How interpretable is explainable?

The development of a framework to assess how interpretable Explainable Artificial Intelligence is for laypeople

By

David Lensen

A Master Thesis

Submitted to the Faculty of Technology, Policy, and Management

At the Delft University of Technology

In the process to obtain the degree of Master of Science

To be defended publicly on Thursday, June 8th, 2023, at 14:00

Student number: 4960432

Project duration: January 1st, 2023 – June 8th, 2023

Thesis committee:

Dr. A.Y. (Aaron) Ding	Chair & First supervisor
Prof. dr. M.E. (Martijn) Warnier	Second supervisor
Dr. M. (Marcus) Westberg	Advisor

An electronic version of this thesis is available at http://repository.tudelft.nl/

TU Delft

Preface

Before you lies the master thesis *"How interpretable is explainable?"* which has been written to fulfil the graduation requirements of the Complex Systems Engineering and Management program at the Delft University of Technology. I was engaged in arranging, researching, and writing this thesis from January 2023 to June 2023.

First, I would like to thank all members of my thesis committee, not only regarding the contents of this report, but moreover their thoughtfulness in considering my personal process throughout the entire project. Starting with dr. Aaron Ding, who, as chair and first supervisor, played an incredibly important role in this process. From the very beginning, I knew I could count on him for support on all levels necessary. And not only during this thesis. Starting halfway through my bachelor, Aaron has been there for me during multiple key projects. Thank you, Aaron. Secondly, I want to thank prof. dr. Martijn Warnier for being the second supervisor in this thesis committee and providing me with a fresh perspective on key aspects of the contents of this thesis. At the most crucial meetings, Martijn has shared incredibly important thoughts and insights on my work. Thank you for that. Thirdly, I would like to thank dr. Marcus Westberg. Thank you for the excellent supervision of my day-to-day process. Thank you for answering the ongoing stream of questions I had. I very much appreciate your willingness to discuss any aspect of the thesis, and other matters, in the kindest and most supportive way possible.

Second, the end product that lies before you would not have been possible without the help of 12 XAI experts from all over the world and over 200 survey respondents. These experts in particular helped me to evaluate the framework and frankly, a lot of other aspects of my thesis. Their enthusiasm for the topic is greatly appreciated: thank you all. Lastly, I would also like to thank my family and friends for their ongoing support. Especially during my thesis in this case, but moreover throughout my entire education.

David Lensen
Delft, May 2023

Executive Summary

Explainable AI (XAI) systems have gained prominence in recent years due to the increasing demand for transparency, trustworthiness, and understandability in AI models. XAI is a form of AI where the users of the model can understand the reasoning behind the decision or classification made by the model (Vilone & Longo, 2021) as opposed to the black-box form of AI (Castelvecchi, 2016). When designing an XAI, it is of great importance to ensure that the end user of the XAI will be able to satisfactorily interpret the explanation. Currently, there are numerous frameworks available to assess to what extent XAI is interpretable for expert interpreters (Jin et al., 2022), however, metrics for assessing to what extent XAI is interpretable for laypeople do not yet exist. This research aims to address this gap by identifying key interpretability factors and examining their trade-offs from the perspectives of both XAI developers and laypeople end users. Therefore, the main research question of this research is:

"How can XAI developers assess to what extent XAI is interpretable for laypeople?".

Consequently, the primary research objective is to develop a comprehensive framework for layperson XAI interpretability that encompasses essential factors and their relationships. In pursuit of an extensive answer to this research question, the Design Science Research Methodology has been applied (Peffers et al., 2007). Starting by combining the conclusions of multiple literature reviews to inform the development of the preliminary XAI interpretability framework, including a comprehensive list of factors and their relationships. The framework is accompanied by a set of key principles for XAI interpretability.

Since this framework is intended to be used by XAI developers and experts, and used on XAI's that are intended for laypeople, the framework should be validated on both sides. Firstly, semi-structured interviews were conducted with 12 experts in the field of XAI to gather their feedback on the preliminary framework. The interviews aimed to validate and refine the framework, and to identify potential shortcomings. The feedback from the experts was used to revise the framework and refine the set of key principles. Secondly, a survey is distributed among laypeople considering a specific use case. This survey provides insights in understanding their preferences regarding trade-offs among various interpretability factors. The results were, again, used to further refine the framework and key principles. This results in the final framework and set of key principles that can be used for assessing how interpretable the explanation part of XAI is to laypeople.

In the end, the core of the deliverable of this research is the theoretical framework. This includes critical factors such as simplicity, transparency, comprehensiveness, complexity, clarity, generalizability, trustworthiness, abnormality, explanation fidelity, model fidelity, intentionality, relevance, affordance, coherence with prior beliefs, and actionability. The final framework can be seen in Figure 17 in chapter 8. Surrounding the framework, and deepening the relationships inside the framework, are the key principles. They are presented as actionable guidelines and should be interpreted by XAI developers and researchers as necessary to read before designing or working on an XAI. The key principles emphasize the importance of trustworthiness, relevance, simplicity, clarity, coherence, intentionality, actionability, fidelity, contextualization, and ethical considerations in designing XAI explanations for laypeople. Striking the right balance among these factors while taking into account contextual factors and potential trade-offs is crucial for achieving optimal interpretability. Moreover, engaging with stakeholders and addressing ethical issues are vital in developing interpretable and responsible XAI systems. Regular evaluations and iterative improvements ensure that explanations continue to evolve and meet users' needs effectively. The final set of key principles can also be found in chapter 8 and the practical guidelines derived from these can be found in chapter 10.

Finally, the implications of the research findings will be discussed, which provide valuable insights for advancing XAI research and system design. The refined interpretability framework and key principles serve as a foundation for both novice and experienced XAI researchers and developers, encouraging a more methodical approach to XAI system design and fostering interdisciplinary research among experts in AI, human-computer interaction, psychology, and philosophy. By enhancing user trust and understanding, these findings can promote the responsible adoption of AI systems in various industries and sectors, such as healthcare, finance, and transportation. Furthermore, the insights gained can inform the development of policies and regulations governing AI technologies, supporting the creation of more effective guidelines and standards that promote responsible AI practices.

Keywords: XAI, interpretability, understandability, evaluation, laypeople, framework, guidelines

Table of Contents

Chapter 1. Introduction 11

1.1. Accountability problem for AI 11

1.2. Black box implications for AI 11

1.3. Introduction to XAI 11

1.4. Introduction to the knowledge gap 12

1.5. Research question 14

1.6. Scope of research 15

1.7. Scientific relevance and impact of research 16

1.8. Program-specific relevance 17

Chapter 2. Methodology 18

2.1. Research approach 18

2.2. Research questions 18

2.2.1. Methodology sub-question 1 19

2.2.2. Methodology sub-question 2 19

2.2.3. Methodology sub-question 3 20

2.2.4. Methodology sub-question 4 21

2.2.5. Methodology sub-question 5 21

2.3. Research flow diagram 23

2.4. Timeline of project 23

Chapter 3. Background 25

3.1. AI defined 25

3.2. XAI classification 25

3.2.1. LIME (Local Interpretable Model-Agnostic Explanations) 26

3.2.2. SHAP (Shapley Additive Explanations) 26

3.2.3. LRP (Layer-wise Relevance Propagation) 27

3.3. Explanations explained 27

Chapter 4. Interpretability of XAI for experts 29

4.1. Clarity 29

4.2. Transparency 30

4.3. Relevance 31

4.4. Trustworthiness 31

4.5. Overlap with human intuition 32

4.6. Intuitive understandability 32

4.7. Information Transfer Rate 33

4.8. Recall Response Time 34

Chapter 5. Interpretability of any explanation for laypeople 35

5.1. Principles from Thagard 35

5.2. Probability 36

5.3. Model fidelity 36

5.4. Abnormality *37*

5.5. Intentionality (and functionality) *37*

Chapter 6. Creation of framework **39**

6.1. Conclusion on sub-question 1 and 2 ..... *39*

6.2. First framework prototype *41*
6.2.1. Incorporating factors 41
6.2.2. Factor examination and analysis 41
6.2.3. Thematic grouping and prioritization 41
6.2.4. Identifying overlaps and bridging gaps 41
6.2.5. Iterative Refinement 41

6.3. Key principles of the first framework prototype *44*

Chapter 7. Evaluation of framework part I **46**

7.1. Nature of first evaluation round *46*

7.2. Nature of experts and developers *47*

7.3. Evaluation of first framework prototype *47*
7.3.1. Individual factors 48
7.3.2. Operationalization and practicality 50
7.3.3. Context 51
7.3.4. Measuring constructs 52
7.3.5. Level of abstraction 53
7.3.6. Other interesting comments 53

