e.g., because it was too small or too long)	P1, P4	✓
Understanding why an exercise is recommended increases eager to solve it	P1	
***What-if* explanation**		
Arrow between the labels clarifies the progress	P2, P3, P5, P6	
Seeing progress is motivating	P4, P6	✓
Translation of the label clarifies the progress	P1	✓
Recommender system		
General understanding that recommendations are based on previous performance	P2, P3, P5, P6	
General understanding that recommendations are based on mastery level	P3, P4, P6	✓
Expects test to calibrate initial mastery	P1, P5	
Still wants to choose exercises (of a specific difficulty level) themselves	P1, P5	✓
Believes recommendations are influenced by other learners' answers	P1	
Expects mainly recommendations for topics with low mastery level	P3	
General comments about the platform		
Personalisation is useful (e.g., to work more independently, to know own strengths and weaknesses)	P2, P3, P5	✓
Wants to re-make exercises (e.g., by clicking dots in the *why* explanation)	P2, P3, P6	✓
Would not always look at explanations, but wants them on-demand	P4	
Would especially look at explanations when making many mistakes	P6	
Relation between recommendations and explanations is unclear	P1	✓

Table 8.3: Details of the educational experts who participated in our focus groups.

ID	Gender and Background
P1	F – deputy director at high school, educational author, implemented ICT tools in college education
P2	F – college researcher, didactic support person for digital learning, former elementary school teacher
P3	F – scientific employee on educational technologies and partnering with industry, didactician, former educational author
P4	M – educational support person and ICT coordinator at a college, former secondary school teacher
P5	M – high school teacher of languages using ICT in class, pedagogical worker for special care education, educational author
P6	M – manager at different educational publishers, former high school teacher of classical languages and history, pedagogue
P7	M – product owner at educational publisher, former educational author, former primary school teacher

students and were afraid the many dots might lead students to the false belief that they would have to solve all exercises instead of just a subset. Interestingly, participants in both G1 and G2 hesitated whether students need to understand the rationale behind recommendations (*why*) as they need to practice anyway. P6, for example, said: *"Students just want to know which exercises they should solve. Then solve them, period."* Instead, all participants agreed that motivating students should be the main objective in the context of practicing and therefore found it more important to visualise the impact of correctly solving exercises (*what-if*). For example, P4 stated: *"I don't know what the added value is of representing all exercises. I think it's more interesting for [students] to see: if I do this, than that will be the effect on my level."* In addition, P5 put it as follows: *"We must avoid students just brainlessly completing exercises without understanding why they are making them. Not why the system proposed them, but 'what's in it for me', why should I make those exercises."* This relates to another prominent theme: students should mainly see their progress and how they can achieve specific goals. P7, for example, phrased it as follows: *"Students should actually only know why they have a certain mastery level, what steps they should take to get to another level, and which options they have [to get there]."* In sum, participants found the *what-if* explanation more relevant than the *why* explanation for students.

P4 furthermore alluded to giving students more control over exercises' difficulty instead of fully depending on recommendations: *"Can students click on a non-recommended exercise [in the why explanation] to try it anyway? Maybe they just guessed five times and therefore have an insufficient level, while they*

actually do master [the content]. And maybe then they'll say like: 'okay, I'll do my best now'." Interestingly, **P5** also suggested this interaction in the think-alouds. Participants in **G2**, however, cautioned that generalising such control to freely choosing topics and individual exercises might be counterproductive: they questioned whether young students have sufficient self-direction to make such pedagogically important decisions without guidance.

Takeaways for teachers Whereas participants did not deem the *why* explanations suitable for students, they saw lots of potential in them for teachers for at least four reasons. First, it could educate teachers about how exercises are being recommended: *"AI-driven decision-making is still very unfamiliar to many teachers, and this really makes it visual"* (**P3**). Second, teachers could use visualisations similar to our *why* explanation to monitor both students' progress and exercises' difficulties. The former allows teachers to decide which students need personal guidance; the latter supports identifying problematic exercises. Third, the visual *why* explanation could support teachers when talking to parents: *"[to tackle questions such as] 'Why does my child need to solve these exercises?', this allows to perfectly explain that [the exercise series] are put together à la tête du client. [...] It brings nuance to the dialogue, which can sometimes be hard"* (**P3**). Fourth, to stimulate spaced practice, **P5** proposed an original idea: teachers could manually draw learning paths between different axes in the visualisation, indicating at which mastery level students can switch to another topic. Alternatively, the visualisation could depict automatically generated learning paths, which teachers could overrule if necessary.

Discussing the broader scope of educational recommender systems, participants in **G2** envisioned that teachers are ideally positioned to determine students' initial mastery level. The same holds for exercises' initial difficulty levels when the system iteratively estimates difficulties. Moreover, participants in **G1** noted that teachers can use learning analytics collected on e-learning platforms to foster dialogue with parents; provided that they are being trained to interpret such analytics. Together with participants' remarks on how teachers could use our visual *why* explanation, these comments underline that teachers remain important in the context of recommendation-driven e-learning.

8.4.4 Prototype 3 and Think-Aloud Studies

Based on the rich feedback from students and pedagogical experts, we iterated a third time over our prototype, addressing or supporting the ticked comments in Tables 8.2 and 8.4. The resulting interfaces in Figure 8.4 show several drastic changes regarding the graphical design and functionalities. Most noticeably, we

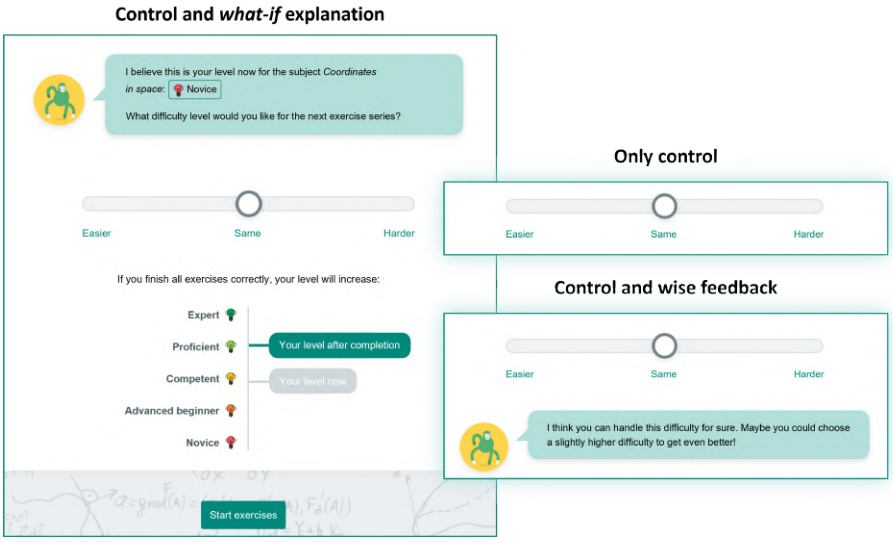


Figure 8.4: Screenshots of our third prototype. Based on the feedback in our think-aloud studies and focus groups, we dropped the *why* explanation, added a slider to control the difficulty level of recommended exercise series, redesigned the *what-if* explanation, and added wise feedback. Both the *what-if* explanation and the wise feedback interactively updated based on the chosen slider value.

decided to drop the *why* explanation: while we could have further improved its visual design to resolve the issues adolescents experienced, our studies showed that the explanation simply did not fulfill their primary need of motivation. Thus, we focused on refining the *what-if* explanation instead. Inspired by the alludations to a control mechanism, we also added a slider with which students can steer the difficulty level of recommended exercise series. In addition, given the educational experts' stress on motivation, we added motivational sentences inspired by wise feedback.

To assess the usability of our new prototype, we conducted another round of think-aloud studies with 6 middle school students (P1–P6, all 7th grade). During the studies, we noticed that the participants were not very fluent in Dutch and read slowly. Our findings summarised in Table 8.5 should be interpreted accordingly. Overall, participants did not face major usability issues: the functioning of the control mechanism and the overall flow of selecting and solving exercises were clear. Yet, observing participants' interactions and registering their remarks gave us some interesting insights.

Takeaways about the control mechanism All participants found it very intuitive to control exercises' difficulty with a slider. **P4**, for example, put this explicitly into words: *"[I] choose harder [...] because the exercises I had now were a little easy."* Furthermore, **P2** explained how the slider made them think more about their level: *"[on the interface without slider] it seems you just have to read the text and start right away and on [the interface with slider] you get to choose hard and so on and then you can start."* This demonstrates how the control mechanism encouraged participants to reflect upon which difficulty levels they could handle. For **P6**, the control mechanism was even the most important aspect in the interfaces, underlining what others seemed to concur with: adolescents highly appreciate learner control.

Takeaways about the *what-if* explanation Five participants identified the interface with the *what-if* explanation as their favourite, mainly because it showed them their progress, was not fully textual, and was colourful. For example, **P4** said: *"I can see where I am with my level, and if I solve another exercise, I can also see if I get to another level, or stay at the same. [...] If I'm too low, this can show I need to work my way up. [...] I like that."* Related to this is that not everyone understood the levels' names, but the coloured icons accompanying them clarified their meaning. This shows the potential advantage of visualisations and supportive visual elements over mere text for learners who are linguistically not so advanced.

Takeaways about the wise feedback Four participants confirmed they found the wise feedback motivating. **P6** stated that the wise feedback stimulated them to indicate a higher difficulty level: *"I wanted to choose easier, but then I saw [the wise feedback], so I made it a little harder. [I kind of like it] because otherwise you always choose easier."* Yet, participants did not always blindly follow the advice: for example, at some point, **P4** read the wise feedback but deliberately chose a lower difficulty to avoid more wrong answers. In addition, **P2** and **P3** admitted they found the text rather long and **P4** raised the question whether the feedback should promote harder exercises after an incorrect answer. These findings show how wise feedback can indeed persuade adolescents to reasonably push their boundaries if they are up for it. At the same time, however, it is unclear whether adolescents would always read the textual feedback and whether they would prefer feedback that is better tailored to their historical performance.

Table 8.5: Main findings of our think-aloud studies with the prototype in Figure 8.4, ordered by overall theme and frequency. The second column shows which participants raised the comment. Ticked comments have been addressed or supported in the next iteration, described in Section 8.4.5.

Comment	Participants	
Control mechanism		
Meaning and functioning of the slider is immediately clear	P1, P2, P3, P4, P5, P6	✓
Not reading or only glancing at text above the slider	P1, P2, P3, P4, P5, P6	✓
Slider supports reflecting on own mastery level	P2, P3, P4, P5, P6	✓
Slider could be more finegrained	P6	✓
***What-if* explanation**		
Favourite interface (colours, balanced text and visuals, see progress)	P1, P2, P3, P4, P5	✓
The words 'expert', 'competent' and 'proficient' are unclear	P1, P2, P3	✓
Colours clarify the ordering of levels	P2, P5	✓
Precise meaning of levels is (partly) unclear	P2, P5	✓
Wise feedback		
Wise feedback is motivating and supporting	P1, P2, P5, P6	✓
Wise feedback is too long	P2, P3	
Do not promote harder exercises after errors	P4	

8.4.5 Final Prototype

After the second think-aloud studies, we streamlined our interfaces' graphic design to address the collected remarks. Figure 8.1 depicts our final protype. One important change was to alter the overall colour palette from green to blue to avoid confusion with the green colour of mastery levels 4 (Proficient) and 5 (Expert). To reduce preference bias towards the interface with the *what-if* explanation because of its colours, we also turned the platform's background more colourful and visually interesting. Another important change was adding the one-screen tutorial in Figure 8.1 (1), which introduces the platform's mastery levels inspired by the five-stage Dreyfus model (Dreyfus, 2004).

8.5 Results of In-Class Experiment

During the experiment, participants solved 25.2 exercises (SD = 11.9) on average, which corresponds to over 8 series of three exercises. Participants in CONTROL, WHAT-IF, and FEEDBACK changed the default slider value on the control screen 776 out of 1507 times (51%), showing that they were actively using the option to steer exercises' difficulty.

8.5.1 Validation of Measurements

First, we checked the validity of our measurements.

Learning motivation. Since amotivation had bad internal validity ($\omega = 0.54$, CI $= [0.37, 0.66]$) and an exploratory factor analysis with 5 factors yielded much cross-loading and items loading on factors different from the theorised ones, we decided to refactor (KMO $= 0.88$, $\chi^2 = 965$, $p < 0.001$). Scree plots and parallel analysis suggested two factors, and after pruning three questions, we obtained a reasonable factor structure and internal validity for what we refer to as *intrinsic motivation* and *extrinsic motivation*, similar to (Van Houdt et al., 2020) (see Table A.3). Yet, this factoring only explained 44% of the variance. Overall, participants scored higher on extrinsic motivation (mean $= 4.45$, SD $= 1.09$) than intrinsic motivation (mean $= 3.87$, SD $= 1.32$).

Trust. Our data showed good internal validity scores for competence ($\omega = 0.74$, $[0.60, 0.81]$), benevolence ($\omega = 0.82$, $[0.73, 0.88]$), and intention to return ($\omega = 0.84$, $[0.74, 0.90]$). However, to reduce cross-loading and low factor loadings, we refactored until 65% of the variance was explained and the internal validity of competence improved.

Metacognition. We did not have to refactor this construct as internal validity was high and the factor structure seemed robust, explaining 57% of the variance.

Enjoyment. Since internal validity was on the low side ($\omega = 0.70$, CI $= [0.59, 0.95]$) and two items had low loadings, we pruned those items and ended up explaining 56% of the variance with the remaining three items.

Motivation. This construct had a good internal validity ($\omega = 0.80$, CI $= [0.73, 0.85]$), but some factor loadings were low. We pruned items until we explained 61% of the variance.

Performance. To assess performance, we could not solely rely on participants' Elo ratings, because their ratings were not converged yet due to the short duration of our experiment. Therefore, we measured performance in terms of the proportion of correct answers and corrected those values by the estimated difficulty of exercises. Details can be found in Appendix A.4. Figure 8.5 shows that overall performance across the four experimental groups follows a normal distribution, and our correction of the performance scores mainly increased scores in the first two quantiles.

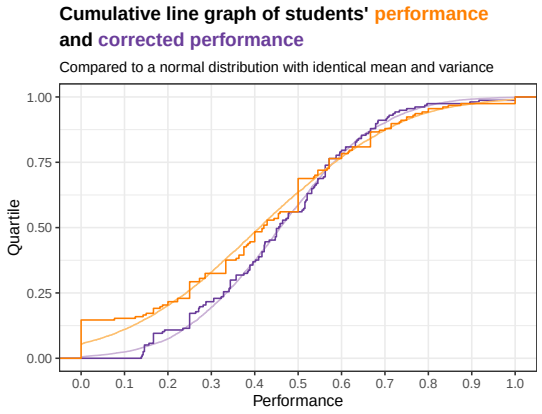


Figure 8.5: Cumulative line graph of the original and corrected performance scores. The correction mainly increased low scores.

8.5.2 Testing Hypotheses About Perceptions

Figure 8.6 shows how participants filled out the post-study questionnaire and how they performed. Visually, average scored did not vary much. Statistical tests supported this: Table 8.1 shows that most of our hypothesised effects could not be confirmed. For one, our subhypothesis in H5 that *what-if* explanations and wise feedback did not increase performance held. Furthermore, Figure 8.7 shows that participants in WHAT-IF spent significantly more time on the slider screen than other groups (for each group, outliers outside the 2σ-interval were removed). This could mean participants in WHAT-IF were more cognitively engaged with the control screen.

8.5.3 Testing Hypotheses About Interactions

After removing outliers outside the 3σ-interval per group, one-sided t-tests showed that participants in WHAT-IF and FEEDBACK interacted with the slider on the control screen significantly more often than participants in CONTROL ($p = 0.02$ and $p = 0.01$, respectively; see Figure 8.7). Thus, the *what-if* explanation and wise feedback stimulated participants to explore different slider values more often. This confirms H6.

The bar charts in Figure 8.8 show that participants especially explored the extreme values (i.e., 'Very easy' and 'Very hard') and values slightly higher than the default value (i.e., between 'Normal' and 'Hard'). Furthermore, the

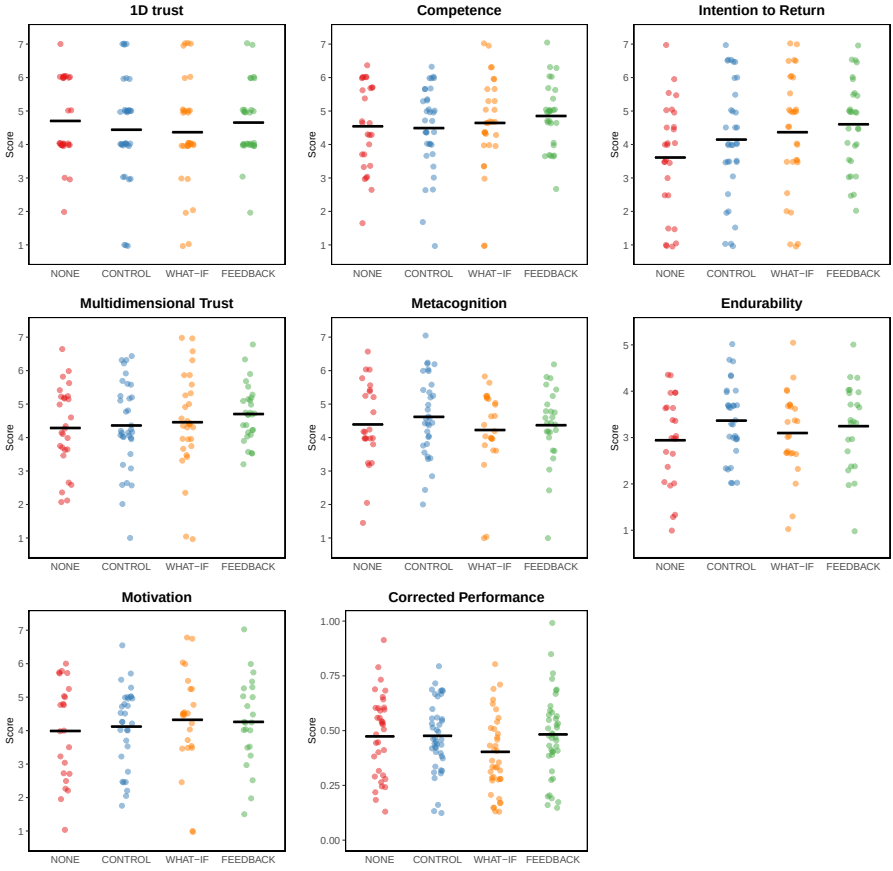


Figure 8.6: Scatter plots of the responses to the questionnaire in Table A.3 for each research group, where the bar indicates the group's average. For visual clarity, dots are slightly jittered horizontally and vertically.

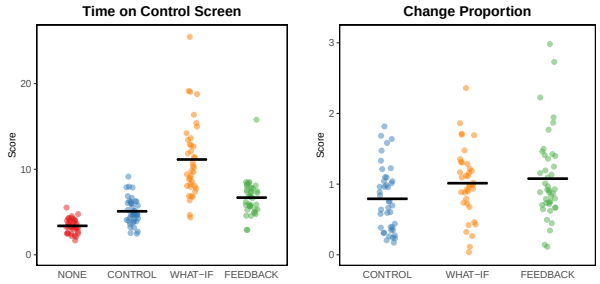


Figure 8.7: Left: time participants spent in the control screen; each dot is the average time of a participant. Right: number of slider changes divided by the number of times participants saw the control screen.

what-if explanation and wise feedback seemed to reinforce this effect compared to participant who only saw the slider. Looking at the eventually chosen slider values, we see that the vast majority stuck to the difficulty proposed by default (i.e., 'Normal'). In WHAT-IF, participants chose less often for the lowest difficulty level, and more often for the middle or highest difficulty, compared to CONTROL. In FEEDBACK there was an even more outspoken shift towards higher difficulty levels, doubling the number of times the highest difficulty level was chosen, compared to CONTROL. Statistical tests showed that these increases were significant in both WHAT-IF ($p = 0.02$) and FEEDBACK ($p = 10^{-6}$), thus confirming H7.

8.5.4 Correlation Analysis

Corrected performance did not correlate with any of the measured constructs, but Figure 8.9 shows several insights about the correlations between all other constructs. First, there was no strong correlation between any of the constructs and participant's intrinsic or extrinsic motivation for learning. Yet, it is interesting that metacognition correlated twice as high with intrinsic motivation compared to extrinsic motivation. This suggests that students who are intrinsically motivated to learn are also more inclined to have a higher metacognition. Furthermore, enjoyment, motivation, and intention to return had the highest inter-correlations, which plausibly suggests that students who found the e-learning platform motivating and engaging are more inclined to return. Finally, competence correlated quite strongly with the other constructs, which for example suggests that participants reported to be more eager to return when they perceived the recommendation system underlying the e-learning platform as more competent.

Figure 8.8: Left: Distribution of the explored slider values, together with explicit differences between WHAT-IF and CONTROL, and WHAT-IF and CONTROL, respectively. Right: similar charts for chosen slider values.

8.6 Discussion

This section discusses the results from our user-centred design process and in-class experiment to answer our research questions. It first proposes some design implications for visual explanations and control mechanisms on e-learning platforms, and then interprets our results concerning the impact of our designs on students' motivation, metacognition, enjoyment, performance, and trust in the platform.

8.6.1 *Why* Explanations not for Adolescents, but for Teachers?