7.4. Second framework prototype *53*
7.4.1. Revisiting and revising factors 54
7.4.2. Context-awareness and adaptability 54
7.4.3. Practicality, measurability, and abstraction levels 54
7.4.4. Factor weights 54
7.4.5. Presentation of the second framework prototype 54

7.5. Key principles of the second framework prototype *57*

7.6. Conclusion and reflection on the second framework prototype *58*

Chapter 8. Evaluation of framework part II **60**

8.1. Nature of second evaluation round *60*
8.1.1. Transparency, complexity, comprehensiveness, simplicity, generalizability and clarity 61
8.1.2. Coherence with prior beliefs and affordance 62
8.1.3. ITR and RRT 62

8.2. Survey practicalities *62*

8.3. Survey results *62*

8.4. Survey conclusions *64*
8.4.1. Explanation set 1 64
8.4.2. Explanation set 2 65
8.4.3. Explanation set 3 66
8.4.4. Explanation set 4 66
8.4.5. Explanation set 5 66
8.4.6. Explanation set 6 67
8.4.7. Aggregated conclusions 67

8.5. Final framework prototype *69*

8.6. Key principles of the final framework *69*

Chapter 9. Discussion **72**

9.1. Shortcomings and recommendations for future research ... *72*
9.1.1. Factor selection ... 72
9.1.2. First framework prototype and first set of key principles ... 72
9.1.3. Laypeople and experts ... 73
9.1.4. Theoretical nature ... 73
9.1.5. Assumptions ... 73
9.1.6. Interview shortcomings ... 74
9.1.7. Survey shortcomings ... 75
9.1.8. Other recommendations for future research ... 75
9.2. Ethical note ... *75*

Chapter 10. Conclusion ... **77**
10.1. From key principles to exemplary actionable guidelines ... *77*
10.2. Comparison with literature ... *79*
10.3. Scientific implications and relevance ... *80*
10.3.1. Advancing XAI research and system design ... 80
10.3.2. Informing comparative studies and benchmarking ... 80
10.3.3. Fostering Interdisciplinary Research ... 80
10.4. Societal Relevance ... *80*
10.4.1. Enhancing trust and adoption of AI systems ... 80
10.4.2. Empowering consumers and encouraging informed decision making ... 81
10.4.3. Promoting accessibility and inclusion ... 81
10.4.4. Informing policy and regulation ... 81
10.5. Relevance to Complex Systems Engineering and Management ... *81*
10.6. Reflection on Design Science Research Methodology ... *82*
10.7. Closing remarks ... *82*

Bibliography ... **83**

Appendix A. Research Flow Diagram ... **92**

Appendix B. Gantt chart ... **93**

Appendix C. Interviewee institutions ... **94**

Appendix D. XAI developer and expert interview ... **97**

Appendix E. Interview summaries ... **102**

Appendix F. Survey for evaluation on laypeople ... **118**
Section 1. Introduction ... *118*
Section 2. Evaluation of explanations ... *118*
Set 1: Comprehensiveness, transparency, simplicity, and generalizability ... 119
Set 2: Complexity, transparency, simplicity and clarity ... 120
Set 3: Abnormality, coherence with prior beliefs and affordance ... 121
Set 4: Intentionality and actionability ... 121
Set 5: Model fidelity and explanation fidelity ... 122
Set 6: Trustworthiness and relevance ... 123
Section 3. Who are you? ... *123*
Section 4. Final statement ... *124*

Appendix G. Survey Data Cleaning ... **125**
Section 1. Laypeople – expert distinction ... *125*

Section 2. Attention check *125*
Section 3. Amount of time taken *126*

Appendix H. Demographic data **128**
Section 1. Gender *128*
Section 2. Age *128*
Section 3. Geographical location *129*
Section 4. Ethnicity *131*
Section 5. Employment status *131*
Section 6. Student status *132*
Section 7. AI & XAI experience *132*

Appendix I. Human Research Ethics: DMP **134**

Appendix J. Human Research Ethics: Informed Consent **140**

Appendix K. Human Research Ethics: Approval **143**

List of Figures

Figure 1. XAI explained 12
Figure 2. Explainability from model perspective and interpretability from user perspective 13
Figure 3. Accuracy and interpretability of numerous AI model types 14
Figure 4. Layperson and expert distinction used throughout this thesis 16
Figure 5. Google's trend analysis on the term 'Explainable AI' 17
Figure 6. Design Science Research Methodology (DSRM) Process Model for Information Systems (Peffers et al., 2007) 18
Figure 7. Venn Diagram visualizing how the requirements from the first framework prototype are based on the two previously conducted literature reviews 21
Figure 8. Evaluation process for the framework based on the methodology 23
Figure 9. Classification of Explainable AI 26
Figure 10. Five categories of scientific explanation (Overton, 2012) 28
Figure 11. General structure of a theory-data explanation (Overton, 2012) 28
Figure 12. First framework prototype. Red arrows represent negative relationships whereas blue arrows represent positive relationships. 42
Figure 13. Main categories of practical conclusions from the expert interviews 48
Figure 14. Importance of individual factors based on expert interviews 49
Figure 15. Second framework prototype 55
Figure 16. Holistic versus reductionist factors 62
Figure 17. Final framework specifically for medical contexts 69
Figure 18. Research Flow Diagram 92
Figure 19. Gantt chart 93
Figure 20. Number of experts respondents in the survey 125
Figure 21. Number of respondents that have failed to complete attention check 126
Figure 22. Histogram of time taken to complete survey 127
Figure 23. Gender distribution 128
Figure 24. Age distribution 128
Figure 25. Geographical distribution 129
Figure 26. Participant distribution across the world (n = 157) 129
Figure 27. Participant distribution across the United States (n = 13) 130
Figure 28. Participant distribution across the United Kingdom (n = 124) 130
Figure 29. Participant distribution across the world (n = 18) 130
Figure 30. Ethnic distribution 131
Figure 31. Employment distribution 131
Figure 32. Student distribution 132
Figure 33. AI & XAI experience distribution 132

List of Tables

Table 1. Literature used for initial literature review 13
Table 2. Some definitions of AI organized into four categories (Russell & Norvig, 2022) 25
Table 3. Summarized assessment factors 39
Table 4. Categorization of interviewees based on expertise, domain, and area 47
Table 5. Relative importance of individual factors used for weight attribution 50
Table 6. Explanation sets specifics for the survey 61
Table 7. Core survey results. U = natural understandability and S = general satisfaction of the explanation 64
Table 8. Survey conclusions 68

List of Abbreviations

AI. Artificial Intelligence
DARPA. Defence Advanced Research Projects Agency
DSRM. Design Science Research Methodology
GDPR. General Data Protection Regulation
ITR. Information Transfer Rate
LIME. Local Interpretable Model-agnostic Explanation
LRP. Layer-wise Relevance Propagation
ML. Machine Learning
RFD. Research Flow Diagram
RRT. Recall Response Time
SFL. Systemic Functional Linguistics
SHAP. Shapley Additive Explanations
XAI. Explainable Artificial Intelligence

Chapter 1. Introduction

1.1. Accountability problem for AI

The number of processes that are being taken over completely by AI is increasing rapidly. A commonly raised question concerns accountability. There are two general views on accountability. The first definition is that accountability refers to the capability of providing a clear and justifiable reason for the actions or decisions made (for example by an AI system). This involves explainability and openness in the decision-making process, allowing for accountability to be established. Secondly, there is accountability as in responsibility: in case of malicious consequences caused by the result yielded by the AI algorithm, who can be held accountable for these consequences? Both definitions of accountability are relevant regarding AI. However, for illustration purposes, the latter definition will be used in this section.

Firstly, assessing accountability at the designer or creator side of the algorithm raises multiple issues. Helen Nissenbaum was one of the first scientists to be concerned about accountability for designers when using computerized systems. As early as in 1996, she wrote a paper in which she described four barriers that obscure accountability in a computerized society. These four barriers are rather self-explanatory: *many hands*, *bugs*, *computer as a scapegoat,* and *ownership without liability* (Nissenbaum, 1996). And still, at the time of writing this report, 26 years later, these four barriers very well illustrate the difficulty in designating accountability when a process is aided by an algorithm (Cooper et al., 2022). Secondly, placing responsibility on the user of the algorithm is difficult, as, in a significant proportion of the cases, the user has no to very little influence on the content of the algorithm (Bader & Kaiser, 2019). Therefore, how can the user be held accountable for actions an unknown algorithm makes?

Currently, most case studies show that the creators of the algorithms sign off their accountability to the users during the acquisition of the product containing the algorithm. For example, when consumers buy a Tesla with 'Full Self-Driving Capability', Tesla simply states that these capabilities are solely included to *assist* the driver (user) and that therefore, the user is responsible at all times (Ferrara, 2016; Tesla, 2022). However, this does not seem morally correct.

1.2. Black box implications for AI

Besides the accountability problem, decision-makers that base their decisions on AI often do not understand the reasoning behind the model. This causes most AI systems to be seen as a black box. This affects two types of users of AI. Firstly, users that do not trust the model (or a prediction made by the model), will not use the model (or the prediction made by the model) if the situation even allows them to disregard the outcome of the model (Ribeiro et al., 2016). Secondly, in some cases, the results of AI models are not solely to assist humans in their decision-making process anymore: the results yielded by the algorithm have already made the decision. In these cases, if there is a human involved in the process, he/she most of the time simply needs to adhere to the results of the algorithm (Bader & Kaiser, 2019).

To illustrate: the European Commission proposed further restricting the use of AI by issuing the 'Artificial Intelligence Act' (European Commission, 2021). The EU wants to ensure that all Europeans can trust the AI they are using through the implementation of the act. They intend to enhance AI's transparency, governance, and insights.

1.3. Introduction to XAI

A solution to morally sign off accountability (or: responsibility) to the user, and to deal with the implications imposed by the black box model, is to make sure AI can be explained. Explainable AI (XAI) is a form of AI where the users of the model can understand the reasoning behind the decision or classification made by the model (D. Gunning et al., 2019; Vilone & Longo, 2021) as opposed to the black-box form of AI (Castelvecchi, 2016), see Figure 1. XAI plays an

important role in understanding the underlying mechanisms and decision-making processes of AI systems. This is particularly important for safety-critical systems, such as autonomous vehicles, healthcare applications, and financial systems, as it can help to identify potential problems and minimize risk (Arrieta et al., 2020). The ability to explain AI decisions can also help to build trust in AI systems. By making AI decisions more transparent, users can gain an understanding of how a system works and can make more informed decisions. Additionally, XAI can help to reduce the risk of bias and errors by providing explanations which can be used to identify and correct any bias or errors (Arrieta et al., 2020). XAI can also be used to improve the accuracy of AI systems by providing explanations which can be used to identify and correct any errors or limitations in the system. By understanding the underlying mechanisms of a system, it is possible to develop more effective algorithms and improve the accuracy of the system (Arrieta et al., 2020). Therefore, explainable AI plays an important role in improving the safety, trustworthiness, and accuracy of AI systems. It is essential for ensuring the reliability and fairness of AI systems and for developing a better understanding of the underlying mechanisms and decision-making processes.

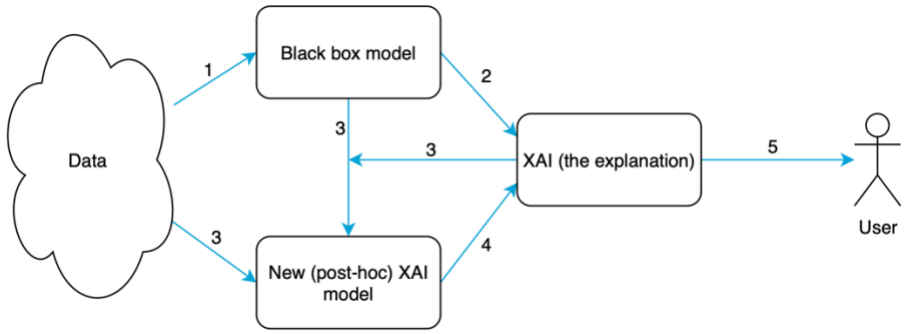

Figure 1. XAI explained

1.4. Introduction to the knowledge gap

The goal of the remainder of this chapter is to conclude with the knowledge gap and the main research question based on a review of scientific literature about XAI. Since trustworthiness is a significant factor within the field of AI, the starting point of the following literature review is in that era. For the literature review, the methodology outlined by Van Wee and Banister (2016) is followed. Only anglophone academic literature written between 2013 and 2023 related to XAI found on Scopus is used for the initial search. The entered search query in Scopus is:

explain AND (algorithm* OR (ai OR (artificial AND intelligence))) AND trust**

From this point onwards, the most important selection criterion is that the focus of the article lies in describing ways to assess the interpretability of XAI. The final choice of papers for this review is made using merely backward snowballing (Wee & Banister, 2016), whilst taking into account how many times the article was cited (considering that articles with more citations are more likely to be scientifically valuable). Eventually, a total of 12 articles that matched all criteria were found. The selected articles are presented in Table 1 below in alphabetical order.

Author(s)	Year	Title
Adadi, A., Berrada, M.	2018	Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)
Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C.	2021	Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review
Bader, V., & Kaiser, S.	2019	Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence

Author(s)	Year	Title
Bohanec, M., Robnik-Šikonja, M., & Kljajić Borštnar, M.	2017	Decision-making framework with double-loop learning through interpretable black-box machine learning models
Chromik, M.	2021	Making SHAP Rap: Bridging Local and Global Insights Through Interaction and Narratives
Liao, Q. V., Singh, M., Zhang, Y., & Bellamy, R. K. E.	2020	Introduction to Explainable AI
Miller, T., Howe, P., & Sonenberg, L.	2017	*Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*
Mohseni, S., Zarei, N., & Ragan, E. D.	2020	*A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems*
Nissenbaum, H.	1996	Accountability in a computerized society
Ribeiro, M.T., Singh, S., Guestrin, C.	2016	"Why should I trust you?" Explaining the predictions of any classifier
Robnik-Šikonja, M., & Kononenko, I.	2008	Explaining Classifications for Individual Instances.
Vilone, G., & Longo, L.	2021	Notions of explainability and evaluation approaches for explainable artificial intelligence

Table 1. Literature used for initial literature review

One of the most essential criteria for explainability regarding AI, is interpretability (Ribeiro et al., 2016). Interpretability means an understanding is created between the input variables and the model response. Interpretability describes to what degree a user can understand the explanation (Biran & Cotton, 2017). One can assess a model as explainable by looking at it from the model perspective, but only a user can assess a model as interpretable (as visualized in Figure 2).

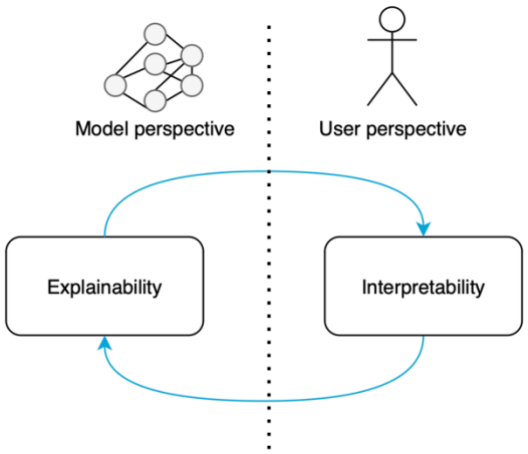

Figure 2. Explainability from model perspective and interpretability from user perspective

Within AI, a trade-off can be observed. First, there are simple linear models that can be easily interpreted by humans. These linear models will most likely not lead to adequate predictions for complex problems. The other option concerns highly non-linear models that provide increasingly well performance on most tasks but are simply too complex for humans to understand. Neural networks for instance often have millions of parameters which simply exceed human capabilities. A graphical overview of common AI model types including their accuracy and interpretability is provided in Figure 3.