Our user studies suggested that *why* explanations do not fill pressing needs of adolescents in the context of an e-learning platform that recommends exercises. The adolescents we spoke did not seem to strongly require understanding recommendations and only one person mentioned that such understanding might increase eagerness to solve them. As alluded to by teachers and pedagogical experts, this lacking need for explainability might be due to the traditional school system that simply imposes tasks on young students. Furthermore, our proposed *why* explanation sometimes reinforced an inaccurate mental model

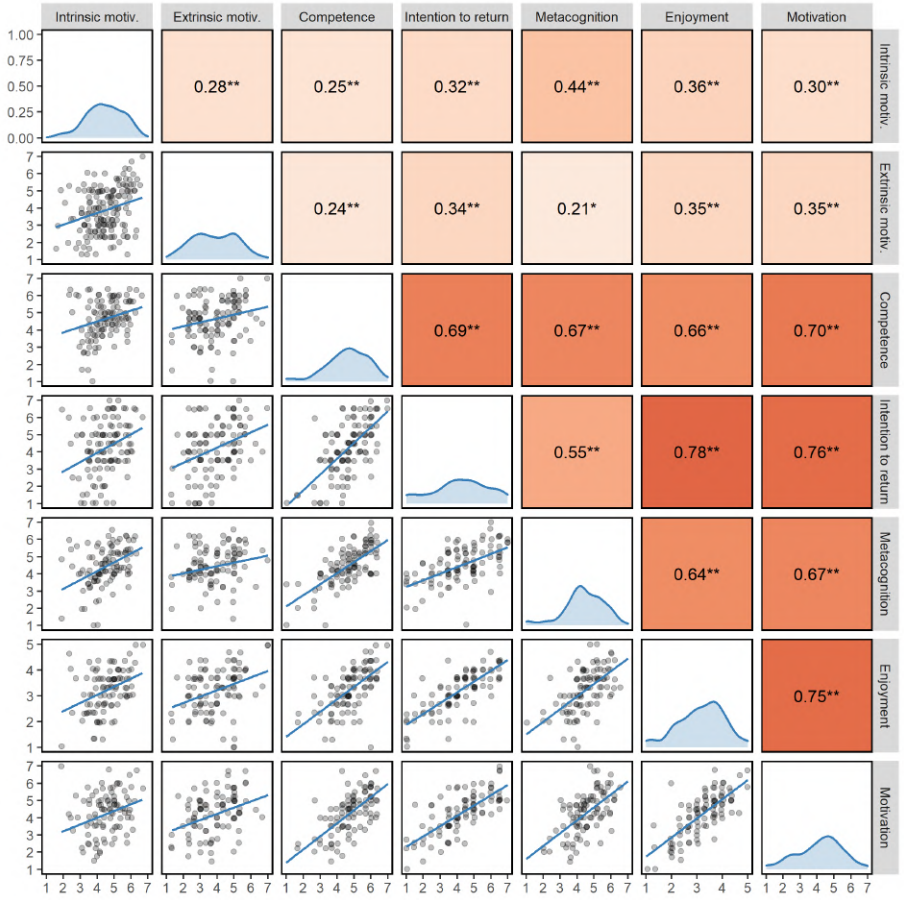


Figure 8.9: Relations between all measured constructs. Lower triangle: scatter plots with regression lines. Diagonal: density plots of constructs. Upper triangle: correlations coloured by value ($^*p < 0.01$, $^{**}p < 0.001$).

because its visualisation was not clear at first glance. This could have been due to low visual literacy or heuristic thinking: to make faster decisions, people resort to intuitive and low-effort thinking, but this can make AI novices more prone to misunderstanding explanations (Wang et al., 2019a). Preventing misunderstandings with textual annotations, as suggested by Szymanski et al. (2021), brought no immediate solace as we found those to be overlooked by adolescents, potentially because of low reading proficiency.

Overall, even though our *why* explanation seemed to become clearer once adolescents paid closer attention and became more familiar with the visualisation, we did not further evaluate it as it did not fulfil learners' needs. For contexts where *why* explanations become more prominent, we recommend to follow a visual approach as the graphics and colours drew adolescents' attentions, and to keep textual annotations concise and linguistically simple. We also highlight the benefit that *why* explanations could have for teachers, both for understanding and steering the recommendations, as for improving communication with parents. This fits well within learning analytics (Bodily et al., 2018b).

8.6.2 Fostering Motivation With *What-If* Explanations and Wise Feedback

Teachers and pedagogical experts were clearly concerned with motivating students, which is in line with broader attempts to foster motivation on e-learning platforms; for example with gamification or support communities (Naidoo, 2020; Ooge, 2019). Our qualitative studies showed that our *what-if* explanation could fulfil this need for motivation: both adolescents and pedagogical experts praised it for showing the potential positive impact on mastery level as it instilled a goal to work towards. This resonates with motivation theory on performance-approach goal orientation, which is learners' goal to demonstrate and prove ability (Leondari and Gialamas, 2002).

However, our randomised controlled experiment did not show an increase in reported motivation for learning for participants who saw *what-if* explanations. Surprisingly, the same held for participants who saw wise feedback, which is specifically designed for increasing motivation through self-efficacy. We see several possibilities to contextualise these effects. First, motivation was indeed not affected and our study illustrates how people's expectations do not always match with their perceptions afterwards. Second, our *what-if* explanation actually did motivate participants to solve exercises, but since our measurement instrument focused on motivation for overall learning, we missed that effect. Third, our experiment did not last long enough to instil large motivational differences.

8.6.3 Learner Control Is Not a Panacea

While our designs initially only focused on transparency through visual explanations, it was interesting that both adolescents and teachers raised a need for control over the recommended exercises. While their suggestions to fill this need involved manually selecting exercises, we implemented a slider through which students could manipulate the difficulty of exercise series. This proved to be an adequate level of control for students. Our in-class experiment revealed two interesting insights.

First, our experiment did not align with previous research or hypotheses that adding learner control can lead to more motivation, enjoyment, metacognition (Ooge et al., 2023), or performance (Long and Aleven, 2016, 2017). Yet, our results matched with previous findings that merely adding control does not increase trust in the platform (Ooge et al., 2023).

Second, our logging data of how adolescents interacted with the control slider showed they typically chose extreme values ('Very easy' and 'Very hard') or the central value ('Normal'). This is rather surprising as adolescents explicitly asked for more granular control levels during our design evaluations. While it is interesting to see that both *what-if* explanations and wise feedback realised the upwards shift we intended, it seems less desirable if chosen difficulties therefore 'overshoot' to levels that may be too high for learners. Thus, future designs could consider adaptively restricting the allowed steering, or providing more prominent warnings when chosen difficulty levels do not match with the learner's estimated mastery level or track record.

8.6.4 Limitations and Future Work

Our research has several limitations which restrict how well our findings generalise. First, our study was conducted during a single school period, making it too brief to detect effects that arise on the longer term. Since especially trust and motivation are calibrated in the long term (Holliday et al., 2016), our non-conclusive results might be unsurprising. Second, the 8 classes that participated in our study differed in how well they focused on the study. While some classes worked very concentrated, others were more chaotic with interaction between participants. It is unclear to what degree this may have biased the results, but it could explain the high variance in almost all self-reported measures. Third, wise feedback may be more effective in terms of motivation when not given at the start of every single exercise series and when adapted to the learners' track record. We opted against this to make sure participants saw the feedback sufficiently often during the study, and provided several versions for

each difficulty level instead. In addition, we based the phrasings of the wise feedback on the literature (Yeager et al., 2017), but did not consult teachers, for example. Fourth, we noticed that many participants were not fluent in Dutch, which might have caused inaccurate measurements and ignoring textual information in our interfaces. We tried to accommodate this limitation by rigorously refactoring the measured concepts, but future researchers should be careful with considering the newly composed scales as 'validated'. Fifth, we only included young adolescents in our study, around 11–13 years old. Given the rapid changes in adolescence, future studies should compare our results with those of older adolescents.

8.7 Conclusion

While studies on how to design e-learning platforms for adolescents are abundant, XAI studies on how to design effective explanations and learner control mechanisms for them are not. During our design process with three major iterations, involving both adolescents, teachers, and pedagogical experts, we derived many design lessons regarding these aspects for adolescents. In particular, we found that *why* explanations do not necessarily fulfil explainability needs for young learners, but could be very useful for teachers. Furthermore, *what-if* explanations were received well in view of motivation. Our randomised controlled experiment with 170 adolescents from grades 7 and 8 did not show strong evidence for increased motivation, trust, metacognition, enjoyment, or learning performance under shared learner control, whether or not accompanied by *what-if* explanations or wise feedback. Yet, the latter two did affect how adolescents used the control mechanisms, leading to possible future research paths on how learners best collaborate with AI systems.

Acknowledgements

Thank you to all participants and teachers. Special thanks to Joke Vandepitte for recruiting students and teachers, and Bram Faems for recruiting the pedagogical experts and helping with the conducting the focus groups. Thanks to Peter Brusilovsky for sharing his inspiring paper (Brusilovsky, 2023) before publication. Thanks to Vero Vanden Abeele, Luciana Monteiro Krebs, and Vincent Aleven for discussing our preliminary results with us.

The Human Side of Chapter 8

Everything, Everywhere, All at Once

I started contributing to the project that led to this paper in November 2021 and had a great collaboration with Arno for about a year, in which we conducted the first four user studies. Sadly, Arno interrupted his PhD before we could roll out a final evaluation. Maxwell fortunately jumped in to help me and we decided to go for an ambitious randomised controlled study. This, however, required lots of implementation work, practical arrangements with the participating school, and careful tweaks in the research plan. Doing all that in less than a month time was extremely stressful, especially because the study would be rolled out on a rather large scale and many teachers had freed up a class period for us. The night before the study, while frantically trying to finish the implementation, I had a panic attack. If it weren't for my partner who had been anxiously sleeping with one eye open, I would have collapsed. To make matters worse, I even missed the train to get to the school in the morning. Luckily, Maxwell took care of the practicalities while I took the next train, and in the end the final study was a success. It took a huge amount of stress and adrenaline, but I'm proud we pushed ourselves that month. Afterwards, we joked about our research adventure and asked Midjourney to illustrate it with the prompts underneath. In case you're wondering, especially the image in the upper left seems a pretty accurate depiction of what I must have looked like.

A man driving his bike like crazy in the middle of the night. He looks very tired and his hair is a mess. He is late for the train, which he sees leaving in the distance. The man is panicking.

Two researchers are working together. They are having the best time of their life, they are smiling. There is a rainbow in the sky, the sun is shining. There are sparkles everywhere.

Songs on repeat:

- *Démons (live orchestral)* by Angèle and Damso
- *Water Water* by Empress Of
- *All Night* by Maceo Plex, and Oscar and the Wolf
- *Hold On* by En Vogue

Images generated by Midjourney – December 2022

January 2023 ushered in a month-long non-stop working rush, while the end of my PhD hung above my head like the sword of Damocles. While I was trying to analyse data collected during the final experiment, I travelled to Australasia for a research visit and the IUI 2023 conference (see Page 191), Paris for a seminar in which Katrien and I presented our work, and Hamburg for the CHI 2023 conference. In the meantime, I was mentoring the thesis of master's students, teaching, applying for travel and postdoc funding, arranging another research stay (see later), applying for a postdoc and faculty position, contributing to four full papers which had to be submitted, reviewing a bunch of papers for conferences, contributing to new research with colleagues, starting my PhD text, and much more.

The Panthéon in Paris – April 2023

End of May 2023, I travelled to Pittsburgh (United States) for a three-month research stay at Vincent Aleven's lab in Carnegie Mellon University. It would be the ultimate ending of my PhD. The "small" maze-like university campus was spectacular, I made friends for life, and the city was simply enchanting under the summer heat. Working and living in Pittsburgh was an incredible experience and I would need a whole chapter to tell you everything about the dozens of bridges that misled me all the time, the "stop, squash, scrape" lanternflies and fireflies, the deer and rabbits in the streets, the impressive storms after hot days, and the incredible people I met. Yet, being away from my partner Yens and family was also mentally challenging sometimes, especially because I was under a lot of work pressure. After two months, it was thus a relief to take a week off in New York with Yens. NYC was a thrill so it was hard to capture it in a single photograph. I selected this one because it shows a glimpse of the many different architectural styles, the green natural elements, the typical large windows in skyscrapers, and the ubiquitous art.

MoMA museum in New York – July 2023

In terms of work, my time in Pittsburgh was a continuation of the multi-tasking marathon I had been running for months. Besides conducting new research funded by the Research Foundation Flanders (FWO), I finally finished the analysis for this paper and wrote the whole text together with the rest of the previously unpublished parts in this thesis. Even though it took many nights with little to no sleep, it felt really good to finish the project I started contributing to over 2.5 years ago. The bench in the picture shows the place where the final parts came together. Looking at the Cathedral of Learning in the distance and the deer and squirrels around me from time to time really helped me get over the fear of not finishing; or worse, finishing imperfectly.

Songs on repeat:

- *Jon Batiste Interlude* by Lana Del Rey
- *All Is Full Of Love* by Björk
- *Deep end* by Lykke Li
- *Don't Delete The Kisses* by Wolf Alice

Schenley Park in Pittsburgh – August 2023

Part IV

Conclusions

Chapter 9

Research Contributions and Future Directions

This final chapter sums up all research contributions of this thesis and zooms out to contextualise these contributions in relation to existing research and provide future research directions.

9.1 Research Contributions

Our research contributes to multiple fields, including human-centred explainable AI, interactive information visualisation, and application domains such as healthcare, agrifood, and education. Overall, we showed how AI explainability can be established through visual analytics (Part I), visualisation-supported justification (Part II), and visualisation-supported control (Part III). We did this by reviewing the existing literature, developing new visualisation-supported explanations and control mechanisms in close collaboration with real end-users of AI systems, and conducting user studies to better understand how our new explainability methods affect people's perceptions regarding AI systems.

Visual Explanations Tailored to People and Contexts

RQ1 asked how visual explanations tailored to a target audience and application domain can make AI models more transparent. We tackled this question by studying the existing literature and designing several visual explanations,

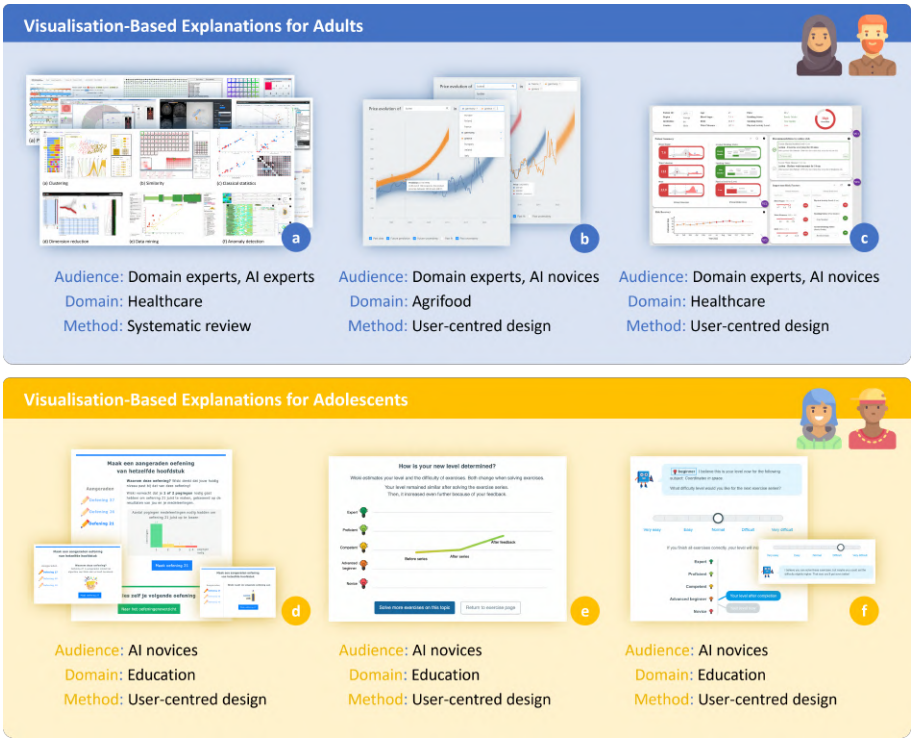


Figure 9.1: Summary of our research contributions concerning visualisation-based explanations tailored to adults or adolescents in healthcare, agrifood, and education. (Credits: people by flaticon.com.)

as shown in Figure 9.1. For the latter, we tailored visualisations towards both domain experts and AI novices in agriculture, healthcare, and education. Moreover, besides targeting adults as most XAI research does, we also worked with adolescents.

In Chapter 4, **we systematically reviewed the existing literature on visual analytics**, restricting ourselves to healthcare (see Figure 9.1a). We found that visual analytics can explain advanced algorithms through visualising their outcomes, interacting with these visualisations, shepherding (i.e., controlling) the algorithmic process to show algorithms' behaviour under different settings, and directly explaining the algorithm with visualisations. These methods are not strictly distinct; for example, interacting with visualised outcomes can be a form of control, and visualisations can be a suitable format to display algorithm-centred explanations such as feature importance. In other words,

there is a fine line between getting insights in the inner logic of algorithms and their outcomes. Regarding the target audience, we found that the vast majority of current visual analytics systems in healthcare target healthcare professionals (i.e., domain experts). Patients are also increasingly involved in health, but since they are typically AI novices and unfamiliar with data analysis, highly exploratory and information-heavy visual analytics interfaces are likely too complex for them. Thus, our review invites researchers to develop explainability solutions for AI novices in a human-centred way, drawing inspiration from visual analytics for domain experts.

Next, **we designed and implemented five visualisation-supported solutions for explainability**. Our first solution in Chapter 5 consisted of a simple visual analytics system for agrifood, which showed product price evolutions and corresponding price predictions (see Figure 9.1b). We operationalised explainability of the prediction model with three functionalities: comparing raw and prediction data for different countries, seeing the model's past performance, and seeing uncertainty in the prediction model. Furthermore, users could enable or disable visual components to focus on the information and insights they needed. Our research underlined the importance of tailoring visual analytics systems towards the application context, users' experience with predictive modelling, and tasks. Our second solution, briefly presented on Page 159, supports healthcare professionals in the context of monitoring patients' risk of diabetes onset (see Figure 9.1c). Risk predictions are visually explained with data-centric, feature-importance, and example-based explanations.

Our three other solutions for explainability were targeting adolescents in an educational context. All these solutions were **iteratively designed in close collaboration with our target group** of adolescents, teachers, and other education stakeholders. In Chapter 6, we justified next recommended exercises with a textual *why*-statement and a bar chart of how many attempts other learners needed to solve the exercise correctly (see Figure 9.1d). As such, we explained that the recommendation algorithm considered learners' mastery level and conducted collaborative filtering. In Chapter 7, a line graph visualised how the system changed learners' estimated mastery after solving series and exerting control over their mastery level (see Figure 9.1e). This allowed for model inspection. In Chapter 8, finally, we combined control over the difficulty level of the next recommended exercise series with a *what-if* explanation that indicated how solving a series of the chosen difficulty would impact the assessed mastery level (see Figure 9.1f). Additionally, textual feedback justified whether chosen difficulty levels were recommended or not. We also designed a *why*-explanation, which turned out to be promising for adults to understand the internal recommendation process. To conclude, it seems appropriate to stress that our visualisation-supported explanations are the outcome of an intensive

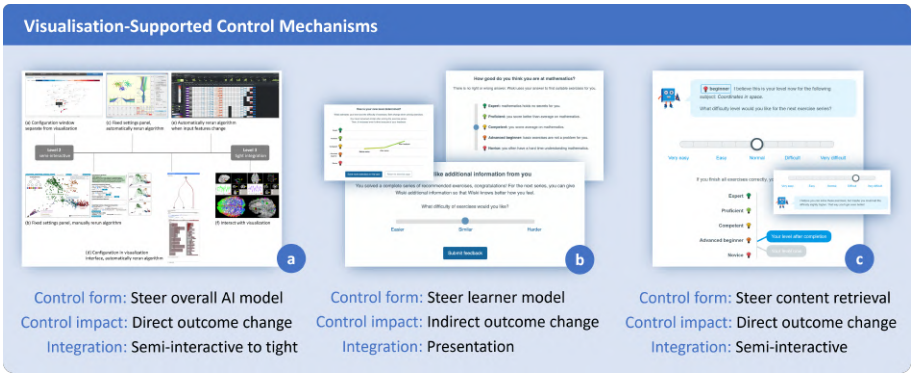


Figure 9.2: Summary of our research contributions concerning visualisation-supported control mechanisms. Each item mentions the form of control, whether outcomes are changed directly or indirectly, and how tightly visualisations are integrated with underlying algorithms.

user-centred design process and are tailored towards adolescents aged 12–18. The latter explains why the visualisations are less advanced than the ones in the first solution, which target adults. Moreover, targeting adolescents seems an important contribution as this age group is often overlooked in XAI research, even though they too are frequently exposed to AI algorithms such as recommendation algorithms on social media.

Combining Control Over AI and Visual Explanations

RQ2 asked how people can control AI models with additional feedback, supported by interactive visual explanations. We first studied existing control methods in healthcare in Chapter 4 and then developed two visualisation-supported control mechanisms for education in Chapters 7 and 8. Figure 9.2 summarises these contributions.

In Chapter 4, **our review showed that less than half of current visual analytics systems in healthcare allows to control underlying algorithms and very few integrate explanations**. On the positive side, visual analytics systems that facilitate algorithmic control span the full spectrum between semi-interactive and tight integration (Turkay et al., 2014) of visualisations and underlying algorithms (see Figure 9.2a). In other words, visual analytics can be used to directly change algorithmic outcomes by modifying parameters and altering the processed data. So far, however, there are few examples of combining control mechanisms with explanations such as feature importance or

sensitivity analysis. This may be related to our finding that most visual analytics systems are either backed by classical statistics or clustering algorithms, which are typically quite interpretable. Thus, our review revealed that the intersection of control-supporting visual analytics and XAI can be further explored and adopted in healthcare.

In Chapter 7 and Chapter 8, **we presented two new ways to control automatically personalised content selection** in a learning context. Both control mechanisms were applied for recommender systems based on Elo ratings, but can be generalised to other AI methods that estimate learners' mastery level and the difficulty of learning content.