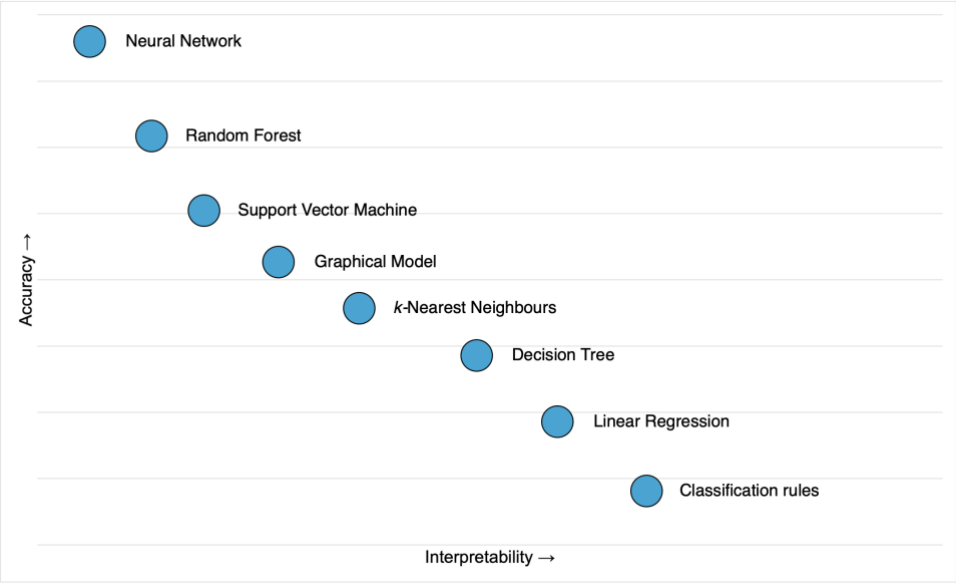

Figure 3. Accuracy and interpretability of numerous AI model types

A brief introduction to some of the researched factors for assessing the interpretability of an XAI method will be provided in this section. There are two types of factors available: human-centric (often qualitative) or mathematic (quantitative) (Antoniadi et al., 2021). Human-centric factors are measured by looking at the direct influence on the people that use the model (Vilone & Longo, 2021). One of the most commonly mentioned human-centric factors concerns the user determining the pragmatic influences of the explanation (Antoniadi et al., 2021; Chromik, 2021; Miller et al., 2017; Mohseni et al., 2020). Pragmatic influences can be usefulness, generalism, coherence, simplicity, relevance, etc. Another commonly mentioned factor by many authors is one where the user rates to what degree the explanation meets or overlaps with the explanation that the user would have given (Chromik, 2021; Mohseni et al., 2020). Most of the currently available mathematic factors tend to measure the actual XAI model's performance, as opposed to the interpretability of the explanation. For example, the most commonly used mathematical factor is the accuracy of the model's prediction (Liao et al., 2020). However important for the functionality of the model itself, they are not directly relevant to the interpretability of XAI.

As the above literature suggests, several factors have been developed to measure to what extent XAI makes AI interpretable. However, it should be noted that most (if not all) of these studies take interpretability by experts as their benchmark (Bohanec, et al., 2017; Robnik-Šikonja & Kononenko, 2008). Therefore, these factors and frameworks cannot be directly applied to assessing interpretability for laypeople (non-experts, regular people, ordinary users/consumers). However, XAI should also allow models to be interpretable to ordinary consumers. An example of a layperson using AI can be seen in the healthcare industry: medical apps for online self-diagnosis (Symptomate, 2022). For patients, it is of great importance that the explanation given together with the diagnosis can be interpreted adequately. The problem is that there is not one framework available to assess that degree of adequacy: this is the knowledge gap for this research. Therefore, the key deliverable of this research will concern a framework suitable for helping in assessing to what extent XAI is interpretable to these ordinary users. This process is also visualized in figure 1, namely the process represented by arrow number 5. Hence, a research question has been drawn in the next section.

1.5. Research question

The main research question that will be answered in this report can be stated as follows:

"How can XAI developers assess to what extent XAI is interpretable for laypeople?"

Accordingly, the key deliverable of the research will concern a framework that can be directly used by XAI developers, researchers, and other stakeholders in the design and evaluation of XAI systems, to ensure that their XAI explanations are effective, efficient, and accessible for all: to assess to what extent their explanation of an AI is interpretable to laypeople. The first classification that can be made about the above research question concerns the epistemological perspective of the research. The epistemological perspective of the research is of *constructivist nature* since it can be answered by constructing an understanding of the world to create assumptions about reality. Furthermore, it is inductive research with an exploratory nature (Hasa, 2020; Mackenzie, 2011), since a framework will be created in absence of a comparable framework.

1.6. Scope of research

There are several relevant dimensions of this research that would benefit from a clearly set scope. Namely: the deliverable itself, the target group, the explanation purpose, and the explanation methods.

- The deliverable requires some scoping on its own. In section 1.5, the deliverable is mentioned to be a framework. For this research, the core of the framework will be XAI evaluation principles *and/or* guidelines to facilitate XAI developers. The framework will be considered the outer shell of the deliverable, whereas the inner core of the deliverable will concern the principles/guidelines.

- The target group for the framework can be interpreted in two different ways. On the one hand, the framework is intended to be *used by* XAI developers to assess to what extent a generated explanation is interpretable. On the other hand, the framework is intended to be *used on* laypeople instead of experts. Within this research, laypeople are defined as people that are *not* experts on the field in which the XAI operates (the domain). Since this distinction is quite important throughout this research, a more specific definition is also provided. There is a difference in judgement between laypeople and experts that can be attributed to two sources (Bolam et al., 2003; Ganzach, 1994):
    - Differences in information processing: Experts have better representations of tasks in their area of expertise, they process information more efficiently, their search for information is more relevant to the task at hand, and they tend to use less information in their judgments (Ganzach, 1994).
    - Differences in integrating information: Laypersons and experts rely on different theories in combining information to form a judgment. The integration rules that guide experts' judgments may be more linear or configural, while laypersons may have more intuitive or subjective approaches. The experts may have a collective rationality in their subject of expertise, which is distinct from the laypersons' intuition (Ganzach, 1994).

  When looking at the previously mentioned example of the self-diagnosis app: a doctor would be a domain expert using the AI, whilst a patient is a layperson. The exact distinction between a layperson and an expert is rather vague and field-dependent (Newman, 2014).

  It should be noted that the terminology will remain consistent throughout this entire thesis. A layperson is considered someone with no to little domain knowledge (the bottom side of Figure 4). An XAI expert is considered someone with a sufficient level of XAI expertise (right side of Figure 4). And a domain expert is someone with a sufficient body of knowledge regarding the domain of the XAI (for example the medical domain, the top side of Figure 4). Figure 4 presents a 2 by 2 matrix with domain knowledge on the y-axis and XAI knowledge on the x-axis.

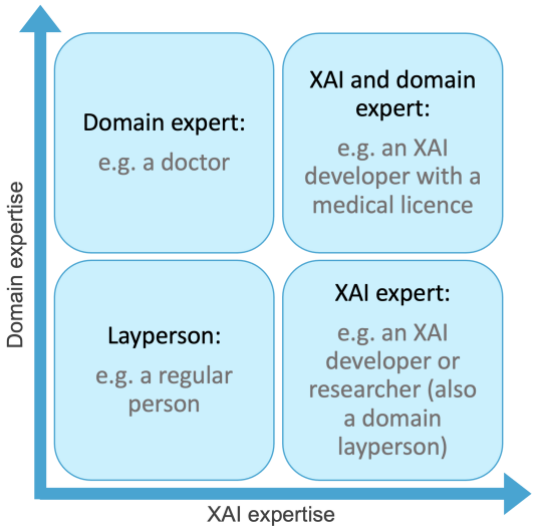

Figure 4. Layperson and expert distinction used throughout this thesis

- This research will focus on post-hoc (or extrinsic) models (see Figure 9 in section 3.2). A post-hoc model provides explanations for its predictions after the fact, rather than during the prediction process. Unlike other XAI models, which are designed to be transparent and interpretable from the outset, post-hoc XAI models are typically trained as black-box models and then modified or analysed after the prediction is made in order to provide explanations.

- Lastly, the explanation type is of importance. Explanations can be presented in, for example, a textual format (natural language descriptions that explain the reasoning behind the AI's decision), graphically (visual representations, such as graphs, charts, or diagrams, to show the relationships between features, variables, or concepts), image-based (for example heatmaps, saliency maps, or attention maps that highlight areas in the input image that were most influential in making the decision), or interactive (allow users to interact with the AI model, exploring different scenarios or inputs to see how the model's decisions change) (Guidotti et al., 2018; Linardatos et al., 2020; Shevskaya, 2021). For the purpose of this research, textual explanations are most relevant. Therefore, this research will only concern textual explanations.

1.7. Scientific relevance and impact of research

As can be concluded from Google's trend analysis in Figure 5 below, XAI is a rapidly emerging research field. This is mainly caused by all the advantages XAI has over non-explainable AI as illustrated in this chapter. Furthermore, governmental organizations have started regulating the use of AI systems. The main marker of the great spark of XAI research is in late 2016 when DARPA published a report about the importance of XAI (DARPA, 2016). DARPA (Defence Advanced Research Projects Agency) is an agency of the US Department of Defence, responsible for the development of new technologies for military use. It conducts cutting-edge research in fields like AI, robotics, biotechnology, cybersecurity, and microelectronics, working with universities, companies, and other government organizations. DARPA has a long history of supporting ground-breaking work, including the development of the Internet and GPS, and continues to shape the future of technology through its ongoing research and development programs.

Thereafter, in May 2018, the European Union replaced the 1995 EU Data Protection Directive and strengthens EU data protection law by taking on the General Data Protection Regulation (GDPR). The GDPR aims to protect the privacy and personal data of EU residents and to give individuals greater control over their personal information. It sets out strict rules for the collection, processing, storage, and use of personal data, and gives individuals the right to access, correct, and delete their personal data. The European Commission built on the GDPR

by proposing to restrict the use of AI through proposing the 'Artificial Intelligence Act' in April 2021 (European Commission, 2021). The act aims to ensure that AI systems used in the EU are trustworthy, transparent, and respect individuals' rights. The act establishes specific obligations for organizations using AI, including the need for human oversight, and establishes a framework for the assessment and regulation of high-risk AI applications. They intend to enhance AI's transparency, governance, and insights. To ensure that AI systems are more comprehensively compliant with this act, an option for AI developers would be to ensure explainability of their AI systems. Once an AI system is 'explainable', the likelihood for the AI system to be fully compliant with this act is inherently quite high.

As this research aims to provide a framework that assesses interpretability, this research is closely linked to the enhancement of XAI. That is why this research will form one of the necessary steps to consequently increase the quality of XAI.

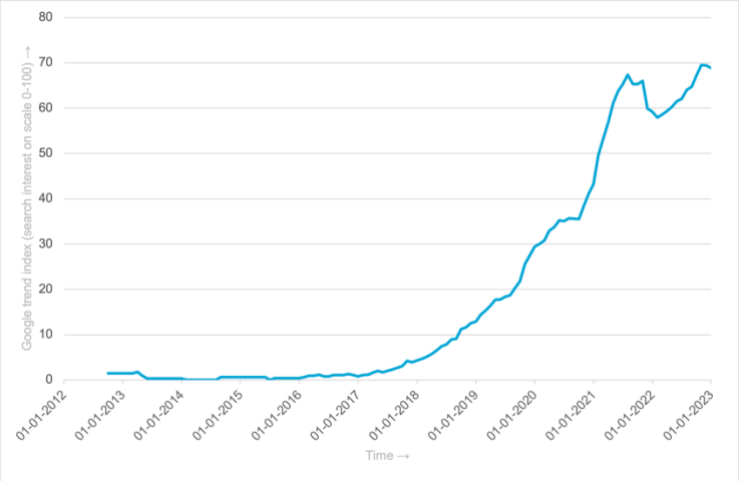

Figure 5. Google's trend analysis on the term 'Explainable AI'

1.8. Program-specific relevance

In the master's program of Complex Systems Engineering and Management, students learn to think about technologies as part of a socio-technical system. Technologies will not be assessed solely based on technological aspects. This research does just that. This research assesses XAI in the context of ethics, human behaviour, regulations, and technology. Furthermore, this research aims to design in a socio-technical system, aimed at an effective intervention. This combination results in a high level of complexity. Furthermore, the XAI topic is very closely linked to the Information and Communication track within the master program. It incorporates many elements taught within all track-specific courses, such as AI, robotics, accountability, and the societal impact of ICT.

Chapter 2. Methodology

2.1. Research approach

The main objective of the research is to develop a framework that can be used in assessing to what extent XAI is interpretable for laypeople. Since the objective is to create a design, the research will follow a design approach (Peffers et al., 2007). Peffers and his colleagues created a process model specifically for the design science research methodology, needed to successfully carry out design science research and have a "mental model" for its presentation. Adhering to this model will ensure that the approach follows scientifically accepted standards and guarantees repeatability. This model can be found in Figure 6.

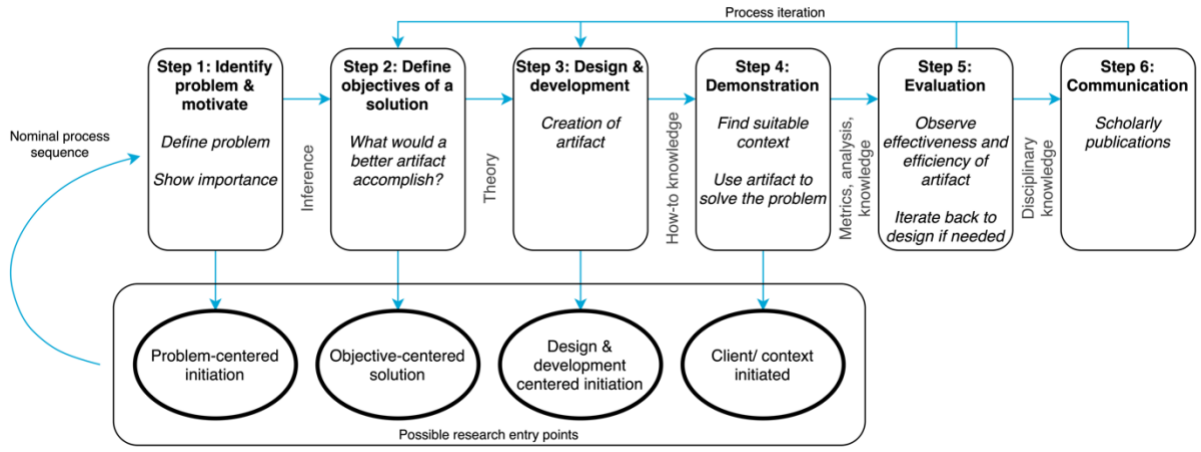

Figure 6. Design Science Research Methodology (DSRM) Process Model for Information Systems (Peffers et al., 2007)

As can be seen in Figure 6, several steps need to be undertaken to complete design science research. A problem-centred approach is the basis for this research (the idea for the research resulted from observing the problem itself), therefore starting with step 1. That first step (*identify the problem and motivate*) is rather introductory and will already be dealt with in the earlier stages of this research. The second stop (defining *the objectives of a solution*) will require multiple extensive literature reviews, which can be combined to the actual design of the framework, which is the third step (*design and development*). To allow evaluation of the framework (step 5), first, a demonstration of the framework (step 4) is necessary. Therefore, expert interviews combined with user cantered surveys seem a suitable approach. Lastly, the entire design should be communicated. This will happen by publishing the report in the form of a master's thesis. On a final note, it should be mentioned that the entire process is iterative. This means that after having completed a step, it is encouraged to go back to the previous steps to assess the validity of that step, bearing in mind the newly acquired knowledge from the latter step.

2.2. Research questions

As was concluded in chapter 1, the main research question that will be answered in this report can be stated as follows:

"How can XAI developers assess to what extent XAI is interpretable to laypeople?"

To answer this main research question with the process model from Figure 6 in mind, several sub-questions have been drafted. First of all, the main research question is split up into two components. The first component mostly concerns the interpretability of XAI part (sub-question 1), and the second component concerns the interpretability to laypeople part (sub-question 2). Once these first two sub-questions have been researched. A comparison can be made between interpretability with regard to XAI on the one hand and interpretability by laypeople

on the other hand, upon which conclusions should be drawn (sub-question 3), which will result in the creation of a framework. Lastly, this research intends to assess how well this framework can be applied in practice (sub-question 4 and 5).

2.2.1. Methodology sub-question 1

To fully understand how XAI explanations are evaluated by laypeople, the problem needs to be split up into two parts, as explained in the previous paragraph. On the one hand, assessment factors that measure to what extent XAI is interpretable for experts (not laypeople) should be looked at. This will be done by answering the following sub-question, which is in line with the second step of the DSRM model since it is part of defining objectives for the framework:

1. *"What are the assessment factors that measure to what extent XAI is interpretable to experts?"*

The first sub-question has already been an interesting research topic for several years. Therefore, the most logical approach to answering this question is by means of a literature review in combination with text analysis using the R-Studio software. This literature review will aim to bring together all relevant research on the evaluation factors of XAI.

For the literature review, the methodology outlined by Van Wee and Banister (2016) was followed. Only anglophone academic literature related to the interpretability of XAI found on Scopus was used. Finally, only articles that were cited by at least 20 other articles were included in the initial set. The search query entered in Scopus is:

(xai OR (explain* AND ai) OR (explain* AND artificial AND intelligence))
AND (interpretab* OR trust*) AND (metric OR review OR evaluation OR framework)
AND (user OR human)

This keyword combination, combined with the other criteria, produced 93 documents on Scopus. From this point onwards, the most important selection criterion was that the focus of the article was on describing ways to assess the interpretability of XAI. The final choice of papers for this review was made using merely backward snowballing (Wee & Banister, 2016), whilst taking into account how many times the article was cited (considering that articles with more citations are more likely to be scientifically valuable).