Our first control mechanism in Chapter 7 allowed learners to steer their learner model (Brusilovsky, 2023): after finishing an exercise series, learners could lower or raise their system-assessed mastery level with a slider. As such, they could indirectly steer the difficulty of subsequently recommended exercise series. The accompanying visualisation in Figure 9.1e, however, did not support interaction and was thus not integrated with the recommender system.

Our second control mechanism in Chapter 8 allowed learners to steer the content retrieval step (Brusilovsky, 2023). Specifically, before starting an exercise series, a slider gave learners direct control over the difficulty level of the next exercise series. Moreover, the accompanying visual *what-if* explanation in Figure 9.1f was linked to the slider in real time: exploring different slider values immediately updated the explanation. This interactive aspect meets the call for designing and evaluating non-static explanations, which is currently most common in XAI research (Abdul et al., 2018).

Better Understanding Human Perceptions of AI Systems

RQ3 was concerned with how visual explanations and control mechanisms affect people's perceptions of AI systems, specifically in terms of appropriate trust and understanding their outcomes and algorithmic processes. Figure 9.3 summarises the human- and application-grounded experiments (Doshi-Velez and Kim, 2017) we conducted to answer this research question. Overall, our experiments contributed to the XAI state-of-the-art both in terms of new research findings and refined research methods to obtain those findings.

Our human-grounded experiment in Chapter 5 studied the intricate relations between four human-centred metrics: the usefulness of a visual decision support system, needs regarding such a system in the context of price prediction for food products, understanding of the prediction model, and appropriate trust in the prediction model (see Figure 9.3a). Our analysis of both quantitative

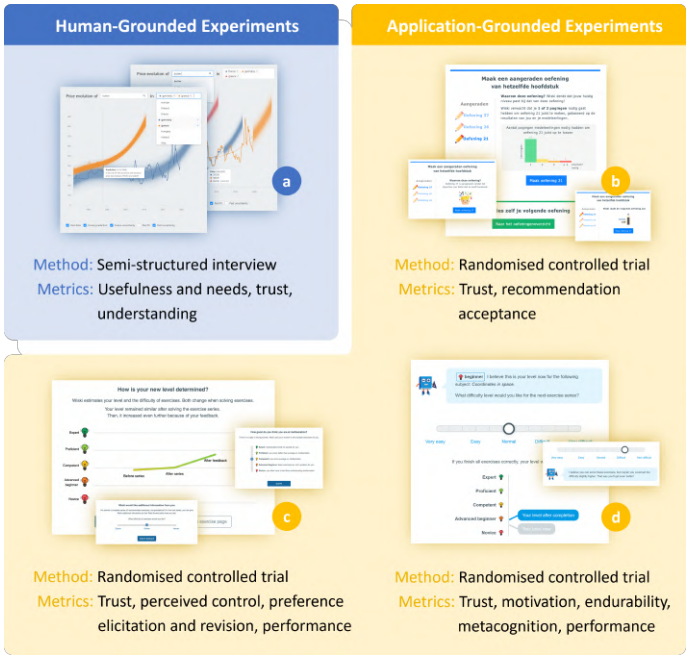


Figure 9.3: Summary of our research contributions concerning human- and application-grounded experiments.

and qualitative data suggested that **usability, usefulness, and model understanding can directly and indirectly affect appropriate trust**. In addition, **perceptions differ between people with diverging experience levels in predictive modelling**. This adds to the rich literature about how people perceive uncertainty visualisation for AI algorithms (Demmans Epp and Bull, 2015; Gutiérrez et al., 2019b; Hullman, 2020; Leffrang and Müller, 2021; Padilla et al., 2021; Sacha et al., 2016; Zhou et al., 2017) and how they interact with visual analytics systems (Cui, 2019; Endert et al., 2017; Saraiya et al., 2006; Savikhin et al., 2011); how these aspects affect their understanding of (Kulesza et al., 2013) and trust in (Han and Schulz, 2020; Hoff and Bashir, 2015; Holliday et al., 2016; Kizilcec, 2016; Schlicker et al., 2022) the AI algorithms; and how this all depends on people's background such as previous experience with AI algorithms (Bayer et al., 2022; Dasgupta et al., 2017; Dikmen and Burns, 2022; Morrison et al., 2023; Nourani et al., 2020; Ooge and Verbert, 2021) and model-centric aspects such as accuracy (Papenmeier et al., 2022; Yin et al., 2019). On a methods level, we also contributed by measuring participants' experience with predictive regression in a way that goes beyond simple self-reporting:

we combined self-reported data with participants' background and jargon use, which we deem useful indirect indicators for experience.

Our three application-grounded experiments in Chapters 6 to 8 were conducted as randomised controlled trials on a fully-operational e-learning platform and in a real class context.

First, our experiment in Chapter 6 (see Figure 9.3b) showed that **visual explanations significantly increased acceptance of recommendations and initial trust** in the e-learning platform. However, initial trust only changed significantly when measured as an average of trusting beliefs, intention to return, and perceived transparency; not when measured with a single question. This suggests that **visual explanations may not be the most important factor for building initial trust**, in contrast to, for example, the platform's appearance and the learning material's quality. Finally, on a methods level, we advanced typical XAI research approaches by using placebo explanations as a baseline and measuring trust fine-grained as a multidimensional construct.

Second, our experiment in Chapter 7 (see Figure 9.3c) showed that **visualising the impact of exercised control significantly increased initial trust** in the e-learning platform. Since there was no such increase for the control mechanism alone, the visualisation caused the trust gain. Arguably, this effect occurred because the visualisation acted as an indirect explanation for the recommendation algorithm: by repeatedly seeing how the e-learning platform estimated and modified their mastery level, adolescents might have better understood why recommendations were suitable for them. Furthermore, **having control over their mastery level seemed to stimulate adolescents' metacognition**: they reflected more upon the underlying recommendation algorithm and whether exercises were tailored to their personal mastery level. In sum, our research suggested potential links between control mechanisms, explanation through visual model inspection, and metacognition.

Third, our experiment in Chapter 8 (see Figure 9.3d) found **no strong evidence that *what-if* explanations lead to higher initial motivation, metacognition, enjoyment, learning performance, or trust** when supporting a learner control mechanism. This did not fully align with previous work or participants' qualitative feedback during interviews, potentially because of the short duration of our experiment. However, interestingly, the ***what-if* explanations stimulated adolescents to choose more difficult exercises**, which shows that they might be a promising technique to encourage learners who underestimate themselves into choosing exercises that better fit their true mastery level. Moreover, we found similar results for motivational feedback inspired by the concept of wise feedback. Studying these aspects paves the way for more research on how XAI can contribute to metacognition or motivation.

9.2 Impact

Beyond research findings and proof-of-concepts for visualisation-based explanations, our research might have impacted the research community in the long term and society in a broader sense as well:

- Our paper (Ooge et al., 2022b) corresponding to Chapter 4 was recognised by the journal publisher as one of the most downloaded papers in the 12 months following online publication. This sparks hope that in the near future visual analytics will be researched more widely as a technique to explain advanced algorithms. In addition, we hope the visual analytics approaches in our review either allow healthcare practitioners to determine whether advanced algorithms can safely be adopted or foster further interdisciplinary dialogue with XAI researchers.
- Our paper (Ooge and Verbert, 2022) corresponding to Chapter 5 was marked by the publisher as a feature paper, meaning it "represents most advanced research with significant potential for high impact in the field." We hope the aspiration encapsulated in this recognition becomes a reality to help spread the adoption of human-centred approaches in agrifood for designing visual decision-support systems. Our research namely suggests that such approaches can lead to systems that better meet people's needs and foster appropriate trust, which ultimately contributes to increased uptake.
- Multiple teachers and schools have expressed interest in further using the e-learning platform we built upon in Chapters 6 to 8. Additionally, we collaborated with industrial partners in educational technology such that our research methods and outcomes can seep through into their current products, which are being used on a large scale. Both events show we helped pave the way towards an exciting new educational approach where learning content is personalised in a controllable and transparent way.

In sum, through our research, we have promoted XAI, visualisation, and human-centred practices in healthcare, agrifood, and education. We feel this will contribute to closer interdisciplinary research efforts and solutions for explainability that align with people's needs.

9.3 Critical Reflections and Future Directions

This section critically discusses several overall challenges for XAI, links them to some of the limitations in our research as indicated in the corresponding sections in Chapters 4 to 8, and proposes future research directions that could tackle them.

Definitions and Measurement Instruments

"Define your terms [...] or we shall never understand one another," wrote the philosopher Voltaire in the eighteenth century (Voltaire, 1977). What was true then, still holds today, especially in scientific research. However, XAI seems to be struggling with definitions on both a *conceptual* and a *fundamental* level.

On a *conceptual* level, XAI researchers have not yet agreed upon definitions for many of the human-centred concepts they study, for example, model understanding or trust in AI systems. Also in our work, we have not built upon frameworks from the learning sciences to clearly define self-reflection and metacognition. This is problematic in the long run as it hampers comparing results and building on previous findings. We discuss the case of trust in more detail. Many papers claim to study trust, but upon closer inspection of their measurement instruments, they actually equate trust with confidence (Dasgupta et al., 2017), actual and perceived model accuracy (Chuang et al., 2012; Mohseni et al., 2020; Nourani et al., 2019), satisfaction (Gedikli et al., 2014), perceived transparency (Gedikli et al., 2014), or acceptance and rejection of model outcomes (Papenmeier et al., 2019; Yin et al., 2019; Zhang et al., 2020). The lack of widely-accepted definitions also results in a plethora of measurement instruments to assess XAI-related concepts. For example, some studies measure trust with a single Likert-type question (Bussone et al., 2015; Dasgupta et al., 2017; Holliday et al., 2016; Krause et al., 2018b; Millecamp et al., 2019; Nourani et al., 2020), implicitly assuming that trust is a monolithic concept. Other studies measure trust with multiple Likert-type questions, either ad hoc because the questions correlate or refer to a general definition for trust (Kizilcec, 2016; Uggirala et al., 2004; Yang et al., 2020a), or based on underlying theory of trust as a multidimensional concept constituted by constructs such as competence, benevolence, and integrity (Bayer et al., 2022; Cramer et al., 2008; Dikmen and Burns, 2022). In the latter case, however, there is again "little agreement on the specific constructs that constitute trust" (McKnight et al., 2002). Yet, since we feel multidimensional measurements are the most precise and insightful, we adopted this approach in our work in Chapters 5 to 8. While our operationalisation of trust seems a step in the right direction, it does not

directly account for the core concept of vulnerability (Vereschak et al., 2021), implying that our results might not transfer to application contexts with higher stakes.

On a *fundamental* level, XAI lacks widely-accepted definitions for concepts at its core, including transparency and explanations (Doshi-Velez and Kim, 2017; Lipton, 2018). In addition, remember from Section 2.3 that XAI researchers often use terms such as 'explainability', 'interpretability', 'transparency', 'understandability', 'intelligibility', 'explicability', and 'comprehensibility' interchangeably (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020). Possibly more concerning is that it might be misleading to say XAI methods are "explaining" AI models (Rudin, 2019): the algorithmic XAI approaches in Section 2.4 are not faithful to the models they "explain" and may thus inaccurately represent them. In other words, "explanations" do not truly describe how AI models works. If we are to make decisions based on real-world data with AI models that already approximate the real world, we should question whether it is always appropriate to add a second layer of approximation with simplifying "explanations." Put more poetically: do we want a black box on top of a black box, and if so, when and what for?

The next subsections will elaborate on this question from different perspectives, but before proceeding, I wish to soften the critical reflections in this section a bit. In the end, it is very challenging to define human-centred aspects unambiguously, especially in an interdisciplinary field. Ending on a positive note, we are hopeful that the current Babel-like confusion of tongues will fade as XAI matures as a research field. Researchers have already been working towards more unified definitions and corresponding measurement instruments (Donoso-Guzmán et al., 2023; Gulati et al., 2017, 2019; Hoff and Bashir, 2015; Hoffman et al., 2019; Jacovi et al., 2021; Jian et al., 2000; Madsen and Gregor, 2000; Vereschak et al., 2021), and we foresee they will continue to do so in the future.

Goals for XAI Beyond Model Understanding and Trust

Much of the research presented in this thesis involved trust perceptions. However, our findings support researchers who argue that fostering (appropriate) trust should not be the only goal of XAI and there should be more focus on *utility* (Davis et al., 2020). This means the sum of humans and AI explanations should be bigger than the human or AI on their own: explanations should be designed such that they improve model debugging and validation, model selection, mental model and model understanding, executing tasks together with AI, and model steering (Davis et al., 2020). To expand upon this list, I propose XAI researchers should also study wider concepts such as metacognition and

motivation, which are very relevant goals for learners in education and patients who need to alter their lifestyle, for example.

In our work, we made first steps to study how visual explanations affect metacognition, motivation, learning performance, and enjoyment. While qualitative data in Chapters 7 and 8 showed hopeful signs for how XAI can stimulate learners' metacognition and motivation, we could not confirm this with quantitative data in a short-term randomised controlled experiment in Chapter 8 for the case of *what-if* explanations. Yet, much like trust evolves (Holliday et al., 2016; Nourani et al., 2020; Ooge and Verbert, 2021), motivation can rise and fall within individuals over time (Ryan, 2012). Future work should therefore study motivation trajectories in long-term experiments and investigate the relation with explanations and control mechanisms more carefully. To this end, ongoing collaborations with researchers from Carnegie Mellon University study more closely how *what-if* explanations affect motivation, metacognition, and mastery orientation under different learner control levels. The core idea regarding *what-if* explanations is to not only show potential improvement in learning in the best case, but also potential decline in the worst case, and the expected change based on learning analytics of similar learners. The rationale behind this set-up is to start exploring the trade-off between stimulating motivation while staying realistic: for example, *what-if* explanations that show large progress could be perceived as most motivating at first glance, but become demotivating once it becomes clear they set impossible goals. Overall, our initial studies argue for XAI beyond model understanding and trust, similar to recent research by Conijn et al. (2023).

Explanations Have Issues Too: Foster Cognitive Engagement

XAI is typically promoted as a technique to mitigate biases, and improve model understanding and trust. Even though these goals indeed seem reasonable, they obscure an important pitfall robustly found in recent work: explanations can lead to unwarranted trust or distrust in AI models (Liao and Varshney, 2022). For example, explanations can positively affect trust but also lead to over-reliance (Bussone et al., 2015) and reinforce cognitive biases (Bertrand et al., 2022); and AI novices can prefer explanation representations with which they perform worse (Szymanski et al., 2021). Thus, inadequate explanations can have adverse effects when people believe they understand AI systems even though they do not (Weber et al., 2021). This is related to what is called the "illusion of explanatory depth" (Rozenblit and Keil, 2002): people generally tend to overestimate how well they understand complex phenomena.

Fortunately, researchers are starting to address this problem by studying trust

calibration, i.e., the ways how people distinguish when to trust or distrust AI systems (Han and Schulz, 2020; Zhang et al., 2020). It seems errors in trust calibration occur when people do not understand how systems work, do not know their capabilities, are overwhelmed, lack situation awareness, or feel a loss of control (Naiseh et al., 2021). In such situations, people are not cognitively engaging with explanations to build correct mental models and calibrate their trust. This aligns with a more general tendency: people are often reluctant to engage in what they perceive as effortful (Kool and Botvinick, 2018), resulting in less-informed trust decisions (Naiseh et al., 2021). Admittedly, our work in Chapters 6 to 8 also did not consider whether adolescents really cognitively engaged with our visual explanations. Specifically, we did not use eye-tracking to validate that adolescents indeed analysed them, and we did not measure their understanding of the recommender system through answering questions about the recommendation model. In our defence, determining whether trust in recommender systems is warranted seems challenging without a ground truth for 'good' recommendations.

Overall, this discussion motivates why we as XAI researchers should be more careful with how we frame the benefits of XAI. Too often, papers mention slogans similar to "we need explanations to increase trust and enable humans to understand and appropriately trust AI." Yet, there is a difference between what we ideally hope to achieve and what experiments with real people find. To avoid biases, explanations should be cognitively engaging and to realise that, the way forward is making explanations interactive and combining them with control mechanisms, similar to the approach in Chapter 8. The next subsection elaborates on a more general framework that supports this belief.

Recommendation-Driven vs Hypothesis-Driven Explaining

Current explanations are typically static (Abdul et al., 2018) and one-off, not considering user input or preferences beyond initial configurations (Sokol and Flach, 2020). Yet, this mismatches with how people justify things in conversations: this typically happens iteratively, with participants frequently asking questions similar to "do you understand?" (Hind, 2019). While Miller (2019) has already pointed out this interactive, dialogue-like nature of explanations several years ago, XAI techniques seem to lag behind. This leads to situations where AI complemented with explanations seems like a one-way road where AI models try to persuade people to adopt its outcomes. Miller (2023) calls this the *recommendation*-driven paradigm and Van Cauwenberge et al. (2022) captured this paradigm beautifully as: many roads lead to Rome, but AI only shows one road.

Recently, Miller (2023) has argued that the recommendation-driven way of implementing XAI is "dead." Instead, he pushed the idea of dialogue-like explanations and control over decision-making further by arguing that XAI should move to a *hypothesis*-driven paradigm for decision support. Concretely, Miller proposed that XAI should no longer focus on justifying AI recommendations in the way we also pursued in Chapter 6. Instead, the focus should lie on generating and presenting evidence that supports or refutes human judgements, explaining trade-offs between different options. This puts the control over which hypotheses are investigated back into the hands of human decision-makers, essentially turning around the idea of trying to bring "humans in the loop" into "machines in the loop" (Green and Chen, 2019) during decision-making. Overall, this aligns with the challenge on interactive, cognitively engaging explanations, discussed in the previous subsection. Thus, it seems vital to study why, when, and for whom it is desirable to design explainable AI systems according to a hypothesis-driven paradigm.

The question remains how the idea of a hypothesis-driven paradigm can be translated into practice. Research has started to work in this direction by trying to nudge people to engage deeper in System 2 thinking (Liao and Varshney, 2022), i.e., slower and analytical thinking (Kahneman, 2011). For example, Buçinca et al. (2021) used different strategies to force people into engaging more thoughtfully with explanations and found this reduced overreliance on an AI model. Our work in Chapter 8 has also contributed to this line of research by interactively combining learner control with explanations. Future research could push our idea further by not providing a default value on the control slider and forcing learners to actively consider which difficulty level suits them. From an interaction point of view, conversational techniques could be promising to actively engage people with explanations through natural language dialogues (Lakkaraju et al., 2022). Combining this with visualisations could combine the best of both worlds, which is why future collaborations with researchers from Hong Kong Baptist University will study how visualisation-supported chatbots impact people's decision-making.

9.4 Taking a Final Step Back

This thesis showed how XAI plays a pivotal role within the broader AI ecosystem, offering people insights into the decision-making process of complex AI models. However, we should reflect upon how the XAI techniques we design and evaluate can be integrated into real applications. For example, it seems plausible that major tech companies will not voluntarily embrace transparency solutions that could reveal potential biases in their products as it may jeopardise their revenues.

Furthermore, another challenge is ensuring that the general public gains more awareness of AI's potential pitfalls: automation bias seems deeply rooted and it is unclear whether explanations alone can counter this. In this respect, education plays a vital role in nurturing people's realistic expectations of AI.

Zooming out more, explanations and control mechanisms alone seem insufficient to solve the entirety of AI's ethical and practical challenges. Therefore, the ultimate goal is not providing perfect explanations, but ensuring what is nowadays called *reliable, safe, and trustworthy AI* (Shneiderman, 2020). This acknowledges that besides explainability, there is a need for robustness, safety, fairness, accountability, privacy, and data governance (Hamon et al., 2022). It is in this broader interdisciplinary framework that we should critically study trade-offs between different goals. For example, there is a trade-off between transparency and security as explanations are potentially susceptible to extraction attacks that expose privacy (Yan et al., 2022), and transparency and accountability might conflict (Ananny and Crawford, 2018; Lima et al., 2022).

Overall, it seems the rapid pace with which AI evolves brought us to a point where whole societies are reflecting more about aspects we probably should have thought about for much longer. Which metrics should we use to assess technology? Should we launch AI on massive scales without properly testing it or thoroughly considering its societal impacts? Is it acceptable to tolerate huge power imbalances regarding massive data collection and deployment of incredibly complex AI models? Is it worth deploying AI systems that consume enormous amounts of energy and thus contribute to ruining the world-wide climate? Is automation even the answer to all problems? While XAI is but one piece of the complex AI puzzle and certainly has its own set of limitations, I find it truly inspiring to witness and contribute to the interdisciplinary efforts towards augmenting human capabilities with AI. Especially XAI in the sense of a hypothesis-driven paradigm seems to have a bright future.

Acknowledgements

The past four years have undoubtedly been the most exciting years of my life. Pursuing a PhD offered me the privilege to live my passion for research and education, travel around the globe, and learn from hundreds of talented people. I have experienced periods of extreme joy and fulfilment and have grown as a writer, speaker, mentor, teacher, programmer, ambassador, and analyst. But the past four years have also been turbulent ones. The PhD constantly challenged me to improve my work and myself without slipping into self-doubt, exhaustion, or stagnation. I want to wholeheartedly thank everyone who raised my enthusiasm and energy levels, shared moments of sheer happiness, and helped me balance perfectionism and acceptance.

Thank you to my family. My dear parents, grandparents, and sister Joke. Your decade-long support and sacrifices allow me to follow my heart. Without the safe and stimulating environment you create, I wouldn't be as brave and free of serious sorrows, and an academic career would never have been an option. I know you often need to miss me because of my career choices, but I will always be there if you need me. Yens, five years ago, I asked you whether you were ready for me, and I couldn't believe you answered "yes." But time and again, you proved that you were: when you cycled me to the hospital after the broccoli debacle, when you survived months of lockdown in a one-person studio with me, when you held me in moments of anxiety and mental breakdown. You bring out the best in me, and I hope I can support your personal growth and PhD journey as well as you have mine.

Thank you to everyone on my examination committee. Katrien, I am grateful that you were my promotor. From the start, you provided a safe financial environment, plenty of research opportunities, and the space to work how and when I felt most comfortable. I am especially thankful that you didn't prey upon my perfectionism and work ethos and were generously sharing your academic network and experience. Tinne and Vero, thank you for following my PhD trajectory all the way and acting as constructively critical yet encouraging

sounding boards. Denis and Tias, thank you for providing an algorithmic counterweight to the human-centred armada.

Thank you to my direct colleagues in the Augment lab: Robin for your hands-on advice and positive feedback, Houda for enlightening my days with your laughter, Ivania for bringing enthusiasm and fun into the office, Aditya for ensuring a continuous stream of Darjeeling tea, Maxwell for your positivism and casual after-work meetups. Nyi-Nyi, Martijn, Francisco, Tom, Diego, Arno, Leen, Raphael: I missed your presence in the lab after your leave and still cherish all our moments together. Also, thanks to my many other friends at KU Leuven: Mingxiao, sharing our office has been a pleasure, and I am grateful for all the lovely talks we had; Jihae, Robbe, Song, Taiyu, Adal, you made me feel at home in a second lab and our adventures abroad and lunch/tea breaks fed my moral; Shirin your warmth and passion for research made me feel understood; Thomas, your drive inspired me; Lucy, you were a fantastic Easter bunny and chess master; Alex, you are even more fun than your musical wall.

During my travels abroad and conferences, I met so many wonderful people and shared so many beautiful moments with them that it would take dozens of pages to name everyone who left an impression on me. Therefore, a shoutout to the friends with whom I spent the most time: Viva, you're a rockstar, and I wish the CHI conferences weren't the only opportunity for us to hang out together in person so I could enjoy more of your humour and wisdom; Clara, what you did for me in Sydney moved me, and I loved our trip in your hometown Paris. Special thanks to everyone who surrounded and supported me during my unforgettable stays in Maribor, Hong Kong and Pittsburgh: Gregor Stiglic, Lucija Gosak, Primoz Kocbek, Li Chen, Yucheng Jin, Xianglin Zhao, Weixin Chen, Jingwen Xu, Xinglin Pan, Vincent Aleven, Conrad Borchers, Kexin Yang, Meng Xia, Ken Holstein, John Stamper, Adam Perer, Youli Chang, Katelyn Morrisson, Peter Brusilovsky, and Kamil Akhuseyinoglu. Jeremiah and Ethan, thanks for the dynamic hackathon adventure. Jordan, my stay in Pittsburgh wouldn't have been the same without you, and I owe you so much for all your generosity and friendship. Deniz and Mesut Erhan, your house was my safe haven and without it, my thesis wouldn't have been of the quality I pursued.

I thank all students who participated in my classes and renewed my energy with their enthusiasm. Sho, Kenan, Mario, Jeffrey, Barbara, Leen, Anissa, Joran: it was a pleasure guiding your master theses and I'm so grateful for all the things you taught me. I also thank the hundreds of study participants whose time and feedback form the base of my research achievements, and the often-forgotten people who work behind the scenes to manage the Department of Computer Science at KU Leuven. Special thanks to Karin Michiels for processing my dozens of receipts for reimbursement and arranging flights and hotels. Finally, I thank you, dear reader. I hope my work and experiences inspire you.

Appendix A

Questionnaires and Details

A.1 Pre- and Post-Study Questionnaires

Table A.1: The questionnaire that participants answered at the end of the study in Chapter 6. All questions were evaluated on a 7-point range. The group names in italics are for reference; participants did not see them. After each group, participants could motivate their answers and give additional comments in a text field.

No.	English original	Dutch translation
Competence		
Q1	Wiski is like an expert (for example, a teacher) for recommending math exercises.	Wiski is zoals een expert (bv. een leerkracht) in wiskunde-oefeningen aanraden.
Q2	Wiski has the expertise (knowledge) to estimate my math level.	Wiski heeft de expertise (kennis) om mijn wiskundeniveau te kunnen inschatten.
Q3	Wiski can estimate my math level.	Wiski kan mijn wiskundeniveau inschatten.
Q4	Wiski understands the difficulty level of math exercises well.	Wiski begrijpt de moeilijkheidsgraad van wiskunde-oefeningen goed.
Q5	Wiski takes my math level into account when recommending exercises.	Wiski houdt rekening met mijn wiskundeniveau om oefeningen aan te raden.
Benevolence		
Q6	Wiski prioritizes that I improve in math.	Wiski zet op de eerste plaats dat ik vorderingen maak in wiskunde.
Q7	Wiski recommends exercises so that I improve in math.	Wanneer Wiski oefeningen aanraadt, doet Wiski dat zodat ik vorderingen maak in wiskunde.

Continued on next page

Table A.1 – Continued from previous page

No.	English original	Dutch translation
Q8	Wiski wants to estimate my math level well.	Wiski wil mijn wiskundeniveau goed inschatten.
Integrity		
Q9	Wiski recommends exercises as correctly as possible.	Wiski raadt oefeningen op een zo correct mogelijke manier aan.
Q10	Wiski is honest.	Wiski is eerlijk.
Q11	Wiski makes integrous recommendations.	Wiski maakt oprechte aanbevelingen.
Trust (one-dimensional)		
Q12	I trust Wiski to recommend me math exercises.	Ik vertrouw Wiski om mij wiskunde-oefeningen aan te raden.
Intention to return		
Q13	If I want to solve math exercises again, I will choose Wiski.	Als ik nog eens online wiskunde-oefeningen maak, dan kies ik voor Wiski.
Q14	If I want to be recommended math exercises again, I will choose Wiski.	Als ik nog eens wiskunde-oefeningen aangeraden wil krijgen, dan kies ik voor Wiski.
Perceived transparency		
Q15	I find that Wiski gives enough explanation as to why an exercise has been recommended.	Ik vind dat Wiski genoeg uitleg geeft over waarom een oefening aangeraden is.
General questions		
Q16	I do NOT want any explanations about why an exercise has been recommended when I use Wiski.	Wanneer ik Wiski gebruik, wil ik GEEN uitleg over waarom een oefening wordt aangeraden.
Q17	I find an explanation for why an exercise is recommended more important than for why a movie is recommended.	Ik vind uitleg krijgen over waarom een oefening wordt aangeraden belangrijker dan waarom een film wordt aangeraden.
Q18	I am NOT happy with the level of math exercises Wiski recommended.	Ik ben NIET blij met het niveau van de oefeningen die Wiski aanraadde.
Q19	I find it important to receive explanations when something (exercise/movie/product/...) has been recommended.	In het algemeen vind ik het belangrijk om uitleg te krijgen wanneer iets (oefening/-film/product/...) wordt aangeraden.

Table A.2: The questionnaire that participants filled out at the end of the study in Chapter 7. All questions were evaluated on a 7-point range, and questions Q19, Q20, and Q25 were reverse-scored. The italic group names are for reference; participants did not see them.

No.	English version	Dutch version
Competence		
Q1	Wiski is like an expert (for example, a teacher) for recommending maths exercises.	Wiski is zoals een expert (bv. een leerkracht) in wiskunde-oefeningen aanraden.

Continued on next page

Table A.2 – Continued from previous page

No.	English version	Dutch version
Q2	Wiski has the expertise (knowledge) to estimate my maths level.	Wiski heeft de expertise (kennis) om mijn wiskundeniveau te kunnen inschatten.
Q3	Wiski can estimate my maths level.	Wiski kan mijn wiskundeniveau inschatten.
Q4	Wiski understands the difficulty level of maths exercises well.	Wiski begrijpt de moeilijkheidsgraad van wiskunde-oefeningen goed.
Q5	Wiski takes my maths level into account when recommending exercises.	Wiski houdt rekening met mijn wiskunde-niveau om oefeningen aan te raden.
Benevolence		
Q6	Wiski prioritises that I improve in maths.	Wiski zet op de eerste plaats dat ik vorderingen maak in wiskunde.
Q7	Wiski recommends exercises so that I improve in maths.	Wanneer Wiski oefeningen aanraadt, doet Wiski dat zodat ik vorderingen maak in wiskunde.
Q8	Wiski wants to estimate my maths level well.	Wiski wil mijn wiskundeniveau goed inschatten.
Integrity		
Q9	Wiski recommends exercises as correctly as possible.	Wiski raadt oefeningen op een zo correct mogelijke manier aan.
Q10	Wiski is honest.	Wiski is eerlijk.
Q11	Wiski makes integrous recommendations.	Wiski maakt oprechte aanbevelingen.
Trust (one-dimensional)		
Q12	I trust Wiski to recommend me maths exercises.	Ik vertrouw Wiski om mij wiskunde-oefeningen aan te raden.
Intention to return		
Q13	If I want to solve maths exercises again, I will choose Wiski.	Als ik nog eens online wiskunde-oefeningen maak, dan kies ik voor Wiski.
Q14	If I want to be recommended maths exercises again, I will choose Wiski.	Als ik nog eens wiskunde-oefeningen aangeraden wil krijgen, dan kies ik voor Wiski.
Transparency		
Q15	I understood why the exercises were recommended to me.	Ik begreep waarom de oefeningen aan mij werden aanbevolen.
Q16	Wiski helps me understand why the exercises were recommended to me.	Wiski helpt mij te begrijpen waarom de oefeningen aan mij werden aanbevolen.
Q17	Wiski explains why the exercises are recommended to me.	Wiski legt uit waarom de oefeningen aan mij worden aanbevolen.
Control		
Q18	I feel in control of telling Wiski what I want.	Ik heb het gevoel dat ik Wiski kan vertellen wat ik wil.
Q19	I don't feel in control of telling Wiski what I want.	Ik heb niet het gevoel dat ik Wiski kan vertellen wat ik wil.
Q20	I don't feel in control of specifying and changing my preferences.	Ik heb niet het gevoel dat ik controle heb over het omschrijven en veranderen van mijn voorkeuren.

Continued on next page

Table A.2 – Continued from previous page

No.	English version	Dutch version
Q21	Wiski seems to control my decision process rather than me.	Wiski lijkt mijn keuzeproces te controleren in plaats van ikzelf.
Preference elicitation		
Q22	Wiski provides an adequate way for me to express my preferences.	Wiski laat me op een geschikte manier mijn voorkeuren aangeven.
Q23	I found it easy to tell Wiski about my preferences.	Ik vond het gemakkelijk om Wiski over mijn voorkeuren te vertellen.
Q24	It is easy to learn to tell Wiski what I like.	Het is gemakkelijk om te leren hoe ik Wiski kan vertellen wat ik leuk vind.
Q25	It required too much effort to tell Wiski what I like.	Het kostte te veel moeite om Wiski te vertellen wat ik leuk vind.
Preference revision		
Q26	Wiski provides an adequate way for me to revise my preferences.	Wiski laat me op een geschikte manier mijn voorkeuren aanpassen.
Q27	I found it easy to make Wiski recommend different things to me.	Ik vond het gemakkelijk om Wiski mij verschillende dingen te laten aanbevelen.
Q28	It is easy to train Wiski to update my preferences.	Het is gemakkelijk om Wiski te leren mijn voorkeuren aan te passen.
Q29	I found it easy to alter the recommended exercises due to my preference changes.	Ik vond het gemakkelijk om de aanbevolen oefeningen te wijzigen met mijn voorkeurs-veranderingen.
Q30	It is easy for me to inform Wiski if I dislike/like recommended exercises.	Het is voor mij gemakkelijk om Wiski te laten weten of ik de aanbevolen oefeningen leuk/niet leuk vind.
Q31	It is easy for me to get a new set of recommended exercises.	Het is voor mij gemakkelijk om een nieuwe reeks aanbevolen oefeningen te krijgen.

Table A.3: The pre- and post-study questionnaires that participants filled out at the end of the study in Chapter 8. All questions were evaluated on a 7-point range, except for ENDUR1–ENDUR5, for which a 5-point range was used. Questions ENDUR3 and MOTIV4 were reverse-scored. The italic group names are for reference; participants did not see them.

No.	Question	Load.	Com.
Intrinsic motivation ($\omega = 0.65$, CI $= [0.51, 0.73]$)			
SMS1	Because it gives me pleasure to learn	0.65	0.40
SMS2	Because it is very interesting to learn how I can improve	0.66	0.43
SMS3	Because I find it enjoyable to discover new things	0.66	0.42
SMS4	Because learning reflects the essence of whom I am	0.60	0.39
SMS6	Because learning is an integral part of my life	0.72	0.47
SMS7	Because it is one of the best ways I have chosen to develop other aspects of myself	0.66	0.43
SMS8	Because I have chosen learning as a way to develop myself	0.77	0.56

Continued on next page

Table A.3 – Continued from previous page

No.	Question	Load.	Com.
SMS9	Because I found it is a good way to develop aspects of myself that I value	0.57	0.43
SMS11	Because I feel better about myself when I learn	0.67	0.50
SMS5	Because through learning, I am living in line with my deepest principles		
SMS10	Because I would feel bad about myself if I did not take the time to learn		
SMS12	Because I would not feel worthwhile if I did not learn		
Extrinsic motivation ($\omega = 0.87$, CI $= [0.84, 0.90]$)			
SMS13	Because people I care about would be upset with me if I did not learn	0.57	0.37
SMS14	Because people around me reward me when I learn	0.48	0.23
SMS15	Because I think others would disapprove of me if I did not learn	0.82	0.61
SMS16	I used to have good reasons to learn, but now I am asking myself if I should continue		
SMS17	I don't know anymore; I have the impression that I am incapable of succeeding in learning		
SMS18	It is not clear to me anymore; I don't really think my place is in learning		
Trust (one-dimensional)			
1DT	I trust Wiski to recommend me maths exercises.		
Competence ($\omega = 0.81$, CI $= [0.73, 0.87]$)			
COMP2	Wiski has the knowledge to estimate my maths level	0.52	0.48
COMP3	Wiski takes my maths level into account when recommending exercises	0.74	0.76
BEN2	Wiski wants to estimate my maths level as well as possible	0.79	0.54
COMP1	Wiski is as good as a teacher in recommending exercises		
BEN1	Wiski prioritises that I improve in maths		
Intention to return ($\omega = 0.84$, CI $= [0.75, 0.90]$)			
ITR1	If I practice maths exercises online again and I want recommended exercises, I will choose Wiski	0.88	0.83
ITR2	I would use Wiski again in the future	0.79	0.64
Metacognition ($\omega = 0.87$, CI $= [0.80, 0.91]$)			
MCOGN1	This screen made me reflect upon my maths level	0.73	0.53
MCOGN2	This screen made me reflect upon whether I made progress	0.75	0.56
MCOGN3	This screen made me reflect upon how Wiski recommends exercises	0.73	0.54
MCOGN4	This screen made me reflect upon whether Wiski recommends suitable exercises	0.73	0.54
MCOGN5	This screen made me reflect upon whether I reach my learning goals	0.81	0.66
Endurability ($\omega = 0.78$, CI $= [0.65, 0.85]$)			
ENDUR2	I consider my experience with Wiski was a success	0.89	0.79
ENDUR4	My experience with Wiski was rewarding	0.69	0.48
ENDUR5	I would recommend practising on Wiski to my friends	0.64	0.41
ENDUR1	Practising on Wiski was worthwile		

Continued on next page

Table A.3 – Continued from previous page

No.	Question	Load.	Com.
ENDUR3	My experience with Wiski did not work out the way I had planned		
Motivation ($\omega = 0.86$, CI $= [0.80, 0.90]$)			
MOTIV1	Wiski motivated me to make more exercises than usual	0.88	0.78
MOTIV2	Because of Wiski I want to understand maths more	0.70	0.49
MOTIV3	I find practising with Wiski more fun than making exercises from a text book	0.79	0.63
MOTIV5	Wiski stimulated me to put more effort into maths	0.73	0.53
MOTIV4	I did not find Wiski motivating		

A.2 Elo Rating System

Our implementation of the Elo rating system in Chapter 8 was heavily inspired by the variant typical for chess ratings. The Beginner–Expert levels were inspired by the Dreyfus model (Dreyfus, 2004) and their range corresponded to the interval $[1000, 2000]$, which roughly corresponds to typical Elo scores for novice (1000) and expert (2000) chess players.

As explained in Section 7.2.4, Elo ratings are updated each time a learner l answers an exercise e:

$$\text{Elo}(l) = \text{Elo}(l) + k_{\text{learner}}\,(X_{le} - P(X_{le} = 1)),$$

$$\text{and} \quad \text{Elo}(e) = \text{Elo}(e) - k_{\text{exercise}}\,(X_{le} - P(X_{le} = 1)).$$

While $k_{\text{learner}} = k_{\text{exercise}}$ in Chapters 6 and 7, our implementation now used different values, depending on how many exercises l had solved and l's Elo rating:

```
if (number_solved < 20 & elo_learner < 2000) {
  k_learner = 40;
} else if (elo_learner < 2000) {
  k_learner = 20;
} else {
  k_learner = 10;
}

if (number_solved<10) {
  k_exercise = 40;
} else if (number_solved<20) {
  k_exercise = 20;
} else {
  k_exercise = 10;
```

}

While these parameters can be optimised (Wauters et al., 2010), we inspired our adaptation scheme on values proposed by the international chess federation FIDE. Further improving upon Chapters 6 and 7, we computed the probability $P(X_{le} = 1)$ considering that exercises have a multiple-choice format (Pelánek, 2016):

$$P(X_{le} = 1) = \frac{1}{k} + \left(1 - \frac{1}{k}\right) \frac{1}{1 + 10^{-(\text{Elo}(l) - \text{Elo}(e))/400}}, \tag{A.1}$$

where k is the number of options in exercise e.

A.3 Wise Feedback

Table A.4: Variants of the wise feedback on our platform in Chapter 8 for different intervals of difficulty levels. Each variant conveys high standards for the learners' performance, but also includes a belief in their potential to reach that standard (Yeager et al., 2017).

Difficulty	Wise Feedback
$[0.0, 0.2]$	I think you can handle exercises that are much harder. I expect a lot from you and am sure you can do it! This level is very low for you. I expect you can solve a more difficult series. I believe in you! These are very easy exercises. You can surely make them, but I believe you can handle a harder level. Then you'll grow faster!
$[0.2, 0.4]$	I believe you can handle exercises that are more difficult. I believe you can grow even further! I think you can handle more difficult exercises. That way you will grow faster! You can definitely solve easier exercises, but I believe you can handle slightly more difficult exercises. You can do it!
$[0.4, 0.6]$	I think you can handle this difficulty for sure. Maybe you could choose a slightly higher difficulty to get even better! This is a level I believe you can handle. Maybe you can choose a slightly harder level to grow faster. I believe you can solve these exercises, but maybe you could set the difficulty slightly higher. That way you'll get even better!
$[0.6, 0.8]$	This difficulty is challenging, but the bar is high and I trust in your abilities! This is a slightly more challenging level, but I definitely believe you can solve these exercises correctly! I trust that you can solve these difficult exercises. That way, you will also grow faster.
$[0.8, 1.0]$	This difficulty seems very challenging for you. If you think you can handle it, I totally support you! Wow, a challenge! You can always try, I believe in you!

Continued on next page

Table A.4 – Continued from previous page

Difficulty	Wise Feedback
	This is a very difficult level. But if you think you can handle the exercises, I totally believe in you!

A.4 Learning Performance Correction

In Chapter 8, we did not measure participants' learning performance solely in terms of correct and wrong answers because this does not consider the exercises' difficulty. For example, suppose learner A and B both solved three exercises correctly, where A's exercises were easy and B's were hard (difficulty is estimated by the exercises' Elo ratings). We opted to assess B's performance as higher and thus needed to think of a correction strategy.

We initially planned to average participants' Elo ratings for the topics they practised, but that strategy was suboptimal as participants had not practised long enough for their Elo ratings to converge. Moreover, Elo ratings only have relative meaning, so a rating of, say, 1000 does not have a pedagogy-based "easy" or "hard" interpretation. Correcting based on topic difficulties did not work either: topics proved to have exercises of varying difficulty, which made averaging per topic rather useless as all topics got similar average difficulties (see Figure A.1). Therefore, our performance score for each learner is the average of their performance on all the exercises they solved, corrected by those exercises' difficulty. The correction is based on the *cumulative density function* (cdf) of exercises' difficulty as measured by their Elo ratings and is depicted in Figure A.2. Specifically, a pair (z, p) of a difficulty $z \in [1000, 2000]$ and performance $p \in \{0, 1\}$ is transformed as follows:

$$T(z, p) = \big(z, (1 - \beta)\, p + \alpha\, \mathrm{cdf}(z)\big),$$

where α and β are freely chosen parameters. We chose $\alpha = \beta = 1/4$ because those values seemed reasonable, but future work can assign different values and even choose $\alpha \neq \beta$. Under our transformation, a wrong answer for the hardest exercise would yield a performance of $1/4$ instead of 0, and a correct answer for the easiest exercise $3/4$ instead of 1. Finally, a learner's corrected performance

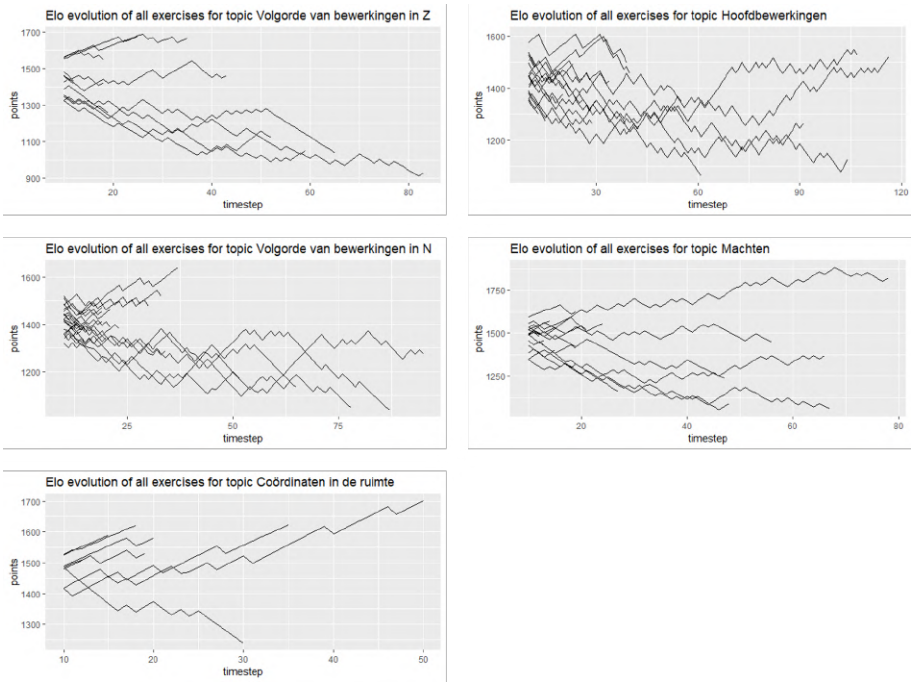


Figure A.1: Elo rating evolutions of exercises in the five most practised maths topics. There is a wide variety in exercises' difficulty.

would then become:

$$\begin{aligned}\text{corrected performance} &= \frac{1}{n}\sum_{k=1}^{n}\big((1-\beta)\,p_k + \alpha\,\mathrm{cdf}(z_k)\big)\\ &= (1-\beta)\,\bar{p} + \frac{\alpha}{n}\sum_{k=1}^{n}\mathrm{cdf}(z_k),\end{aligned}$$

where $\bar{p}$ is the average performance over all exercises $e_1, \ldots, e_n$ solved by that learner, with respective difficulties $z_1, \ldots, z_n$ and performances $p_1, \ldots, p_n$.