2.2.2. Methodology sub-question 2

On the other hand, assessment factors that measure to what extent any explanation (non-XAI) is interpretable to laypeople should be reviewed. When aiming to design XAI that is truly able to provide an interpretable explanation to people, it is fair to say that looking at humans explaining decisions to other humans is a good way to improve the analysis (Miller, 2019). This will be done by answering the following sub-question, which is in line with the second step of the DSRM model since it is part of defining objectives for the framework:

2. *"What are the assessment factors that measure to what extent any explanation is interpretable to laypeople?"*

The second sub-question is also a topic that has been researched thoroughly, specifically in the field of the social sciences. Therefore, another literature review will be performed to answer this sub-question. This literature review will aim to bring together all relevant research on interpretability factors for regular people regarding any explanation (measuring how one can assess how well something can be interpreted by someone).

For this literature, the same methodology was followed (Wee & Banister, 2016). Only anglophone academic literature related to interpretability factors for regular people found on Scopus was used. The search query entered in Scopus is:

explain* AND (interpretab* OR trust*) AND (metric OR review OR evaluation OR framework) AND (user OR human)

From this point onwards, the most important selection criterion was that the focus of the article was on interpretability factors. The final choice of papers for this review was made using backward and forward snowballing (Wee & Banister, 2016), whilst taking into account how many times the article was cited. Given that this type of research is not very tied to technological developments or otherwise time-sensitive variables, the year in which the article was published is not considered.

2.2.3. Methodology sub-question 3

After answering the first two sub-questions, the results should be combined to lay out the principles and requirements for the first framework prototype. This is summarized in the third sub-question:

3. *"What are the requirements for the framework prototype as concluded from comparing the results from sub-question 1 and sub-question 2 regarding the main research question?"*

These requirements are found through a systematic and structured process, rather than being discovered haphazardly. The first two sub-questions help to define the scope and goals of the framework prototype, while the third sub-question brings the focus onto the specific requirements that are needed to make the first framework prototype a reality. The process of combining the results from the first two sub-questions to lay out the principles and requirements for the framework prototype is a critical step in the design process: it is the core of the design process. It involves synthesizing the information gathered from the first two sub-questions to form a comprehensive understanding of the requirements for the framework prototype. This process can be broken down into several steps:

1. Before beginning the process of combining the results, it is important to thoroughly review the answers to the first two sub-questions. This will help ensure that all relevant information is taken into account when forming the principles and requirements for the framework prototype.
2. Next, look for common themes and patterns in the results from the first two sub-questions. This will help to identify areas where the requirements are aligned and where additional clarification may be needed. Factors that only occur in the analysis of either the first or the second sub-question should also be included, if inclusion can be justified. This will cause the final set of factors/requirements to be in some sense a weighted average of both sub-question 1 and sub-question 2.
3. Using the information gathered from the first two sub-questions, formulate a set of principles and requirements for the framework prototype. These should be specific and clearly defined to ensure that the framework prototype can be successfully developed.
4. Once the principles and requirements have been formulated, it is important to validate them to ensure that they are accurate and complete considering the purpose of the final framework (evaluating XAI interpretability for laypeople). This initial validation will be done based on the previous literature review.
5. Based on the results of the validation process, update the principles and requirements as needed. This will help to ensure that the framework prototype is developed in accordance with the most current and accurate information.

In conclusion, combining the results from the first two sub-questions to lay out the principles and requirements for the framework prototype is a critical step in the design process. It requires a systematic and thorough approach to ensure that all relevant information is taken into account and that the requirements for the framework prototype are clearly defined. As stated earlier, this process is the core of the design process (step 3 in the DSRM model) and can be summarized in the third sub-question. Furthermore, it is visualized in the Venn diagram in Figure 7.

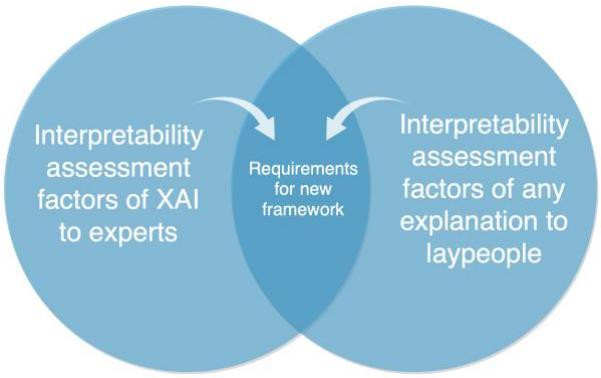

Figure 7. Venn Diagram visualizing how the requirements from the first framework prototype are based on the two previously conducted literature reviews

2.2.4. Methodology sub-question 4

Finally, the framework prototype should be evaluated. This represents steps 4 and 5 of the DSRM process model. However, the framework will require multiple rounds of evaluation. As stated in section 1.6, the framework will be *used by* XAI developers and *used on* laypeople. Therefore, the framework will require evaluation on both sides. Namely expert interviews to assess the workability and efficiency of the framework from the viewpoint of an XAI developer and a use case presented in the form of a survey where the framework will be applied to will show how the framework operates in practice on the laypeople. Any elements of the prototype that require changes, can thereafter be altered accordingly. The first evaluation round (expert evaluation) is presented in the fourth research question:

4. *"How is the first framework prototype evaluated by experts and XAI developers in order to determine its strengths, limitations, and potential areas for improvement?"*

A total of twelve interviews will be performed with a diversified group of both XAI experts and XAI researchers. The interviews will be recorded, with consent of the interviewee. After the interviews are finished, they will be summarized in appendix E. At that point, the recording will be deleted, to protect the privacy of the interviewees. Using Atlas-TI, the summaries of the interviews will be thoroughly analysed, and conclusions can be drawn. These conclusions form the basis to reshape the first framework prototype and therefore facilitate the shift towards the second framework prototype.

2.2.5. Methodology sub-question 5

As stated in section 2.2.4, the second round of evaluation is by applying the framework in practice. This will be in the form of a use case presented in a survey. Data gathered from the survey for this second evaluation round should both be reliable and valid. This second evaluation round (evaluation in practice on laypeople) is presented in the fifth and final sub-research question:

5. *"How can the second framework prototype be applied in practice and thus evaluated?"*

Practically, the second framework prototype will be applied in the healthcare sector: to laypeople using XAI to help in determining causes for symptoms. This is a classic example of a layperson using an XAI. This will be done using a survey that needs to be filled in by at least 200 respondents, created using Qualtrics and mainly distributed using the platform Prolific.

By distributing the survey on the Prolific platform, participants are being paid to complete it, which introduces external motivation (financial compensation) as opposed to internal motivation (a genuine desire to contribute to sound research). This distinction in motivation could potentially lead to biased or less reliable results, as some participants may rush through the survey or provide low-quality responses just to receive the payment.

To mitigate this potential risk, an estimation has been made that the survey is estimated to take approximately 8 minutes to complete, as indicated by Qualtrics. This duration is a benchmark to ensure that participants are dedicating an appropriate amount of time and attention to the survey questions, which should result in more thoughtful and accurate responses. To further reduce the risk of biased or unreliable data, the decision has been made to exclude all respondents who took less than 5 minutes to complete the survey. This threshold is set to filter out participants who may have rushed through the survey without carefully considering their answers. By removing these respondents from the dataset, the quality of the remaining responses should be higher, leading to more reliable insights and conclusions.

Secondly, as a final measure to mitigate the potential risk of processing low-quality results, an attention check has been added to the survey. This has been done by adding the following question to the survey:

How do you view the following statement: I have never used a computer-like device before?

a. *Yes, this is true for me.*
b. *No, this is not true for me.*

Since all surveys are being filled in on either a smartphone, tablet, or computer (all computer-like devices), everyone should be answering B to this question. All responses that have not selected answer B, will automatically be discarded from the final dataset.

This process of excluding participants based on their response time aims to minimize the impact of external motivation on the survey results and to encourage participants to be more engaged and attentive when providing their answers. The goal is to collect high-quality data that genuinely reflects the opinions and experiences of the participants, leading to more accurate and robust findings from the research.

Since both sub-question 4 and sub-question 5 concern research on human subjects (interviews and surveys), it is of great importance to ensure that that process is ethically correct. To that end, TU Delft's Human Research Ethics Committee must review all plans and approve them. The results of this ethics process can be found in appendix I, J, and K.

In summary, the evaluation of the framework can be visualized in 3 phases (see Figure 8). First of all, the first framework prototype is created via a literature review (according to sub-questions 1 and 2. The second framework prototype is consequently created via expert interviews. Lastly, that second framework prototype needs to be validated in practice on laypeople, via a survey. This will result in the creation of the final framework. This evaluation process progress figure will be used throughout the thesis to maintain a clear overview of the process.