For the cdf, we only considered exercises that were solved at least 10 times and for which the Elo ratings were rather converged, i.e., the last 10 ratings were maximally 200 apart. As a consequence, we could not assess the performance for 9 participants who had only solved exercises that did not meet our restrictions.

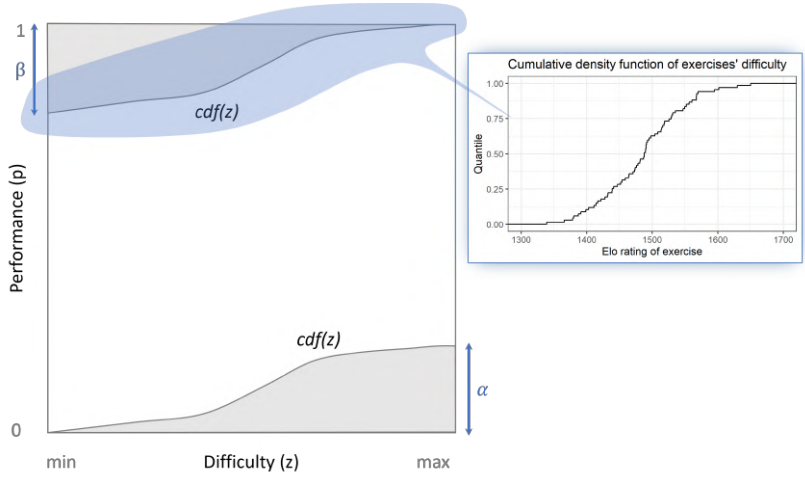


Figure A.2: The cumulative density function for the exercises' Elo ratings in our experiment in Chapter 8.

Bibliography

Abbasloo, A., Wiens, V., Schmidt-Wilcke, T., Sundgren, P., Klein, R., and Schultz, T. (2019). Interactive formation of statistical hypotheses in diffusion tensor imaging. In *Eurographics Workshop on Visual Computing for Biology and Medicine, VCBM 2019*, pages 33–43.

Abdi, S., Khosravi, H., Sadiq, S., and Gasevic, D. (2019). A Multivariate Elo-Based Learner Model for Adaptive Educational Systems. In *International Educational Data Mining Society*, page ED599177, Montreal, Canada. International Educational Data Mining Society.

Abdi, S., Khosravi, H., Sadiq, S., and Gasevic, D. (2020). Complementing educational recommender systems with open learner models. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 360–365. Association for Computing Machinery, New York, NY, USA.

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Abdullah, S., Rostamzadeh, N., Sedig, K., Garg, A., and McArthur, E. (2020). Visual analytics for dimension reduction and cluster analysis of high dimensional electronic health records. *Informatics*, 7(2).

Abras, C., Maloney-krichmar, D., and Preece, J. (2004). User-Centered Design. In *In Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*. Publications.

Accorsi, P., Lalande, N., Fabrègue, M., Braud, A., Poncelet, P., Sallaberry, A., Bringay, S., Teisseire, M., Cernesson, F., and Le Ber, F. (2014). HydroQual: Visual analysis of river water quality. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 123–132.

Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Afchar, D., Melchiorre, A., Schedl, M., Hennequin, R., Epure, E., and Moussallam, M. (2022). Explainability in Music Recommender Systems. *AI Magazine*, 43(2):190–208.

Afzal, S., Maciejewski, R., and Ebert, D. (2011). Visual analytics decision support environment for epidemic modeling and response evaluation. In *VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings*, pages 191–200.

Aher, S. B. and Lobo, L. M. R. J. (2013). Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. *Knowledge-Based Systems*, 51:1–14.

Ahmad, M. A., Eckert, C., Teredesai, A., and McKelvey, G. (2018). Interpretable Machine Learning in Healthcare. *IEEE Intelligent Informatics Bulletin*, 19(1):1–7.

Akçapınar, G., Altun, A., and Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, 16(1):40.

Ali, M., Alqahtani, A., Jones, M., and Xie, X. (2019). Clustering and Classification for Time Series Data in Visual Analytics: A Survey. *IEEE Access*, 7:181314–181338.

Alsaad, R., Malluhi, Q., Janahi, I., and Boughorbel, S. (2019). Interpreting patient-Specific risk prediction using contextual decomposition of BiLSTMs: Application to children with asthma. *BMC Medical Informatics and Decision Making*, 19(1).

Amatriain, X., Pujol, J. M., Tintarev, N., and Oliver, N. (2009). Rate it again: Increasing recommendation accuracy by user re-rating. In *Proceedings of the Third ACM Conference on Recommender Systems - RecSys '09*, page 173, New York, New York, USA. ACM Press.

Ananny, M. and Crawford, K. (2018). Seeing without knowing: Limitations of the transparency and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989.

Andjelkovic, I., Parra, D., and O'Donovan, J. (2016). Moodplay: Interactive Mood-based Music Discovery and Recommendation. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 275–279, Halifax Nova Scotia Canada. ACM.

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., and Mooney, C. (2021). Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences*, 11(11):5088.

Armstrong, L. J. and Nallan, S. A. (2016). Agricultural decision support framework for visualisation and prediction of Western Australian crop production. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1907–1912.

Augasta, M. G. and Kathirvalavakumar, T. (2012). Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. *Neural Processing Letters*, 35(2):131–150.

Ayoub Shaikh, T., Rasool, T., and Rasheed Lone, F. (2022). Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Computers and Electronics in Agriculture*, 198:107119.

Badam, S. K., Zhao, J., Sen, S., Elmqvist, N., and Ebert, D. (2016). TimeFork: Interactive Prediction of Time Series. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5409–5420, San Jose California USA. ACM.

Bahri, A. and Corebima, A. D. (2015). THE CONTRIBUTION OF LEARNING MOTIVATION AND METACOGNITIVE SKILL ON COGNITIVE LEARNING OUTCOME OF STUDENTS WITHIN DIFFERENT LEARNING STRATEGIES. *Journal of Baltic Science Education*, 14(4):487–500.

Bakken, S. (2019). Advancing biomedical and health informatics knowledge through reviews of existing research. *Journal of the American Medical Informatics Association*, 26(4):273–275.

Bandura, A., editor (1995). *Self-Efficacy in Changing Societies*. Cambridge University Press, Cambridge ; New York.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6):574–594.

Barlowe, S., Yang, J., Jacobs, D., Livesay, D., Alsakran, J., Zhao, Y., Verma, D., and Mottonen, J. (2013). A visual analytics approach to exploring protein flexibility subspaces. In *IEEE Pacific Visualization Symposium*, pages 193–200.

Barocas, S., Selbst, A. D., and Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of*

the 2020 Conference on Fairness, Accountability, and Transparency, pages 80–89, Barcelona Spain. ACM.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Barria-Pineda, J. (2020). Exploring the Need for Transparency in Educational Recommender Systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 376–379. Association for Computing Machinery, New York, NY, USA.

Barria-Pineda, J. and Brusilovsky, P. (2019). Making Educational Recommendations Transparent through a Fine-Grained Open Learner Model. *Los Angeles*.

Barria-Pineda, J., Guerra-Hollstein, J., and Brusilovsky, P. (2018). A Fine-Grained Open Learner Model for an Introductory Programming Course. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, pages 53–61, New York, NY, USA. Association for Computing Machinery.

Basole, R. C., Bellamy, M. A., and Park, H. (2017). Visualization of Innovation in Global Supply Chain Networks. *Decision Sciences*, 48(2):288–306.

Bayer, S., Gimpel, H., and Markgraf, M. (2022). The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, 32(1):110–138.

Behrisch, M., Krüeger, R., Lekschas, F., Schreck, T., Gehlenborg, N., and Pfister, H. (2018). Visual Pattern-Driven Exploration of Big Data. In *2018 International Symposium on Big Data Visual and Immersive Analytics, BDVA 2018*.

Berkovsky, S., Taib, R., and Conway, D. (2017). How to Recommend? User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 287–300, New York, NY, USA. Association for Computing Machinery.

Bertrand, A., Belloum, R., Eagan, J. R., and Maxwell, W. (2022). How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 78–91, Oxford United Kingdom. ACM.

Bertrand, A., Eagan, J. R., and Maxwell, W. (2023). Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 943–958, New York, NY, USA. Association for Computing Machinery.

Bhattacharya, A., Ooge, J., Stiglic, G., and Verbert, K. (2023). Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 204–219, New York, NY, USA. Association for Computing Machinery.

Bitew, S. K., Hadifar, A., Sterckx, L., Deleu, J., Develder, C., and Demeester, T. (2022). Learning to Reuse Distractors to Support Multiple Choice Question Generation in Education. *IEEE Transactions on Learning Technologies*, pages 1–16.

Bodily, R., Ikahihifo, T. K., Mackley, B., and Graham, C. R. (2018a). The design, development, and implementation of student-facing learning analytics dashboards. *Journal of Computing in Higher Education*, 30(3):572–598.

Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., and Verbert, K. (2018b). Open learner models and learning analytics dashboards: A systematic review. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 41–50, Sydney New South Wales Australia. ACM.

Boggust, A., Hoover, B., Satyanarayan, A., and Strobelt, H. (2022). Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–17, New York, NY, USA. Association for Computing Machinery.

Bögl, M., Aigner, W., Filzmoser, P., Gschwandtner, T., Lammarsch, T., Miksch, S., and Rind, A. (2014). Visual analytics methods to guide diagnostics for time series model predictions. In *Proceedings of the 2014 IEEE VIS Workshop on Visualization for Predictive Analytics*, volume 1.

Bollen, D., Knijnenburg, B. P., Willemsen, M. C., and Graus, M. (2010). Understanding choice overload in recommender systems. In *Proceedings of the Fourth ACM Conference on Recommender Systems - RecSys '10*, page 63, Barcelona, Spain. ACM Press.

Bonnett, L. J., Snell, K. I. E., Collins, G. S., and Riley, R. D. (2019). Guide to presenting clinical prediction models for use in clinical settings. *BMJ*, page l737.

Borland, D., Wang, W., Zhang, J., Shrestha, J., and Gotz, D. (2020). Selection bias tracking and detailed subset comparison for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):429–439.

Bostandjiev, S., O'Donovan, J., and Höllerer, T. (2012). TasteWeights: A visual interactive hybrid recommender system. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 35–42, New York, NY, USA. Association for Computing Machinery.

Botha, C. P., Preim, B., Kaufman, A., Takahashi, S., and Ynnerman, A. (2012). From individual to population: Challenges in Medical Visualization. *arXiv:1206.1148 [physics]*.

Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., and Detyniecki, M. (2022). Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th International Conference on Intelligent User Interfaces*, pages 807–819, Helsinki Finland. ACM.

Brachman, M., Pan, Q., Do, H. J., Dugan, C., Chaudhary, A., Johnson, J. M., Rai, P., Chakraborti, T., Gschwind, T., Laredo, J. A., Miksovic, C., Scotton, P., Talamadupula, K., and Thomas, G. (2023). Follow the Successful Herd: Towards Explanations for Improved Use and Mental Models of Natural Language Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 220–239, New York, NY, USA. Association for Computing Machinery.

Branwen, G. (2011). The Neural Net Tank Urban Legend.

Braun, V. and Clarke, V. (2012). Thematic analysis. In *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, APA Handbooks in Psychology®, pages 57–71. American Psychological Association, Washington, DC, US.

Braun, V., Clarke, V., Hayfield, N., and Terry, G. (2018). Thematic Analysis. In Liamputtong, P., editor, *Handbook of Research Methods in Health Social Sciences*, pages 1–18. Springer Singapore, Singapore.

Britton, E., Fisher, P., and Whitley, J. (1998). Quarterly Bulletin February 1998. Technical report, Bank of England.

Brockwell, P. J. and Davis, R. A. (2016). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer International Publishing, Cham.

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., and McClelland, I. L., editors, *Usability Evaluation in Industry*, volume 189. Taylor & Francis, UK, London.

Brunker, A., Catchpoole, D., Kennedy, P., Simoff, S., and Nguyen, Q. (2019). Two-dimensional immersive cohort analysis supporting personalised medical treatment. In *Proceedings - 2019 23rd International Conference in Information Visualization - Part II, IV-2 2019*, pages 34–41.

Brusilovsky, P. (2023). AI in Education, Learner Control, and Human-AI Collaboration. *International Journal of Artificial Intelligence in Education*.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):188:1–188:21.

Buijsman, S. (2020). *AI: Alsmaar Intelligenter*. Bezige Bij b.v., Uitgeverij De.

Bull, S. (2020). There are Open Learner Models About! *IEEE Transactions on Learning Technologies*, 13(2):425–448.

Bull, S. and Kay, J. (2007). Student Models that Invite the Learner In: The SMILI:() Open Learner Modelling Framework. *International Journal of Artificial Intelligence in Education*, 17(2):89–120.

Bull, S. and Kay, J. (2010). Open Learner Models. In Kacprzyk, J., Nkambou, R., Bourdeau, J., and Mizoguchi, R., editors, *Advances in Intelligent Tutoring Systems*, volume 308, pages 301–322. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bull, S. and McKay, M. (2004). An open learner model for children and teachers: Inspecting knowledge level of individuals and peers. In *International Conference on Intelligent Tutoring Systems*, pages 646–655. Springer.

Bull, S. and Pain, H. (1995). 'Did I say what I think I said, and do you agree with me?': Inspecting and Questioning the Student Model. page 13.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):205395171562251.

Bussone, A., Stumpf, S., and O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169, Dallas, TX, USA. IEEE.

Caban, J. J. and Gotz, D. (2015). Visual analytics in healthcare – opportunities and research challenges. *Journal of the American Medical Informatics Association*, 22(2):260–262.

Cai, C. J., Jongejan, J., and Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 258–262, Marina del Ray California. ACM.

Cao, N., Gotz, D., Sun, J., and Qu, H. (2011). DICON: Interactive visual analysis of multidimensional clusters. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2581–2590.

Carriere, J., Shafi, H., Brehon, K., Pohar Manhas, K., Churchill, K., Ho, C., and Tavakoli, M. (2021). Case Report: Utilizing AI and NLP to Assist with Healthcare and Rehabilitation During the COVID-19 Pandemic. *Frontiers in Artificial Intelligence*, 4.

Carroll, J. M. (1997). Human–computer interaction: Psychology as a science of design. *International Journal of Human-Computer Studies*, 46(4):501–522.

Cavallo, M. and Demiralp, C. (2019). Clustrophile 2: Guided Visual Clustering Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):267–276.

Chatzimparmpas, A., Martins, R., Jusufi, I., Kucher, K., Rossi, F., and Kerren, A. (2020a). The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum*, 39(3):713–756.

Chatzimparmpas, A., Martins, R. M., Jusufi, I., and Kerren, A. (2020b). A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233.

Chen, L., editor (2008). *User Decision Improvement and Trust Building in Product Recommender Systems*. EPFL, Lausanne.

Chen, L. and Pu, P. (2012). Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1-2):125–150.

Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z., and Zha, H. (2019). Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 765–774, Paris France. ACM.

Cheng, F., Ming, Y., and Qu, H. (2021). DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447.

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., and Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Glasgow Scotland Uk. ACM.

Chishtie, J., Marchand, J.-S., Turcotte, L., Bielska, I., Babineau, J., Cepoiu-Martin, M., Irvine, M., Munce, S., Abudiab, S., Bjelica, M., Hossain, S., Imran, M., Jeji, T., and Jaglal, S. (2020). Visual analytic tools and techniques in population health and health services research: Scoping review. *Journal of Medical Internet Research*, 22(12).

Chopra, K. and Wallace, W. (2003). Trust in electronic environments. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003.*, pages 10 pp.–, Big Island, HI, USA. IEEE.

Christian, B. (2021). The Alignment Problem: Machine Learning and Human Values. *Perspectives on Science and Christian Faith*, 73(4):245–247.

Chuang, J., Ramage, D., Manning, C., and Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452, Austin Texas USA. ACM.

Cisternas, I., Velásquez, I., Caro, A., and Rodríguez, A. (2020). Systematic literature review of implementations of precision agriculture. *Computers and Electronics in Agriculture*, 176:105626.

Clark, N., Hafner, M., Kouril, M., Williams, E., Muhlich, J., Pilarczyk, M., Niepel, M., Sorger, P., and Medvedovic, M. (2017). GRcalculator: An online tool for calculating and mining dose-response data. *BMC cancer*, 17(1):698.

Clark, R. C. and Mayer, R. E. (2011). E-learning and the Science of Instruction.

Cohen, G. L., Steele, C. M., and Ross, L. D. (1999). The Mentor's Dilemma: Providing Critical Feedback Across the Racial Divide. *Personality and Social Psychology Bulletin*, 25(10):1302–1318.

Commission, E. (2023). Proposal for a regulation of the European Parliament and of the Council. Technical report.

Conati, C., Porayska-Pomsta, K., and Mavrikis, M. (2018). AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling.

Conijn, R., Kahr, P., and Snijders, C. (2023). The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. *Journal of Learning Analytics*, 10(1):37–53.

Cortez, P. and Embrechts, M. J. (2011). Opening black box Data Mining models using Sensitivity Analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 341–348.

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., and Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455–496.

Creswell, J. W. and Creswell, J. D. (2017). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications.

Cui, W. (2019). Visual Analytics: A Comprehensive Overview. *IEEE Access*, 7:81555–81573.

Daher, J. B., Brun, A., and Boyer, A. (2017). A Review on Explanations in Recommender Systems. Technical Report, LORIA - Université de Lorraine.

Dahl, O. H. and Fykse, O. (2018). Combining Elo Rating and Collaborative Filtering to improve Learner Ability Estimation in an e-learning Context. Master's thesis, NTNU.

Dandl, S., Molnar, C., Binder, M., and Bischl, B. (2020). Multi-Objective Counterfactual Explanations. In Bäck, T., Preuss, M., Deutz, A., Wang, H., Doerr, C., Emmerich, M., and Trautmann, H., editors, *Parallel Problem Solving from Nature – PPSN XVI*, volume 12269, pages 448–469. Springer International Publishing, Cham.

Dang, T., Murray, P., and Forbes, A. (2015). PathwayMatrix: Visualizing binary relationships between proteins in biological pathways. *BMC Proceedings*, 9.

Dasgupta, A., Lee, J.-Y., Wilson, R., Lafrance, R. A., Cramer, N., Cook, K., and Payne, S. (2017). Familiarity Vs Trust: A Comparative Study of Domain Scientists' Trust in Visual Analytics and Conventional Analysis Methods. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):271–280.

Davis, B., Glenski, M., Sealy, W., and Arendt, D. (2020). Measure Utility, Gain Trust: Practical Advice for XAI Researchers. In *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pages 1–8, Salt Lake City, UT, USA. IEEE.

Deci, E. L. and Ryan, R. M. (2012a). Motivation, personality, and development within embedded social contexts: An overview of self-determination theory. *The Oxford handbook of human motivation*, 18(6):85–107.

Deci, E. L. and Ryan, R. M. (2012b). Self-determination theory. *Handbook of theories of social psychology*, 1(20):416–436.

Demmans Epp, C. and Bull, S. (2015). Uncertainty Representation in Visualizations of Learning Analytics for Learners: Current Approaches and Opportunities. *IEEE Transactions on Learning Technologies*, 8(3):242–260.

Denden, M., Tlili, A., Essalmi, F., Jemni, M., Chang, M., Kinshuk, and Huang, R. (2019). iMoodle: An Intelligent Gamified Moodle to Predict "at-risk" Students Using Learning Analytics Approaches. In Tlili, A. and Chang, M., editors, *Data Analytics Approaches in Educational Games and Gamification Systems*, pages 113–126. Springer Singapore, Singapore.

Deng, H. (2019). Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, 7(4):277–287.

Dereu, L. (2022). De juiste wiskunde aanbevelen op digitale leeromgevingen: het effect van controle over aanbevolen oefeningen. Master's thesis, KU Leuven, Faculty of Engineering Science.

Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., and Li, Y. (2021). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021*, pages 1591–1602. Association for Computing Machinery, New York, NY, USA.

Di Silvestro, L., Burch, M., Caccamo, M., Weiskopf, D., Beck, F., and Gallo, G. (2014). Visual analysis of time-dependent multivariate data from dairy farming industry. In *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 99–106.

Dikmen, M. and Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162:102792.