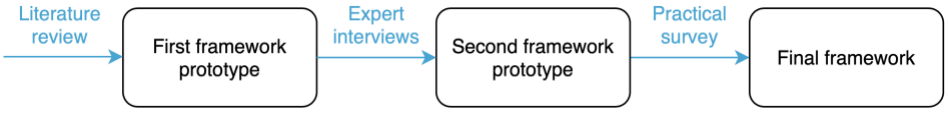

Figure 8. Evaluation process for the framework based on the methodology

Finally, Peffers et al. (2007) suggest communication of the findings. This however does not require a separate research question.

2.3. Research flow diagram

To summarize the overall flow of the research process, a research flow diagram has been created. The thesis project is already split up into chapters. Inside the chapters, the appropriate research steps, sub-questions, research methods and inputs/outputs are shown. The diagram follows the appropriate steps necessary for successful completion of design approach research (Peffers et al., 2007).

The research will start by defining the outline of the research in the introductory chapter of the thesis. This entails specifying the background, the problem at hand, the accordingly appropriate research question, and the scientific relevance. The output of this chapter will be a research outline. This research outline is used as input to define the research methodology. For the methodology, it is relevant to consider the research approach. After having decided upon all methods, the third chapter will form the basis for the development of the objectives of the framework. Firstly, sub-question 1 will be answered through a literature review in chapter 4. Secondly, another literature review will be performed to complete sub-question two in chapter 5. The combination of these two reviews will form the objectives of the framework. Thereafter, chapter 6 will draw conclusions from these objectives and translate them into the main relationships and components of the framework. This will be the core of the design process and will result in the framework prototype.

The seventh chapter will be about the application and therefore evaluation of the framework. This chapter will revolve around sub-question 4, which essentially consists of two components. First of all, the prototype of the framework (output from chapter 6) has to be evaluated according to Peffers et al. (2007), by means of twelve expert interviews. After the evaluation, it is most likely that certain elements require changes. This will be part of the design process. chapter 8 will follow the same structure. However, evaluation will happen by applying the framework in a practical survey, which will be sent out to laypeople. The result of the eighth chapter will be the final framework. The final two chapters are the discussion and conclusion. The final framework will be discussed, conclusions will be drawn (both on the framework and the research process), and recommendations will be finalized. The final research flow diagram is shown in Figure 18 in appendix A.

2.4. Timeline of project

To provide another dimension to the project planning, a Gantt chart is created. This chart incorporates many of the elements as shown in the RFD in Figure 18 in appendix A. However, the Gantt chart in Figure 19 in appendix B adds the dimension of time to it. It shows when every main activity of the project should start, including an estimation of how long the activity will take.

In total, the project may take 21 weeks from start until completion. The activities have been divided into seven main categories. Starting off with project preparation. This will cover the first four weeks and include arranging practical matters, research on XAI interpretability, creating a research proposal and the kick-off meeting. The next five activity categories are very closely linked to each of the five sub-questions. In total, the completion of the five sub-questions will start in week 4 and end in week 17, therefore taking a maximum of 14 weeks to complete.

After all sub-questions have been completed, the project needs to be finalized. This will include the green light meeting, officially submitting the thesis and the thesis defence meeting. Lastly, one activity will happen throughout the entire project duration. Namely, documentation of all findings and thereby iteratively writing the thesis.

Chapter 3. Background

The core of this chapter will concern an extensive background review on the area of research: combining AI with explanations.

3.1. AI defined

A general and high-level definition of Artificial Intelligence is that AI leverages computers and machines to simulate how the human mind makes decisions and solves problems (IBM, 2020). However, there is not just one definition of AI: there are numerous equally correct definitions of AI. Russell and Norvig incorporated eight different definitions of AI into a two-by-two matrix dividing them into thinking or acting, and doing that humanly or rationally, as shown in Table 2 (Russell & Norvig, 2022). This matrix allows us to define every single definition clearly and systematically.

	Humanly	Rationally
Thinking	"The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak & McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting	"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1992) "The study of how to make computers do things at which, at the moment, people are better." (Rich & Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole et al., 1998) "AI ...is concerned with intelligent behaviour in artifacts." (Nilsson, 1998)

Table 2. Some definitions of AI organized into four categories (Russell & Norvig, 2022)

For example, the definitions of acting humanly comply perfectly with the Turing Test. The Turing test is a measure of a machine's ability to exhibit intelligent behaviour that is indistinguishable from a human (Turing, 1950). It was proposed as a way to determine if a machine can truly demonstrate human-like intelligence. The test involves a human evaluator having a conversation with both a human and a machine, without knowing which is which, and then deciding which is the human. If the evaluator is unable to consistently distinguish the machine from the human, the machine is said to have passed the Turing test. The Turing test remains a relevant and widely discussed concept in the field of artificial intelligence and its ethical implications.

The field of AI, in its most basic form, integrates computer science and (big) data to facilitate problem-solving. Artificially intelligent algorithms can be used to build expert systems that make predictions or classifications based on data as input. Sub-fields of AI include machine learning and deep learning, which are both continuously growing in terms of real-life applications.

3.2. XAI classification

To better understand XAI, it is important to agree on a certain classification. A commonly accepted taxonomy (Linardatos et al., 2020; Shevskaya, 2021) classifies XAI using four different areas. This classification is visualized in Figure 9.

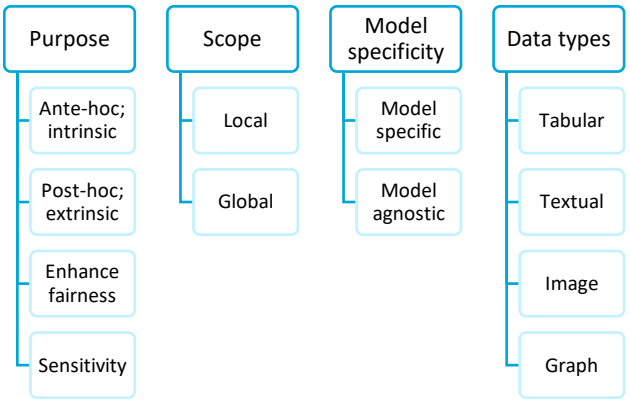

Figure 9. Classification of Explainable AI

Firstly, the purpose of the XAI model. The two most relevant purposes are ante-hoc (or intrinsic) and post-hoc (or extrinsic). An ante-hoc model was developed simultaneously with the algorithm. It concerns a model that was built with the purpose of being explainable. However, a post-hoc model was developed to create an explanation for a black box model. Figure 1 in the introductory chapter visualizes post-hoc XAI. As this research will be conducted from the perspective of post-hoc XAI methods, a brief introduction to the three most well-known post-hoc XAI methods will be provided in this section. It should be noted that there are numerous other methods, such as DTD, DeepLIFT or RISE. However, since LIME, SHAP and LRP are most commonly used in a great variety of applications and domains, only these three will be briefly discussed. This decision does not implicate the research to great extent, however, the perspective in mind enhances replicability.

3.2.1. LIME (Local Interpretable Model-Agnostic Explanations)

LIME is a popular post-hoc XAI method that provides explanations for the predictions made by any machine learning model. The idea behind LIME is to approximate the decision boundary of a black-box model in a small, local region around a particular instance, and then use this approximation to explain the prediction made by the model for that instance. The key advantage of LIME is its model-agnostic approach, which means that it can be used to provide explanations for the predictions made by any machine learning model, regardless of its architecture or the type of data it is trained on. This makes LIME a flexible and widely applicable XAI method. To create a local approximation of a black-box model, LIME perturbs the features of a particular instance and measures the effect of these perturbations on the model's prediction. The explanation generated by LIME is based on the features that have the greatest impact on the prediction, as measured by the magnitude of the perturbations. LIME has been widely used in a variety of applications, including image classification, natural language processing, and predictive analytics. It is especially useful for providing explanations for complex models, such as deep neural networks, that are difficult to interpret directly (Ribeiro et al., 2016).

3.2.2. SHAP (Shapley Additive Explanations)

SHAP is also a post-hoc XAI method that provides explanations for the predictions made by any machine learning model. The key idea behind SHAP is to use the concept of Shapley values from cooperative game theory to explain the contribution of each feature to a prediction. SHAP values measure the contribution of each feature to a prediction, considering the interactions between features. Unlike other XAI methods, such as LIME, which provide local explanations based on perturbations of a single instance, SHAP values provide global explanations that are consistent across all instances in the dataset. SHAP values can be computed efficiently for any machine learning model, regardless of its architecture or the type of data it is trained on. They provide a unified and interpretable way to explain the predictions made by a model and have been shown to be more accurate and consistent than other XAI methods in a variety of settings. In addition, SHAP values have the attractive property of being

consistent with the model's predictions, meaning that the sum of the SHAP values for a particular prediction is equal to the prediction itself. This makes SHAP a particularly useful XAI method for model interpretation, as it provides a unified and interpretable way to understand the impact of each feature on a prediction (Lundberg & Lee, 2017).