Dingen, D., Van't Veer, M., Houthuizen, P., Mestrom, E., Korsten, E., Bouwman, A., and Van Wijk, J. (2019). RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):246–255.

Dixit, P., Garcia Caballero, H., Corvò, A., Hompes, B., Buijs, J., and van der Aalst, W. (2017). Enabling interactive process analysis with process mining and visual analytics. In *HEALTHINF 2017 - 10th International Conference on Health Informatics, Proceedings; Part of 10th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2017*, volume 5, pages 573–584.

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Penguin Books Limited.

Donkers, T., Kleemann, T., and Ziegler, J. (2020). Explaining recommendations by means of aspect-based transparent memories. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 166–176, Cagliari Italy. ACM.

Donoso-Guzmán, I., Ooge, J., Parra, D., and Verbert, K. (2023). Towards a Comprehensive Human-Centred Evaluation Framework for Explainable AI. In *Proceedings of the 1st International Conference on eXplainable Artificial Intelligence*, Lisbon, Portugal.

Doroudi, S. (2022). The Intertwined Histories of Artificial Intelligence and Education. *International Journal of Artificial Intelligence in Education*.

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

Drachsler, H., Verbert, K., Santos, O. C., and Manouselis, N. (2015). Panorama of Recommender Systems to Support Learning. In Ricci, F., Rokach, L., and Shapira, B., editors, *Recommender Systems Handbook*, pages 421–451. Springer US, Boston, MA.

Dreyfus, S. E. (2004). The Five-Stage Model of Adult Skill Acquisition. *Bulletin of Science, Technology & Society*, 24(3):177–181.

Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77.

Dunn, T. J., Baguley, T., and Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3):399–412.

Dunning, D. (2011). Chapter five - The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance. In Olson, J. M. and Zanna, M. P., editors, *Advances in Experimental Social Psychology*, volume 44, pages 247–296. Academic Press, San Diego, CA, USA.

Ehsan, U. and Riedl, M. O. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In Stephanidis, C., Kurosu, M., Degen, H., and Reinerman-Jones, L., editors, *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, volume 12424, pages 449–466. Springer International Publishing, Cham.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. O. (2019). Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274, Marina del Ray California. ACM.

Eiband, M., Buschek, D., Kremer, A., and Hussmann, H. (2019). The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages 1–6, New York, NY, USA. Association for Computing Machinery.

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., and Hussmann, H. (2018). Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*, pages 211–223, Tokyo Japan. ACM.

Ekstrand, M. D., Kluver, D., Harper, F. M., and Konstan, J. A. (2015). Letting Users Choose Recommender Algorithms: An Experimental Study. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 11–18, Vienna Austria. ACM.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Pub, New York.

Endert, A., Ribarsky, W., Turkay, C., Wong, B. W., Nabney, I., Blanco, I. D., and Rossi, F. (2017). The State of the Art in Integrating Machine Learning into Visual Analytics: Integrating Machine Learning into Visual Analytics. *Computer Graphics Forum*, 36(8):458–486.

Everitt, B. and Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer New York, New York, NY.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, Salt Lake City, UT, USA. IEEE.

Fang, D., Kahng, M., Hohman, F., Sharmin, M., Polack, P., Al'Absi, M., Sarker, H., and Chau, D. (2017). Mhealth visual Discovery Dashboard. In *UbiComp/ISWC 2017 - Adjunct Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 237–240.

Farag, Y., Berven, F., Jonassen, I., Petersen, K., and Barsnes, H. (2015). Distributed and interactive visual analysis of omics data. *Journal of Proteomics*, 129:78–82.

Farzan, R. and Brusilovsky, P. (2011). Encouraging user participation in a course recommender system: An impact on user behavior. *Computers in Human Behavior*, 27(1):276–284.

Feldman, R. C., Aldana, E., and Stein, K. (2019). Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know. *Stanford Law & Policy Review*, 30(2):399–420.

Feller, D., Burgermaster, M., Levine, M., Smaldone, A., Davidson, P., Albers, D., and Mamykina, L. (2018). A visual analytics approach for pattern-recognition in patient-generated data. *Journal of the American Medical Informatics Association*, 25(10):1366–1374.

Filgona, J., Sakiyo, J., Gwany, D. M., and Okoronka, A. U. (2020). Motivation in Learning. *Asian Journal of Education and Social Studies*, pages 16–37.

Fitton, D., Read, J. C. C., and Horton, M. (2013). The challenge of working with teens as participants in interaction design. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, page 205, Paris, France. ACM Press.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2):147–168.

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., and Hullman, J. (2021). The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, 22(3):110–161.

Freitas, A. A. (2014). Comprehensible classification models: A position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10.

Fu, M. C. (2016). AlphaGo and Monte Carlo tree search: The simulation optimization perspective. In *2016 Winter Simulation Conference (WSC)*, pages 659–670.

Galici, R., Käser, T., Fenu, G., and Marras, M. (2023). How Close are Predictive Models to Teachers in Detecting Learners at Risk? In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 135–145, Limassol Cyprus. ACM.

Gallego, D., Barra, E., Gordillo, A., and Huecas, G. (2013). Enhanced recommendations for e-Learning authoring tools based on a proactive context-aware recommender. In *2013 IEEE Frontiers in Education Conference (FIE)*, pages 1393–1395, Oklahoma City, OK, USA. IEEE.

García, E., Romero, C., Ventura, S., and de Castro, C. (2009). An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. *User Modeling and User-Adapted Interaction*, 19(1-2):99–132.

Garcia-Martinez, S. and Hamou-Lhadj, A. (2013). Educational Recommender Systems: A Pedagogical-Focused Perspective. In Tsihrintzis, G. A., Virvou, M., and Jain, L. C., editors, *Multimedia Services in Intelligent Environments*, volume 25, pages 113–124. Springer International Publishing, Heidelberg.

Gedikli, F., Jannach, D., and Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4):367–382.

Geurts, A., Sakas, G., Kuijper, A., Becker, M., and von Landesberger, T. (2015). Visual comparison of 3D medical image segmentation algorithms based on statistical shape models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9185:336–344.

Gil, Y., Garijo, D., Khider, D., Knoblock, C. A., Ratnakar, V., Osorio, M., Vargas, H., Pham, M., Pujara, J., Shbita, B., Vu, B., Chiang, Y.-Y., Feldman, D., Lin, Y., Song, H., Kumar, V., Khandelwal, A., Steinbach, M., Tayal, K., Xu, S., Pierce, S. A., Pearson, L., Hardesty-Lewis, D., Deelman, E., Silva, R. F. D., Mayani, R., Kemanian, A. R., Shi, Y., Leonard, L., Peckham, S.,

Stoica, M., Cobourn, K., Zhang, Z., Duffy, C., and Shu, L. (2021). Artificial Intelligence for Modeling Complex Systems: Taming the Complexity of Expert Models to Improve Decision Making. *ACM Transactions on Interactive Intelligent Systems*, 11(2):1–49.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). 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, Turin, Italy. IEEE.

Glennerster, R. and Takavarasha, K. (2013). *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, Princeton, New Jersey.

Göker, M. H. and Thompson, C. A. (2000). The Adaptive Place Advisor: A Conversational Recommendation System.

Gomez, O., Holter, S., Yuan, J., and Bertini, E. (2020). ViCE: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 531–535, Cagliari Italy. ACM.

Goodman, B. and Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, 38(3):50–57.

Gotz, D., Sun, J., Cao, N., and Ebadollahi, S. (2011). Visual cluster analysis in support of clinical decision intelligence. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:481–490.

Gotz, D., Wang, F., and Perer, A. (2014). A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics*, 48:148–159.

Gotz, D., Zhang, J., Wang, W., Shrestha, J., and Borland, D. (2020). Visual analysis of high-dimensional event sequence data via dynamic hierarchical aggregation. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):440–450.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual Visual Explanations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2376–2384. PMLR.

Grandison, T. and Sloman, M. (2000). A survey of trust in internet applications. *IEEE Communications Surveys Tutorials*, 3(4):2–16.

Grant, M. J. and Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies: A typology of reviews, *Maria J. Grant & Andrew Booth. Health Information & Libraries Journal*, 26(2):91–108.

Green, B. and Chen, Y. (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. (2019a). Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems*, 34(6):14–23.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. (2018). Local Rule-Based Explanations of Black Box Decision Systems.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019b). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):1–42.

Gulati, S., Sousa, S., and Lamas, D. (2017). Modelling Trust: An Empirical Assessment. In Bernhaupt, R., Dalvi, G., Joshi, A., K. Balkrishan, D., O'Neill, J., and Winckler, M., editors, *Human-Computer Interaction – INTERACT 2017*, volume 10516, pages 40–61. Springer International Publishing, Cham.

Gulati, S., Sousa, S., and Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10):1004–1015.

Gunning, D. and Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58.

Guo, R., Fujiwara, T., Li, Y., Lima, K., Sen, S., Tran, N., and Ma, K.-L. (2020). Comparative visual analytics for assessing medical records with sequence embedding. *Visual Informatics*, 4(2):72–85.

Guo, S., Xu, K., Zhao, R., Gotz, D., Zha, H., and Cao, N. (2018). EventThread: Visual Summarization and Stage Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):56–65.

Guo, X., Huang, Z., Gao, J., Shang, M., Shu, M., and Sun, J. (2021). Enhancing Knowledge Tracing via Adversarial Training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 367–375, Virtual Event China. ACM.

Gutiérrez, F., Htun, N. N., Schlenz, F., Kasimati, A., and Verbert, K. (2019a). A review of visualisations in agricultural decision support systems: An HCI perspective. *Computers and Electronics in Agriculture*, 163:104844.

Gutiérrez, F., Ochoa, X., Seipp, K., Broos, T., and Verbert, K. (2019b). Benefits and Trade-Offs of Different Model Representations in Decision Support Systems for Non-expert Users. In Lamas, D., Loizides, F., Nacke, L., Petrie, H., Winckler, M., and Zaphiris, P., editors, *Human-Computer Interaction – INTERACT 2019*, Lecture Notes in Computer Science, pages 576–597, Cham. Springer International Publishing.

Ham, D.-H. (2010). The State of the Art of Visual Analytics. In Lee, J. H., Lee, H., and Kim, J.-S., editors, *EKC 2009 Proceedings of the EU-Korea Conference on Science and Technology*, Springer Proceedings in Physics, pages 213–222, Berlin, Heidelberg. Springer.

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., and De Hert, P. (2022). Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. *IEEE Computational Intelligence Magazine*, 17(1):72–85.

Han, W. and Schulz, H.-J. (2020). Beyond Trust Building — Calibrating Trust in Visual Analytics. In *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pages 9–15, Salt Lake City, UT, USA. IEEE.

Harambam, J., Bountouridis, D., Makhortykh, M., and van Hoboken, J. (2019). Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 69–77, Copenhagen Denmark. ACM.

He, C., Parra, D., and Verbert, K. (2016). Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56:9–27.

Heiberger, R. M. and Robbins, N. B. (2014). Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications. *Journal of Statistical Software*, 57(5):1–32.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating Visual Explanations. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, volume 9908, pages 3–19. Springer International Publishing, Cham.

Hennink, M. M. (2014). *Focus Group Discussions*. Understanding Qualitative Research. Oxford University Press, Oxford.

Herlocker, J. L., Konstan, J. A., and Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 241–250, New York, NY, USA. Association for Computing Machinery.

Herm, L.-V., Heinrich, K., Wanner, J., and Janiesch, C. (2022). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, page 102538.

Herold, J., Zhou, L., Abouna, S., Pelengaris, S., Epstein, D., Khan, M., and Nattkemper, T. (2010). Integrating semantic annotation and information visualization for the analysis of multichannel fluorescence micrographs from pancreatic tissue. *Computerized Medical Imaging and Graphics*, 34(6):446–452.

Hijikata, Y., Kai, Y., and Nishida, S. (2012). The relation between user intervention and user satisfaction for information recommendation. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing - SAC '12*, page 2002, Trento, Italy. ACM Press.

Hind, M. (2019). Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):16–19.

Hinterberg, M., Kao, D., Bristow, M., Hunter, L., Port, J., and Görg, C. (2015). PEAX: Interactive visual analysis and exploration of complex clinical phenotype and gene expression association. *Pacific Symposium on Biocomputing*, pages 419–430.

Hoff, K. A. and Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(3):407–434.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2019). Metrics for Explainable AI: Challenges and Prospects. *arXiv:1812.04608 [cs]*.

Hohman, F., Head, A., Caruana, R., DeLine, R., and Drucker, S. M. (2019a). Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow Scotland Uk. ACM.

Hohman, F., Kahng, M., Pienta, R., and Chau, D. H. (2019b). Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2674–2693.

Holliday, D., Wilson, S., and Stumpf, S. (2016). User Trust in Intelligent Systems: A Journey Over Time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 164–168, Sonoma California USA. ACM.

Holstein, K., McLaren, B. M., and Aleven, V. (2019). Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher–AI Complementarity. *Journal of Learning Analytics*, 6(2).

Holte, R. C. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning*, 11(1):63–90.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1312.

Hooshyar, D., Pedaste, M., Saks, K., Leijen, Ä., Bardone, E., and Wang, M. (2020). Open learner models in supporting self-regulated learning in higher education: A systematic literature review. *Computers & Education*, 154:103878.

Hu, J., Perer, A., and Wang, F. (2016). Data Driven Analytics for Personalized Healthcare. In Weaver, C. A., Ball, M. J., Kim, G. R., and Kiel, J. M., editors, *Healthcare Information Management Systems*, pages 529–554. Springer International Publishing, Cham.

Huang, C.-W., Lu, R., Iqbal, U., Lin, S.-H., Nguyen, P., Yang, H.-C., Wang, C.-F., Li, J., Ma, K.-L., Li, Y.-C., and Jian, W.-S. (2015). A richly interactive exploratory data analysis and visualization tool using electronic medical records Clinical decision-making, knowledge support systems, and theory. *BMC Medical Informatics and Decision Making*, 15(1).

Huang, M., Yue, Z., Liang, J., Nguyen, Q., and Luo, Z. (2019). Stroke data analysis through a HVN visual mining platform. In *Proceedings - 2019 23rd International Conference in Information Visualization - Part II, IV-2 2019*, pages 1–6.

Hullman, J. (2020). Why Authors Don't Visualize Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):130–139.

Hund, M., Böhm, D., Sturm, W., Sedlmair, M., Schreck, T., Ullrich, T., Keim, D., Majnaric, L., and Holzinger, A. (2016). Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the Doctor-in-the-loop. *Brain Informatics*, 3(4):233–247.

Hur, C., Wi, J., and Kim, Y. (2020). Facilitating the development of deep learning models with visual analytics for electronic health records. *International Journal of Environmental Research and Public Health*, 17(22):1–14.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *arXiv:2010.07487 [cs]*.

Jameson, A. and Schwarzkopf, E. (2002). Pros and Cons of Controllability: An Empirical Study. In Goos, G., Hartmanis, J., van Leeuwen, J., De Bra, P., Brusilovsky, P., and Conejo, R., editors, *Adaptive Hypermedia and Adaptive Web-Based Systems*, volume 2347, pages 193–202. Springer Berlin Heidelberg, Berlin, Heidelberg.

Jannach, D., Naveed, S., and Jugovac, M. (2017). User Control in Recommender Systems: Overview and Interaction Challenges. In Bridge, D. and Stuckenschmidt, H., editors, *E-Commerce and Web Technologies*, volume 278, pages 21–33. Springer International Publishing, Cham.

Jansen, B. R. J., Hofman, A. D., Savi, A., Visser, I., and van der Maas, H. L. J. (2016). Self-adapting the success rate when practicing math. *Learning and Individual Differences*, 51:1–10.

Jarvis, D. H., Wachowiak, M. P., Walters, D. F., and Kovacs, J. M. (2017). Adoption of Web-Based Spatial Tools by Agricultural Producers: Conversations with Seven Northeastern Ontario Farmers Using the GeoVisage Decision Support System. *Agriculture*, 7(8):69.

Jeong, D. H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., and Chang, R. (2009). iPCA: An Interactive System for PCA-based Visual Analytics. *Computer Graphics Forum*, 28(3):767–774.

Ji, S.-Y., Jeong, D., Hassan, M., and Ilev, I. (2019a). Signature Infrared Bacteria Spectra Analyzed by an Advanced Integrative Computational Approach Developed for Identifying Bacteria Similarity. *IEEE Journal of Selected Topics in Quantum Electronics*, 25(1).

Ji, S.-Y., Najarian, K., Huynh, T., and Jeong, D. (2017). An integration of decision tree and visual analysis to analyze intracranial pressure. *Methods in Molecular Biology*, 1598:405–419.

Ji, X., Shen, H.-W., Ritter, A., MacHiraju, R., and Yen, P.-Y. (2019b). Visual Exploration of Neural Document Embedding in Information Retrieval: Semantics and Feature Selection. *IEEE Transactions on Visualization and Computer Graphics*, 25(6):2181–2192.

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71.

Jin, Y., Tintarev, N., Htun, N. N., and Verbert, K. (2020). Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction*, 30(2):199–249.

Jin, Y., Tintarev, N., and Verbert, K. (2018). Effects of personal characteristics on music recommender systems with different levels of controllability. In *RecSys 2018 - 12th ACM Conference on Recommender Systems*, pages 13–21, Vancouver, British Columbia, Canada. Association for Computing Machinery.

Johansson, U. and Niklasson, L. (2009). Evolving decision trees using oracle guides. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pages 238–244.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousnes*. Harvard University Press, Cambridge, Massachusetts.

Jönsson, D., Bergström, A., Algström, I., Simon, R., Engström, M., Walter, S., and Hotz, I. (2019). Visual Analysis for Understanding Irritable Bowel Syndrome. *Advances in Experimental Medicine and Biology*, 1156:111–122.

Jönsson, D., Bergström, A., Forsell, C., Simon, R., Engström, M., Walter, S., Ynnerman, A., and Hotz, I. (2020). VisualNeuro: A Hypothesis Formation and Reasoning Application for Multi-Variate Brain Cohort Study Data. *Computer Graphics Forum*, 39(6):392–407.

Kadengye, D. T., Ceulemans, E., and Van Den Noortgate, W. (2015). Modeling Growth in Electronic Learning Environments Using a Longitudinal Random Item Response Model. *The Journal of Experimental Education*, 83(2):175–202.

Kaffes, V., Sacharidis, D., and Giannopoulos, G. (2021). Model-Agnostic Counterfactual Explanations of Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 280–285, Utrecht Netherlands. ACM.

Kahneman, D. (2011). *Thinking, Fast and Slow*. macmillan.

Kahng, M., Thorat, N., Chau, D. H. P., Viegas, F. B., and Wattenberg, M. (2019). GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):310–320.

Kakar, T., Qin, X., Rundensteiner, E., Harrison, L., Sahoo, S., and De, S. (2019). Diva: Exploration and validation of hypothesized drug-drug interactions. *Computer Graphics Forum*, 38(3):95–106.

Kamienski, C., Soininen, J.-P., Taumberger, M., Fernandes, S., Toscano, A., Cinotti, T. S., Maia, R. F., and Neto, A. T. (2018). SWAMP: An IoT-based Smart Water Management Platform for Precision Irrigation in Agriculture. In *2018 Global Internet of Things Summit (GIoTS)*, pages 1–6.

Kamilaris, A., Kartakoullis, A., and Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, 143:23–37.

Kasirzadeh, A. and Smart, A. (2021). The Use and Misuse of Counterfactuals in Ethical Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236, Virtual Event Canada. ACM.

Kato, S. (2021). Practicing the Right Math: Enhancing Trust in an E-Learning Platform Using an Explainable Recommender System. Master's thesis, KU Leuven, Faculteit Ingenieurswetenschappen.

Kay, J. (2001). Learner control. *User modeling and user-adapted interaction*, 11:111–127.

Kay, J. and Kummerfeld, B. (2019). From data to personal user models for life-long, life-wide learners. *British Journal of Educational Technology*, 50(6):2871–2884.

Keane, M. T. and Smyth, B. (2020). Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). In Watson, I. and Weber, R., editors, *Case-Based Reasoning Research and Development*, volume 12311, pages 163–178. Springer International Publishing, Cham.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008). Visual Analytics: Scope and Challenges. In Simoff, S. J., Böhlen, M. H., and Mazeika, A., editors, *Visual Data Mining*, volume 4404, pages 76–90. Springer Berlin Heidelberg, Berlin, Heidelberg.

Keim, D. A., Mansmann, F., and Thomas, J. (2010). Visual analytics: How much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter*, 11(2):5–8.

Khakpour, A., Colomo-Palacios, R., and Martini, A. (2021). Visual Analytics for Decision Support: A Supply Chain Perspective. *IEEE Access*, 9:81326–81344.

Khanal, S. S., Prasad, P., Alsadoon, A., and Maag, A. (2020). A systematic review: Machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 25(4):2635–2664.

Khine, M. S., editor (2013). *Application of Structural Equation Modeling in Educational Research and Practice*. SensePublishers, Rotterdam.

Khosravi, H., Denny, P., Moore, S., and Stamper, J. (2023). Learnersourcing in the age of AI: Student, educator and machine partnerships for content creation. *Computers and Education: Artificial Intelligence*, 5:100151.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., and Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074.

Kizilcec, R. F. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2390–2395, San Jose California USA. ACM.

Klemm, P., Frauenstein, L., Perlich, D., Hegenscheid, K., V Ö Lzke, H., and Preim, B. (2014). Clustering socio-demographic and medical attribute data in cohort studies. In *Informatik Aktuell*, pages 181–185.

Klimov, D., Shknevsky, A., and Shahar, Y. (2015). Exploration of patterns predicting renal damage in patients with diabetes type II using a visual temporal analysis laboratory. *Journal of the American Medical Informatics Association*, 22(2):275–289.

Klinkenberg, S., Straatemeier, M., and van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824.

Knight, W. (May/June 2017). The Dark Secret at the Heart of AI. *MIT Technology Review*, 120(3).

Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., and Kobsa, A. (2012a). Inspectability and control in social recommenders. In *Proceedings of the Sixth*

ACM Conference on Recommender Systems - RecSys '12, page 43, Dublin, Ireland. ACM Press.

Knijnenburg, B. P., Reijmer, N. J., and Willemsen, M. C. (2011). Each to his own: How different users call for different interaction methods in recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems - RecSys '11*, page 141, Chicago, Illinois, USA. ACM Press.

Knijnenburg, B. P. and Willemsen, M. C. (2015). Evaluating Recommender Systems with User Experiments. In Ricci, F., Rokach, L., and Shapira, B., editors, *Recommender Systems Handbook*, pages 309–352. Springer US, Boston, MA.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. (2012b). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504.

Kolyshkina, I. and Simoff, S. (2021). Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach. *Frontiers in Big Data*, 4.

Konstan, J. A. and Riedl, J. (2012). Recommender systems: From algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123.

Kool, W. and Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, 2(12):899–908.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., and Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, 10(1):11981.

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., and Getoor, L. (2019). Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 379–390, Marina del Ray California. ACM.

Kovalerchuk, B., Delizy, F., Riggs, L., and Vityaev, E. (2012). Visual Data Mining and Discovery with Binarized Vectors. *Intelligent Systems Reference Library*, 24:135–156.

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4):212–218.

Krause, J., Dasgupta, A., Swartz, J., Aphinyanaphongs, Y., and Bertini, E. (2018a). A Workflow for Visual Diagnostics of Binary Classifiers using

Instance-Level Explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology, VAST 2017 - Proceedings*, pages 162–172.

Krause, J., Perer, A., and Bertini, E. (2014). INFUSE: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1614–1623.

Krause, J., Perer, A., and Bertini, E. (2018b). A User Study on the Effect of Aggregating Explanations for Interpreting Machine Learning Models. page 9.

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

Krishnan, S. and Wu, E. (2017). PALM: Machine learning explanations for iterative debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, HILDA 2017*.

Kulesza, T., Stumpf, S., Burnett, M., and Kwan, I. (2012). Tell me more?: The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–10, Austin Texas USA. ACM.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10, San Jose, CA, USA. IEEE.

Kumar, A., Nette, F., Klein, K., Fulham, M., and Kim, J. (2015). A Visual Analytics Approach Using the Exploration of Multidimensional Feature Spaces for Content-Based Medical Image Retrieval. *IEEE Journal of Biomedical and Health Informatics*, 19(5):1734–1746.

Kunkel, J., Donkers, T., Michael, L., Barbu, C.-M., and Ziegler, J. (2019). Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Glasgow Scotland Uk. ACM.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Kwon, B., Eysenbach, B., Verma, J., Ng, K., De Filippi, C., Stewart, W., and Perer, A. (2018). Clustervision: Visual Supervision of Unsupervised

Clustering. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):142–151.

Kwon, B. C., Anand, V., Severson, K. A., Ghosh, S., Sun, Z., Frohnert, B. I., Lundgren, M., and Ng, K. (2021). DPVis: Visual Analytics With Hidden Markov Models for Disease Progression Pathways. *IEEE Transactions on Visualization and Computer Graphics*, 27(9):3685–3700.

Kwon, B. C., Choi, M.-J., Kim, J. T., Choi, E., Kim, Y. B., Kwon, S., Sun, J., and Choo, J. (2019). RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309.

Lakkaraju, H., Slack, D., Irvine, UC., Chen, Y., and Tan, C. (2022). Rethinking Explainability as a Dialogue: A Practitioner's Perspective.

Lamy, J.-B. and Tsopra, R. (2019). Visual Explanation of Simple Neural Networks using Interactive Rainbow Boxes. In *Proceedings of the International Conference on Information Visualisation*, volume 2019-July, pages 50–55.

Langer, E., Blank, A., and Chanowitz, B. (1978). The Mindlessness of Ostensibly Thoughtful Action: The Role of "Placebic" Information in Interpersonal Interaction. *Journal of Personality and Social Psychology*, 36(6):635–642.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296:103473.

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. (2019). The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 2801–2807, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Laux, J., Wachter, S., and Mittelstadt, B. (2022). Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and the Acceptability of Risk.

Lee, J. D. and See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1):50–80.

Lee, S., Amgad, M., Mobadersany, P., McCormick, M., Pollack, B. P., Elfandy, H., Hussein, H., Gutman, D. A., and Cooper, L. A. D. (2021). Interactive Classification of Whole-Slide Imaging Data for Cancer Researchers. *Cancer Research*, 81(4):1171–1177.

Leech, B. L. (2002). Asking Questions: Techniques for Semistructured Interviews. *Political Science & Politics*, 35(04):665–668.

Leffrang, D. and Müller, O. (2021). Should I Follow this Model? The Effect of Uncertainty Visualization on the Acceptance of Time Series Forecasts. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pages 20–26.

Legg, S., Hutter, M., et al. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, 157:17.

Leondari, A. and Gialamas, V. (2002). Implicit theories, goal orientations, and perceived competence: Impact on students' achievement behavior. *Psychology in the Schools*, 39(3):279–291.

Lewis, D. K. (1986). *Philosophical Papers Volume II*. Oxford University Press, New York.

Lezoche, M., Hernandez, J. E., Alemany Díaz, M. d. M. E., Panetto, H., and Kacprzyk, J. (2020). Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture. *Computers in Industry*, 117:103187.

Li, K., Guo, L., Faraco, C., Zhu, D., Chen, H., Yuan, Y., Lv, J., Deng, F., Jiang, X., Zhang, T., Hu, X., Zhang, D., Miller, L., and Liu, T. (2012). Visual analytics of brain networks. *NeuroImage*, 61(1):82–97.

Li, R., Yin, C., Yang, S., Qian, B., and Zhang, P. (2020). Marrying medical domain knowledge with deep learning on electronic health records: A deep visual analytics approach. *Journal of Medical Internet Research*, 22(9).

Li, Y., Hara, S., Ito, W., and Shimura, K. (2007). A machine learning approach for interactive lesion segmentation. In *Medical Imaging 2007: Image Processing*, volume 6512, pages 1393–1400. SPIE.

Liakos, K., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, 18(8):2674.

Liao, Q. V., Gruen, D., and Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Honolulu HI USA. ACM.

Liao, Q. V., Pribić, M., Han, J., Miller, S., and Sow, D. (2021). Question-Driven Design Process for Explainable AI User Experiences. *arXiv:2104.03483 [cs]*.

Liao, Q. V. and Varshney, K. R. (2022). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences.

Liao, Z., Kong, L., Wang, X., Zhao, Y., Zhou, F., Liao, Z., and Fan, X. (2017). A visual analytics approach for detecting and understanding anomalous resident behaviors in smart healthcare. *Applied Sciences (Switzerland)*, 7(3).

Lima, G., Grgić-Hlača, N., Jeong, J. K., and Cha, M. (2022). The Conflict Between Explainable and Accountable Decision-Making Algorithms.

Lin, M.-H., Chen, H.-C., and Liu, K.-S. (2017). A Study of the Effects of Digital Learning on Learning Motivation and Learning Outcome. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(7):3553–3564.

Linaza, M. T., Posada, J., Bund, J., Eisert, P., Quartulli, M., Döllner, J., Pagani, A., G. Olaizola, I., Barriguinha, A., Moysiadis, T., and Lucat, L. (2021). Data-Driven Artificial Intelligence Applications for Sustainable Precision Agriculture. *Agronomy*, 11(6):1227.

Lindblom, J., Lundström, C., Ljung, M., and Jonsson, A. (2017). Promoting sustainable intensification in precision agriculture: Review of decision support systems development and strategies. *Precision Agriculture*, 18(3):309–331.

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

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Liu, S., Wang, X., Liu, M., and Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., and Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297.

Long, Y. and Aleven, V. (2016). Mastery-Oriented Shared Student/System Control Over Problem Selection in a Linear Equation Tutor. In Micarelli, A., Stamper, J., and Panourgia, K., editors, *Intelligent Tutoring Systems*, Lecture Notes in Computer Science, pages 90–100, Cham. Springer International Publishing.

Long, Y. and Aleven, V. (2017). Enhancing learning outcomes through self-regulated learning support with an Open Learner Model. *User Modeling and User-Adapted Interaction*, 27(1):55–88.

Ltifi, H. and Ayed, M. (2016). Visual intelligent decision support systems in the medical field: Design and evaluation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9605 LNCS:243–258.

Lu, Y., Garcia, R., Hansen, B., Gleicher, M., and Maciejewski, R. (2017). The State-of-the-Art in Predictive Visual Analytics. *Computer Graphics Forum*, 36(3):539–562.

Lundberg, H., Mowla, N. I., Abedin, S. F., Thar, K., Mahmood, A., Gidlund, M., and Raza, S. (2022). Experimental Analysis of Trustworthy In-Vehicle Intrusion Detection System Using eXplainable Artificial Intelligence (XAI). *IEEE Access*, 10:102831–102841.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding. *arXiv:1905.04610 [cs, stat]*.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. page 10.

Luo, J., Wu, M., Gopukumar, D., and Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights*, 8:BII.S31559.

Luo, K., Yang, H., Wu, G., and Sanner, S. (2020). Deep Critiquing for VAE-based Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1269–1278. Association for Computing Machinery, New York, NY, USA.

L'Yi, S., Jung, D., Oh, M., Kim, B., Freishtat, R., Giri, M., Hoffman, E., and Seo, J. (2017). miRTarVis+: Web-based interactive visual analytics tool for microRNA target predictions. *Methods*, 124:78–88.

L'Yi, S., Ko, B., Shin, D., Cho, Y.-J., Lee, J., Kim, B., and Seo, J. (2015). XCluSim: A visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. *BMC Bioinformatics*, 16(11).

Mabbott, A. and Bull, S. (2006). Student Preferences for Editing, Persuading, and Negotiating the Open Learner Model. In Hutchison, D., Kanade, T.,

Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Ikeda, M., Ashley, K. D., and Chan, T.-W., editors, *Intelligent Tutoring Systems*, volume 4053, pages 481–490. Springer Berlin Heidelberg, Berlin, Heidelberg.

Machwitz, M., Hass, E., Junk, J., Udelhoven, T., and Schlerf, M. (2019). CropGIS – A web application for the spatial and temporal visualization of past, present and future crop biomass development. *Computers and Electronics in Agriculture*, 161:185–193.

Madsen, M. and Gregor, S. (2000). Measuring Human-Computer Trust. In *Proceedings of the 11th Australasian Conference on Information Systems*, volume 53, pages 6–8, Brisbane, Australia. Australasian Association for Information Systems.

Males, J., Monclús, E., Díaz, J., Navazo, I., and Vázquez, P.-P. (2020). Interactive framework for the visual exploration of colonic data. *Computers and Graphics (Pergamon)*, 91:39–51.

Malik, S., Du, F., Monroe, M., Onukwugha, E., Plaisant, C., and Shneiderman, B. (2015). Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, volume 2015-January, pages 38–49.

Manouselis, N., Drachsler, H., Verbert, K., and Duval, E. (2013). *Recommender Systems for Learning*. SpringerBriefs in Electrical and Computer Engineering. Springer New York, New York, NY.

Manouselis, N., Drachsler, H., Verbert, K., and Santos, O. C., editors (2014). *Recommender Systems for Technology Enhanced Learning*. Springer New York, New York, NY.

Margolis, H. and Mccabe, P. P. (2006). Improving Self-Efficacy and Motivation: What to Do, What to Say. *Intervention in School and Clinic*, 41(4):218–227.

Mayr, E., Hynek, N., Salisu, S., and Windhager, F. (2019). Trust in Information Visualization. *EuroVis Workshop on Trustworthy Visualization (TrustVis)*, page 5 pages.

McCown, R. (2002). Changing systems for supporting farmers' decisions: Problems, paradigms, and prospects. *Agricultural Systems*, 74(1):179–220.

McKnight, D. H., Choudhury, V., and Kacmar, C. (2002). Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research*, 13(3):334–359.

McNee, S. M., Lam, S. K., Konstan, J. A., and Riedl, J. (2003). Interfaces for Eliciting New User Preferences in Recommender Systems. In Brusilovsky, P., Corbett, A., and de Rosis, F., editors, *User Modeling 2003*, volume 2702, pages 178–187. Springer Berlin Heidelberg, Berlin, Heidelberg.

Merritt, S. M., Heimbaugh, H., LaChapell, J., and Lee, D. (2013). I Trust It, but I Don't Know Why: Effects of Implicit Attitudes Toward Automation on Trust in an Automated System. *Human Factors*, 55(3):520–534.

Michlík, P. and Bieliková, M. (2010). Exercises recommending for limited time learning. *Procedia Computer Science*, 1(2):2821–2828.

Millecamp, M., Htun, N. N., Conati, C., and Verbert, K. (2019). To explain or not to explain: The effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 397–407, Marina del Ray California. ACM.

Millecamp, M., Htun, N. N., Jin, Y., and Verbert, K. (2018). Controlling Spotify Recommendations: Effects of Personal Characteristics on Music Recommender User Interfaces. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 101–109, Singapore Singapore. ACM.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 333–342, New York, NY, USA. Association for Computing Machinery.

Ming, Y., Qu, H., and Bertini, E. (2019). RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):342–352.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246.

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin UK.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group*, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of internal medicine*, 151(4):264–269.

Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., Hu, X., Ji, S., and Ragan, E. D. (2020). Trust Evolution Over Time in Explainable AI for Fake News Detection. page 4.

Mohseni, S., Zarei, N., and Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4):24:1–24:45.

Molnar, C. (2021). *Interpretable Machine Learning*.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Moore, J., Hammerla, N., and Watkins, C. (2019). Explaining Deep Learning Models with Constrained Adversarial Examples. In Nayak, A. C. and Sharma, A., editors, *PRICAI 2019: Trends in Artificial Intelligence*, volume 11670, pages 43–56. Springer International Publishing, Cham.

Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., and Perer, A. (2023). The Impact of Imperfect XAI on Human-AI Decision-Making.

Moschonas, P., Kalamaras, E., Papadopoulos, S., Drosou, A., Votis, K., Bostantjopoulou, S., Katsarou, Z., Papaxanthis, C., Hatzitaki, V., and Tzovaras, D. (2016). Discovering the discriminating power in patient test features using visual analytics: A case study in parkinson's disease. *IFIP Advances in Information and Communication Technology*, 475:600–610.

Moysiadis, V., Sarigiannidis, P., Vitsas, V., and Khelifi, A. (2021). Smart Farming in Europe. *Computer Science Review*, 39:100345.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5-6):527–539.

Müller, J., Stoehr, M., Oeser, A., Gaebel, J., Streit, M., Dietz, A., and Oeltze-Jafra, S. (2020). A visual approach to explainable computerized clinical decision support. *Computers and Graphics (Pergamon)*, 91:1–11.

Munzner, T. (2014). *Visualization Analysis and Design*. A K Peters/CRC Press, 0 edition.

Murray, T. and Arroyo, I. (2002). Toward Measuring and Maintaining the Zone of Proximal Development in Adaptive Instructional Systems. In Cerri, S. A., Gouardères, G., and Paraguaçu, F., editors, *Intelligent Tutoring Systems*,

Lecture Notes in Computer Science, pages 749–758, Berlin, Heidelberg. Springer.

Naidoo, J. (2020). Postgraduate Mathematics Education Students' Experiences of Using Digital Platforms for Learning within the COVID-19 Pandemic Era. *Pythagoras*, 41(1).

Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N., and Ali, R. (2021). Explainable Recommendations and Calibrated Trust: Two Systematic User Errors. *Computer*, 54(10):28–37.

Nauta, M., Van Putten, M., Tjepkema-Cloostermans, M., Bos, J., Van Keulen, M., and Seifert, C. (2020). Interactive explanations of internal representations of neural network layers: An exploratory study on outcome prediction of comatose patients. In *CEUR Workshop Proceedings*, volume 2675, pages 5–11.

Navarro, E., Costa, N., and Pereira, A. (2020). A Systematic Review of IoT Solutions for Smart Farming. *Sensors*, 20(15):4231.

Nguyen, Q., Alzamora, P., Ho, N., Huang, M., Simoff, S., and Catchpoole, D. (2012). Unlocking the complexity of genomic data of RMS patients through visual analytics. In *ICCH 2012 Proceedings - International Conference on Computerized Healthcare*, pages 134–139.

Nguyen, Q., Gleeson, A., Ho, N., Huang, M., Simoff, S., and Catchpoole, D. (2011). Visual analytics of clinical and genetic datasets of acute lymphoblastic leukaemia. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7062 LNCS(PART 1):113–120.

Ni, L., Bao, Q., Li, X., Qi, Q., Denny, P., Warren, J., Witbrock, M., and Liu, J. (2022). DeepQR: Neural-Based Quality Ratings for Learnersourced Multiple-Choice Questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12826–12834.

Nourani, M., Kabir, S., Mohseni, S., and Ragan, E. D. (2019). The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:97–105.

Nourani, M., King, J., and Ragan, E. (2020). The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8:112–121.

Nunes, I., link will open in a new window Link to external site, t., and Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User - Adapted Interaction*, 27(3-5):393–444.

O'Brien, H. L. and Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69.

Ochola, W. O. and Kerkides, P. (2004). An integrated indicator-based spatial decision support system for land quality assessment in Kenya. *Computers and Electronics in Agriculture*, 45(1):3–26.

O'Donovan, J. and Smyth, B. (2005). Trust in recommender systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, pages 167–174, San Diego California USA. ACM.

O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., and Höllerer, T. (2008). PeerChooser: Visual interactive recommendation. In *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*, page 1085, Florence, Italy. ACM Press.

Olson, G. M. and Olson, J. S. (2003). Human-Computer Interaction: Psychological Aspects of the Human Use of Computing. *Annual Review of Psychology*, 54(1):491–516.

Ooge, J. (2019). Het personaliseren van motivationele strategieën en gamificationtechnieken m.b.v. recommendersystemen. Master's thesis, KU Leuven, Faculteit Wetenschappen.

Ooge, J., De Croon, R., Verbert, K., and Vanden Abeele, V. (2020). Tailoring Gamification for Adolescents: A Validation Study of Big Five and Hexad in Dutch. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 206–218, Virtual Event Canada. ACM.

Ooge, J., Dereu, L., and Verbert, K. (2023). Steering Recommendations and Visualising Its Impact: Effects on Adolescents' Trust in E-Learning Platforms. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 156–170, New York, NY, USA. Association for Computing Machinery.

Ooge, J., Kato, S., and Verbert, K. (2022a). Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th International Conference on Intelligent User Interfaces*, IUI '22, pages 93–105, New York, NY, USA. Association for Computing Machinery.

Ooge, J., Stiglic, G., and Verbert, K. (2022b). Explaining artificial intelligence with visual analytics in healthcare. *WIREs Data Mining and Knowledge Discovery*, 12(1):e1427.

Ooge, J. and Verbert, K. (2021). 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, New Orleans, LA, USA. IEEE.

Ooge, J. and Verbert, K. (2022). Visually Explaining Uncertain Price Predictions in Agrifood: A User-Centred Case-Study. *Agriculture*, 12(7):1024.

Osinga, S. A., Paudel, D., Mouzakitis, S. A., and Athanasiadis, I. N. (2022). Big data in agriculture: Between opportunity and solution. *Agricultural Systems*, 195:103298.

Padilla, L. M. K., Powell, M., Kay, M., and Hullman, J. (2021). Uncertain About Uncertainty: How Qualitative Expressions of Forecaster Confidence Impact Decision-Making With Uncertainty Visualizations. *Frontiers in Psychology*, 11.

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., and Gomez, E. (2023). The role of explainable AI in the context of the AI Act. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1139–1150, Chicago IL USA. ACM.

Panniello, U., Gorgoglione, M., and Tuzhilin, A. (2016). In CARSs We Trust: How Context-Aware Recommendations Affect Customers' Trust and Other Business Performance Measures of Recommender Systems. *Information Systems Research*, 27(1):182–196.

Papenmeier, A., Englebienne, G., and Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *arXiv:1907.12652 [cs]*.

Papenmeier, A., Kern, D., Englebienne, G., and Seifert, C. (2022). It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Transactions on Computer-Human Interaction*, 29(4):35:1–35:33.

Papoušek, J. and Pelánek, R. (2017). Should We Give Learners Control Over Item Difficulty? In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 299–303, Bratislava Slovakia. ACM.

Parker, C. (1999). A user-centred design method for agricultural DSS. In *EFITA-99: Proceedings of the Second European Conference for Information Technology in Agriculture. Bonn, Germany*, pages 27–30, Bonn, Germany.

Parker, C. and Sinclair, M. (2001). User-centred design does make a difference. The case of decision support systems in crop production. *Behaviour & Information Technology*, 20(6):449–460.

Parker, CG. and Campion, S. (1997). Improving the uptake of decision support systems in agriculture. In *First European Conference for Information Technology in Agriculture*, pages 129–134. Citeseer.

Pawelczyk, M., Broelemann, K., and Kasneci, G. (2020). Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *Proceedings of The Web Conference 2020*, pages 3126–3132. Association for Computing Machinery, New York, NY, USA.

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education*, 98:169–179.

Pelletier, L. G., Rocchi, M. A., Vallerand, R. J., Deci, E. L., and Ryan, R. M. (2013). Validation of the revised sport motivation scale (SMS-II). *Psychology of Sport and Exercise*, 14(3):329–341.

Petrescu, D. A., Antognini, D., and Faltings, B. (2021). Multi-Step Critiquing User Interface for Recommender Systems. In *Fifteenth ACM Conference on Recommender Systems*, RecSys '21, pages 760–763, New York, NY, USA. Association for Computing Machinery.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. (2021). Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. (2020). FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, New York NY USA. ACM.

Preim, B. and Lawonn, K. (2020). A Survey of Visual Analytics for Public Health. *Computer Graphics Forum*, 39(1):543–580.

Pu, P. and Chen, L. (2006). Trust building with explanation interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces - IUI '06*, page 93, Sydney, Australia. ACM Press.

Pu, P. and Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20(6):542–556.

Pu, P. and Chen, L. (2010). A User-Centric Evaluation Framework of Recommender Systems. In *Proceedings of the ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI)*, volume 612, page 8, Barcelona, Spain. CEUR-WS.org.

Pu, P., Chen, L., and Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 157–164, New York, NY, USA. Association for Computing Machinery.

Qu, Z., Lau, C., Nguyen, Q., Zhou, Y., and Catchpoole, D. (2019). Visual Analytics of Genomic and Cancer Data: A Systematic Review. *Cancer Informatics*, 18.

Radford, A., Jozefowicz, R., and Sutskever, I. (2017). Learning to Generate Reviews and Discovering Sentiment.

Rahdari, B., Brusilovsky, P., and Babichenko, D. (2020). Personalizing Information Exploration with an Open User Model. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 167–176, Virtual Event USA. ACM.

Raidou, R., Kuijf, H., Sepasian, N., Pezzotti, N., Bouvy, W., Breeuwer, M., and Vilanova, A. (2016a). Employing visual analytics to aid the design of white matter hyperintensity classifiers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9901 LNCS:97–105.

Raidou, R., Marcelis, F., Breeuwer, M., Gröller, M., Vilanova, A., and van de Wetering, H. (2016b). Visual analytics for the exploration and assessment of segmentation errors. In *VCBM 2016 - Eurographics Workshop on Visual Computing for Biology and Medicine*, pages 193–202.

Raidou, R., van der Heide, U., Dinh, C., Ghobadi, G., Kallehauge, J., Breeuwer, M., and Vilanova, A. (2015). Visual Analytics for the Exploration of Tumor Tissue Characterization. *Computer Graphics Forum*, 34(3):11–20.

Rethlefsen, M. L., Murad, M. H., and Livingston, E. H. (2014). Engaging Medical Librarians to Improve the Quality of Review Articles. *JAMA*, 312(10):999.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "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. Association for Computing Machinery.

Ribera, M. and Lapedriza, A. (2019). Can we do better explanations? A proposal of User-Centered Explainable AI. *Los Angeles*, page 8.

Riegler, M., Pogorelov, K., Lux, M., Halvorsen, P., Griwodz, C., De Lange, T., and Eskeland, S. (2016). Explorative hyperbolic-tree-based clustering tool for unsupervised knowledge discovery. In *Proceedings - International Workshop on Content-Based Multimedia Indexing*, volume 2016-June.

Rind, A. (2013). Interactive Information Visualization to Explore and Query Electronic Health Records. *Foundations and Trends® in Human–Computer Interaction*, 5(3):207–298.

Riveiro, M. and Thill, S. (2021). "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence*, 298:103507.

Rojo, D., Htun, N. N., Parra, D., De Croon, R., and Verbert, K. (2021). AHMoSe: A knowledge-based visual support system for selecting regression machine learning models. *Computers and Electronics in Agriculture*, 187:106183.

Rose, D. C., Parker, C., Fodey, J., Park, C., Sutherland, W. J., and Dicks, L. V. (2017). Involving stakeholders in agricultural decision support systems: Improving user-centred design. *International Journal of Agricultural Management*, 6(3):10.

Rose, D. C., Sutherland, W. J., Parker, C., Lobley, M., Winter, M., Morris, C., Twining, S., Ffoulkes, C., Amano, T., and Dicks, L. V. (2016). Decision support tools for agriculture: Towards effective design and delivery. *Agricultural Systems*, 149:165–174.

Rostamzadeh, N., Abdullah, S., and Sedig, K. (2020). Data-driven activities involving electronic health records: An activity and task analysis framework for interactive visualization tools. *Multimodal Technologies and Interaction*, 4(1).

Rozenblit, L. and Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5):521–562.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Russell, S. J. and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson, 4th edition edition.

Ryan, R. M., editor (2012). *The Oxford Handbook of Human Motivation*. Oxford Library of Psychology. Oxford University Press, New York.

Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., and Keim, D. A. (2016). The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249.

Salau, L., Hamada, M., Prasad, R., Hassan, M., Mahendran, A., and Watanobe, Y. (2022). State-of-the-Art Survey on Deep Learning-Based Recommender Systems for E-Learning. *Applied Sciences*, 12(23):11996.

Santamaría, R., Therón, R., Durán, L., García, A., González, S., Sánchez, M., and Antequera, F. (2019). Genome-wide search of nucleosome patterns using visual analytics. *Bioinformatics*, 35(13):2185–2192.

Santamaría, R., Therón, R., and Quintales, L. (2008). A visual analytics approach for understanding biclustering results from microarray data. *BMC Bioinformatics*, 9.

Saraiya, P., North, C., Lam, V., and Duca, K. (2006). An Insight-Based Longitudinal Study of Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1511–1522.

Satariano, A. (2023). Europeans Take a Major Step Toward Regulating A.I. *The New York Times*.

Savikhin, A., Lam, H. C., Fisher, B., and Ebert, D. S. (2011). An Experimental Study of Financial Portfolio Selection with Visual Analytics for Decision Support. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10.

Schaffer, J., Hollerer, T., and O'Donovan, J. (2015). Hypothetical Recommendation: A Study of Interactive Profile Manipulation Behavior for Recommender Systems. In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference*, page 6, Hollywood, Florida. AAAI Press.

Schlicker, N., Uhde, A., Baum, K., Hirsch, M. C., and Langer, M. (2022). Calibrated Trust as a Result of Accurate Trustworthiness Assessment – Introducing the Trustworthiness Assessment Model.

Seipp, K., Gutiérrez, F., Ochoa, X., and Verbert, K. (2019). Towards a visual guide for communicating uncertainty in Visual Analytics. *Journal of Computer Languages*, 50:1–18.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Seo, J. and Shneiderman, B. (2002). Interactively exploring hierarchical clustering results. *Computer*, 35(7):80–86.

Shamma, D. A., Lee, M. L., Filipowicz, A. L. S., Denoue, L., Glazko, K., Murakami, K., and Lyons, K. (2022). EV Life: A Counterfactual Dashboard Towards Reducing Carbon Emissions of Automotive Behaviors. In *27th International Conference on Intelligent User Interfaces*, IUI '22 Companion, pages 46–49, New York, NY, USA. Association for Computing Machinery.

Shang, R., Feng, K. J. K., and Shah, C. (2022). Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1330–1340, Seoul Republic of Korea. ACM.

Sharma, S., Henderson, J., and Ghosh, J. (2020). CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–172. Association for Computing Machinery, New York, NY, USA.

Shneiderman, B. (2003). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In Bederson, B. B. and Shneiderman, BEN., editors, *The Craft of Information Visualization*, Interactive Technologies, pages 364–371. Morgan Kaufmann, San Francisco.

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., and Diakopoulos, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, Hoboken, 6th edition edition.

Shortliffe, E. H. (1977). Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 66–69.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences.

Siegel, S. and Castellan, N. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill International Editions Statistics Series. McGraw-Hill.

Simpao, A. F., Ahumada, L. M., Gálvez, J. A., and Rehman, M. A. (2014). A Review of Analytics and Clinical Informatics in Health Care. *Journal of Medical Systems*, 38(4):45.

Singh, S., Ribeiro, M. T., and Guestrin, C. (2016). Programs as Black-Box Explanations.

Sioli, L. (2021). A European Strategy for Artificial Intelligence.

Skeels, M., Lee, B., Smith, G., and Robertson, G. G. (2010). Revealing Uncertainty for Information Visualization. *Information Visualization*, 9(1):70–81.

Snedecor, G. and Cochran, W. (1969). *Statistical Methods*. Iowa State University Press.

Sokol, K. and Flach, P. (2018). Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2018-July, pages 5868–5870.

Sokol, K. and Flach, P. (2020). One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. *KI - Kunstliche Intelligenz*, 34(2):235–250.

Solhaug, B., Elgesem, D., and Stolen, K. (2007). Why Trust is not Proportional to Risk. In *The Second International Conference on Availability, Reliability and Security (ARES'07)*, pages 11–18.

Song, H., Lee, J., Kim, T., Lee, K., Kim, B., and Seo, J. (2017). GazeDx: Interactive Visual Analytics Framework for Comparative Gaze Analysis with Volumetric Medical Images. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):311–320.

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing Uncertainty About the Future. *Science*, 333(6048):1393–1400.

Spitz, L., Niemann, U., Beuing, O., Neyazi, B., Sandalcioglu, I., Preim, B., and Saalfeld, S. (2020). Combining visual analytics and case-based reasoning for rupture risk assessment of intracranial aneurysms. *International Journal of Computer Assisted Radiology and Surgery*, 15(9):1525–1535.

Spooner, T., Dervovic, D., Long, J., Shepard, J., Chen, J., and Magazzeni, D. (2021). Counterfactual Explanations for Arbitrary Regression Models. *arXiv:2106.15212 [cs]*.

Stiglic, G., Kocbek, P., Cilar, L., Fijačko, N., Stožer, A., Zaletel, J., Sheikh, A., and Povalej Bržan, P. (2018). Development of a screening tool using electronic health records for undiagnosed Type 2 diabetes mellitus and impaired fasting glucose detection in the Slovenian population. *Diabetic Medicine*, 35(5):640–649.

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., and Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5):e1379.

Stolper, C., Perer, A., and Gotz, D. (2014). Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1653–1662.

Storms, E., Alvarado, O., and Monteiro-Krebs, L. (2022). 'Transparency is Meant for Control' and Vice Versa: Learning from Co-designing and Evaluating Algorithmic News Recommenders. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):405:1–405:24.

Strobelt, H., Gehrmann, S., Behrisch, M., Perer, A., Pfister, H., and Rush, A. M. (2019). Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):353–363.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. page 19.

Sturm, W., Schreck, T., Holzinger, A., and Ullrich, T. (2015). Discovering medical knowledge using visual analytics - A survey on methods for systems biology and ?omics data -. In *Eurographics Workshop on Visual Computing for Biology and Medicine, VCBM 2015*, pages 71–81.

Sun, D., Feng, Z., Chen, Y., Wang, Y., Zeng, J., Yuan, M., Pong, T.-C., and Qu, H. (2020). DFSeer: A Visual Analytics Approach to Facilitate Model Selection for Demand Forecasting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA. ACM.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Suresh, H., Gomez, S. R., Nam, K. K., and Satyanarayan, A. (2021). Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–16, New York, NY, USA. Association for Computing Machinery.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks.

Szymanski, M., Millecamp, M., and Verbert, K. (2021). Visual, textual or hybrid: The effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*, pages 109–119, College Station TX USA. ACM.

Tan, S., Soloviev, M., Hooker, G., and Wells, M. T. (2020). Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS '20, pages 23–34, New York, NY, USA. Association for Computing Machinery.

Tang, T. and McCalla, G. (2005). Smart Recommendation for an Evolving E-Learning System: Architecture and Experiment. *International Journal on E-Learning*, 4(1):105–129.

TED-Ed (2015). How does anesthesia work? - Steven Zheng.

Thiagarajan, J. J., Kailkhura, B., Sattigeri, P., and Ramamurthy, K. N. (2016). TreeView: Peeking into Deep Neural Networks Via Feature-Space Partitioning.

Thomas, J. and Kielman, J. (2009). Challenges for Visual Analytics. *Information Visualization*, 8(4):309–314.

Tintarev, N. and Masthoff, J. (2007a). Effective explanations of recommendations: User-centered design. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 153–156, New York, NY, USA. Association for Computing Machinery.

Tintarev, N. and Masthoff, J. (2007b). A Survey of Explanations in Recommender Systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 801–810, Istanbul, Turkey. IEEE.

Tintarev, N. and Masthoff, J. (2011). Designing and Evaluating Explanations for Recommender Systems. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 479–510. Springer US, Boston, MA.

Tintarev, N. and Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):399–439.

Tolomei, G., Silvestri, F., Haines, A., and Lalmas, M. (2017). Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 465–474, New York, NY, USA. Association for Computing Machinery.

Torkamaan, H. and Ziegler, J. (2022). Recommendations as Challenges: Estimating Required Effort and User Ability for Health Behavior Change Recommendations. In *27th International Conference on Intelligent User Interfaces*, pages 106–119, Helsinki Finland. ACM.

Tsai, C.-H. and Brusilovsky, P. (2019a). Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 22–30. Association for Computing Machinery, New York, NY, USA.

Tsai, C.-H. and Brusilovsky, P. (2019b). Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 391–396, Marina del Ray California. ACM.

Tsai, C.-H. and Brusilovsky, P. (2021). The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction*, 31(3):591–627.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231.

Turkay, C., Jeanquartier, F., Holzinger, A., and Hauser, H. (2014). On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8401:117–140.

Uggirala, A., Gramopadhye, A. K., Melloy, B. J., and Toler, J. E. (2004). Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics*, 34(3):175–186.

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, 3(31):1026.

van Berkel, N., Skov, M. B., and Kjeldskov, J. (2021). Human-AI interaction: Intermittent, continuous, and proactive. *Interactions*, 28(6):67–71.

Van Cauwenberge, D., Van Biesen, W., Decruyenaere, J., Leune, T., and Sterckx, S. (2022). "Many roads lead to Rome and the Artificial Intelligence only shows me one road": An interview study on physician attitudes regarding the implementation of computerised clinical decision support systems. *BMC Medical Ethics*, 23(1):1–14.

van den Elzen, S. and van Wijk, J. J. (2011). BaobabView: Interactive construction and analysis of decision trees. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 151–160.

Van Houdt, L., Millecamp, M., Verbert, K., and Vanden Abeele, V. (2020). Disambiguating Preferences for Gamification Strategies to Motivate Pro-Environmental Behaviour. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 241–253, Virtual Event Canada. ACM.

Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 32(24):18069–18083.

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., and Klerkx, J. (2013). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing*.

Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., and Duval, E. (2012). Context-Aware Recommender Systems for Learning: A Survey and Future Challenges. *IEEE Transactions on Learning Technologies*, 5(4):318–335.

Vereschak, O., Bailly, G., and Caramiaux, B. (2021). How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):327:1–327:39.

Verma, J., Luo, H., Hu, J., and Zhang, P. (2017). DrugPathSeeker: Interactive UI for exploring drug-ADR relation via pathways. In *IEEE Pacific Visualization Symposium*, pages 260–264.

Viani, N., Botelle, R., Kerwin, J., Yin, L., Patel, R., Stewart, R., and Velupillai, S. (2021). A natural language processing approach for identifying temporal disease onset information from mental healthcare text. *Scientific Reports*, 11(1):757.

Vidotto, G., Massidda, D., Noventa, S., and Vicentini, M. (2012). Trusting Beliefs: A Functional Measurement Study. *Psicologica: International Journal of Methodology and Experimental Psychology*, 33(3):575–590.

Vidovic, M. M.-C., Görnitz, N., Müller, K.-R., and Kloft, M. (2016). Feature Importance Measure for Non-linear Learning Algorithms.

Vieira, C., Parsons, P., and Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, 122:119–135.

Vig, J., Sen, S., and Riedl, J. (2009). Tagsplanations: Explaining recommendations using tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, pages 47–56, New York, NY, USA. Association for Computing Machinery.

Vilone, G. and Longo, L. (2020). Explainable Artificial Intelligence: A Systematic Review. *arXiv:2006.00093 [cs]*.

Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Voltaire (1977). *The Portable Voltaire*. Penguin.

von Landesberger, T., Andrienko, G., Andrienko, N., Bremm, S., Kirschner, M., Wesarg, S., and Kuijper, A. (2013). Opening up the "black box" of medical image segmentation with statistical shape models. *The Visual Computer*, 29(9):893–905.

Wachowiak, M. P., Walters, D. F., Kovacs, J. M., Wachowiak-Smolíková, R., and James, A. L. (2017). Visual analytics and remote sensing imagery to support community-based research for precision agriculture in emerging areas. *Computers and Electronics in Agriculture*, 143:149–164.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*.

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019a). Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15. Association for Computing Machinery, New York, NY, USA.

Wang, F., Kaushal, R., and Khullar, D. (2020). Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine? *Annals of Internal Medicine*, 172(1):59.

Wang, J., Gou, L., Shen, H.-W., and Yang, H. (2019b). DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):288–298.

Wang, T. D., Wongsuphasawat, K., Plaisant, C., and Shneiderman, B. (2011). Extracting Insights from Electronic Health Records: Case Studies, a Visual Analytics Process Model, and Design Recommendations. *Journal of Medical Systems*, 35(5):1135–1152.

Wang, W. and Benbasat, I. (2005). Trust In and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems*, 6(3):72–101.

Wang, X. and Yin, M. (2021). Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, College Station TX USA. ACM.

Wang, Y., Ding, Q., Wang, K., Liu, Y., Wu, X., Wang, J., Liu, Y., and Miao, C. (2021). The Skyline of Counterfactual Explanations for Machine Learning Decision Models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2030–2039, Virtual Event Queensland Australia. ACM.

Wang, Y. D. (2014). Building Trust in E-Learning. *ATHENS JOURNAL OF EDUCATION*, 1(1):9–18.

Wauters, K., Desmet, P., and Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58(4):1183–1193.

Wauters, K., Desmet, P., and Van Noortgate, W. (2010). Monitoring learners' proficiency: Weight adaptation in the elo rating system. In *Educational Data Mining 2011*.

Weber, T., Hußmann, H., and Eiband, M. (2021). Quantifying the Demand for Explainability. In Ardito, C., Lanzilotti, R., Malizia, A., Petrie, H., Piccinno, A., Desolda, G., and Inkpen, K., editors, *Human-Computer Interaction – INTERACT 2021*, volume 12933, pages 652–661. Springer International Publishing, Cham.

West, V., Borland, D., and Hammond, W. (2015). Innovative information visualization of electronic health record data: A systematic review. *Journal of the American Medical Informatics Association*, 22(2):330–339.

Widanagamaachchi, W., Livnat, Y., Bremer, P.-T., Duvall, S., and Pascucci, V. (2017). Interactive Visualization and Exploration of Patient Progression in a Hospital Setting. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2017:1773–1782.

Wolfert, S., Ge, L., Verdouw, C., and Bogaardt, M.-J. (2017). Big Data in Smart Farming – A review. *Agricultural Systems*, 153:69–80.

Wu, D. T. Y., Chen, A. T., Manning, J. D., Levy-Fix, G., Backonja, U., Borland, D., Caban, J. J., Dowding, D. W., Hochheiser, H., Kagan, V., Kandaswamy,

S., Kumar, M., Nunez, A., Pan, E., and Gotz, D. (2019). Evaluating visual analytics for health informatics applications: A systematic review from the American Medical Informatics Association Visual Analytics Working Group Task Force on Evaluation. *Journal of the American Medical Informatics Association*, 26(4):314–323.

Wu, Z., Li, M., Tang, Y., and Liang, Q. (2020). Exercise recommendation based on knowledge concept prediction. *Knowledge-Based Systems*, 210:106481.

Xiao and Benbasat (2007). E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact. *MIS Quarterly*, 31(1):137.

Xing, E., Curtis, R., Schoenherr, G., Lee, S., Yin, J., Puniyani, K., Wu, W., and Kinnaird, P. (2014). GWAS in a box: Statistical and visual analytics of structured associations via GenAMap. *PLoS ONE*, 9(6).

Yan, A., Hou, R., Liu, X., Yan, H., Huang, T., and Wang, X. (2022). Towards explainable model extraction attacks. *International Journal of Intelligent Systems*, 37(11):9936–9956.

Yang, F., Huang, Z., Scholtz, J., and Arendt, D. L. (2020a). How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 189–201, Cagliari Italy. ACM.

Yang, L., Kenny, E. M., Ng, T. L. J., Yang, Y., Smyth, B., and Dong, R. (2020b). Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification.

Yeager, D. S., Lee, H. Y., and Dahl, R. E. (2017). Competence and motivation during adolescence. In *Handbook of Competence and Motivation: Theory and Application*, volume 100, pages 431–448. Guilford Press, New York.

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., Hessert, W. T., Williams, M. E., and Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*, 143(2):804–824.

Yi, J. S., ah Kang, Y., Stasko, J., and Jacko, J. (2007). Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231.

Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). 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, Glasgow Scotland Uk. ACM.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., and Chen, F. (2017a). User Trust Dynamics: An Investigation Driven by Differences in System Performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 307–317, Limassol Cyprus. ACM.

Yu, K.-H., Beam, A. L., and Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10):719–731.

Yu, L., Jiang, H., Yu, H., Zhang, C., McAllister, J., and Zheng, D. (2017b). IVAR: Interactive visual analytics of radiomics features from large-scale medical images. In *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, volume 2018-January, pages 3916–3923.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., and Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 2021:e8812542.

Zhai, Z., Martínez, J. F., Beltran, V., and Martínez, N. L. (2020). Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170:105256.

Zhang, J., Wang, Y., Molino, P., Li, L., and Ebert, D. S. (2019). Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):364–373.

Zhang, Y. and Chen, X. (2020). Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, Barcelona Spain. ACM.

Zhao, H., Zhang, H., Liu, Y., Zhang, Y., and Zhang, X. (2017). Pattern discovery: A progressive visual analytic design to support categorical data analysis. *Journal of Visual Languages and Computing*, 43:42–49.

Zhao, X., Wu, Y., Lee, D. L., and Cui, W. (2019). iForest: Interpreting Random Forests via Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):407–416.

Zhou, J., Arshad, S. Z., Luo, S., and Chen, F. (2017). Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making. In Bernhaupt, R., Dalvi, G., Joshi, A., K. Balkrishan, D., O'Neill, J., and Winckler, M.,

editors, *Human-Computer Interaction – INTERACT 2017*, volume 10516, pages 23–39. Springer International Publishing, Cham.

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5):593.

Zimmerman, B. J. (1990). Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist*, 25(1):3–17.

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing Deep Neural Network Decisions: Prediction Difference Analysis.