3.2.3. LRP (Layer-wise Relevance Propagation)

LRP is a post-hoc XAI method for explaining the predictions made by neural networks. The idea behind LRP is to assign a relevance score to each feature in a neural network, based on its contribution to the prediction for a particular instance. LRP works by propagating the relevance scores backwards through the layers of a neural network, from the output layer to the input layer. The relevance scores are computed based on the activations and weights of the neurons in each layer and take into account the interactions between features. LRP has several desirable properties, such as consistency with the model's predictions, additivity of relevance scores, and the ability to provide both global and local explanations. These properties make LRP a useful tool for understanding the decision-making process of neural networks, and for identifying the most important features in a prediction. LRP has been applied in a variety of domains, including image classification, natural language processing, and predictive analytics. It has been shown to provide meaningful and interpretable explanations for neural networks and has been used to improve the transparency and accountability of AI systems (Montavon et al., 2019).

Besides the purpose of the XAI model, it is relevant to consider whether the XAI explains the entire model (global), or if it explains individual predictions of the model (local). This is referred to as the scope. Thirdly, some XAI systems are applicable to all model types. This can be considered as general software that understands every AI system to a certain extent. This kind of XAI is referred to as model agnostic. On the other hand, there are XAI systems that are specifically designed for, and can only be applied to, a specific model type. These systems are considered to be more detailed. Lastly, the data type which shows the explanation is a relevant factor in classifying XAI. The four most commonly used explanation types are tables, text, images or graphs. As discussed in section 1.6, the scope of this research is limited to textual explanations.

3.3. Explanations explained

Explanatory questions can (logically) be answered through explanations. Pearl and Mackenzie have presented a model for explanatory questions based on their Ladder of Causation (Pearl & Mackenzie, 2018). This model divides explanatory questions into three categories:

1. What-questions, such as "What event occurred?"
2. How-questions, such as "How did the event occur?"
3. Why-questions, such as "Why did the event occur?"

From a reasoning standpoint, why-questions are the most complex as they require the most advanced reasoning skills. What-questions only ask for factual information, potentially relying on associative reasoning to determine other events that occurred based on the observed events. How-questions also request factual information but require interventionist reasoning to identify the set of causes that need to be removed to prevent the event from happening. This could also involve associative reasoning. What if-questions are categorized in the same manner as how-questions since they are simply analysing what would happen under different circumstances. Why-questions are the most difficult as they require counterfactual reasoning to undo events and simulate other non-factual events. Therefore, both associative and interventionist reasoning are required for these questions (Miller, 2019). Given the why-question, an explanation is more specifically defined in twofold. First, there is the explanans: the answer to the question. Second, the explanandum or presupposition: the fact that is being referred to in the question that is asked (Overton, 2012).

For example, consider the question: *'Why am I diagnosed with the flu?'*. This *why*-question requires an explanation for an answer. The answer could be: 'You have the flu, because of the combination of the following symptoms: sudden fever, cough, headache, and tiredness'. The first part ('you have the flu') is the explanandum or presupposition in this case). The rest of the explanation is what is being referred to as the explanans.

Another similar definition of an explanation is provided by Lewis in 1986. He presents an explanation of an event as the provision of information about the causal history leading up to that event. When explaining, someone (*the explainer*) that is in possession of that information (*the explanatory information*), aims to transmit that information to someone else (*explainee*) (Lewis, 1986).

Overton has also defined a structural model for the most complex explanations in science. In order to understand this model, he defined five categories of properties (objects) that are explainable in science. These five objects are:

1. Theories: principles, or a formal system of principles, that can be used as a building block for models.
2. Models: an abstraction of a theory that represents the relationships between kinds and their qualities/attributes.
3. Kinds: an abstraction of any universal class of entities that supports counterfactual reasoning.
4. Entities: no longer an abstraction, but a concrete particular instantiation of a kind.
5. Data: a statement about an entity.

The relationships between these objects are visualized in Figure 10.

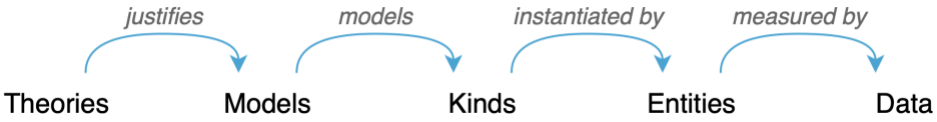

Figure 10. Five categories of scientific explanation (Overton, 2012)

According to Overton, a good explanation of an event at one level must be in relation to, and refer to, at least one other level, and the categories between these two levels must also refer to all intermediate levels. Using these five categories, Overton presents a structure of a theory-data explanation. This type of explanation is arguably the most complex, because the chain of relationships across two levels is the longest since the explainer has to touch upon every intermediary step. This theory-data structure is presented in Figure 11 as an example.

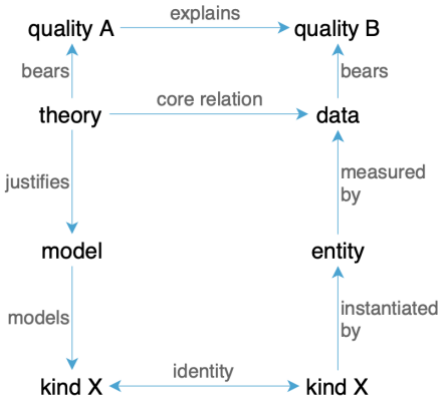

Figure 11. General structure of a theory-data explanation (Overton, 2012)

Chapter 4. Interpretability of XAI for experts

In this report, factors for assessing the interpretability of an explanation will be split up into two categories: objective factors (heuristic-based factors) and human-centric factors (or user-based factors) (Bibal & Frénay, 2016). The first one includes quantitative factors that consist of mathematical entities. The latter one contains studies that evaluated explainability methods with a human as their baseline. These factors involve end-users and exploit their feedback and judgement (Vilone & Longo, 2020).

When measuring human-centric factors, one will rapidly be dependent on interviews and user studies, since that is what can qualitatively 'measure' direct human interactions (Vilone & Longo, 2020). Therefore, a great majority of the available literature aims towards user ratings on various aspects of the explanation, such as usefulness, simplicity or intuitive understandability of the explanation. This is feedback that will be collected after the user's interaction with the explanation. That way, the user can evaluate the given explanation, for example, based on rating several phrases in a survey: "On a scale from one to ten, how much do you agree with the following phrase *'I think that the explanation is very useful'*" (Jesus et al., 2021; Vilone & Longo, 2020). In the following sub-sections, each relevant factor derived from literature will be dealt with separately to ensure overall clarity. This is in accordance with sub-question 1:

"What are the assessment factors that measure to what extent XAI is interpretable to experts?"

4.1. Clarity

Clarity is a key factor that determines the interpretability of XAI (Adadi & Berrada, 2018). Clarity refers to the use of language and presentation that is easy to understand and avoids technical jargon. An explanation is considered clear if it can be understood by a person with a general level of technical knowledge, without the need for specialized training or education. Clear explanations are essential for building trust and understanding between users and AI systems, as well as for ensuring that users are able to understand the decisions made by the AI system and the reasoning behind them. This can help users make informed decisions based on the information provided by the AI system and can also help improve the accuracy and performance of the system over time. To achieve clarity, XAI explanations should be concise and to the point, and should use simple language that is free from technical jargon. It is also important to present the information in a clear and visually appealing format that is easy to understand, such as using simple graphics or images.

Measuring the clarity of an AI explanation can be a challenging task, as it is largely a subjective matter that depends on individual perception and experience. Therefore, clarity is considered a human-centric factor. Consequently, there are several approaches that can be used to assess the clarity of XAI explanations. Some of these approaches include:

1. User surveys: This approach involves conducting surveys or questionnaires to gauge users' understanding and satisfaction with the clarity of XAI explanations. The responses can be analysed to determine the overall level of clarity and to identify areas for improvement.
2. Usability testing: This approach involves having users interact with the XAI system and evaluate the explanations provided based on clarity. The results can be analysed to determine the clarity of the explanations and to identify areas for improvement.
3. Expert review: This approach involves having experts in the field of AI and XAI review the explanations provided by the system and provide feedback on their clarity and effectiveness.

4. Readability metrics: This approach involves using metrics such as the Flesch Reading Ease Score (Flesch, 1948) or Gunning's Fog Index (R. Gunning, 1969) to assess the readability of the XAI explanations. These metrics can provide a numerical score that reflects the level of complexity and ease of understanding of the explanations.

Ultimately, a combination of these approaches may be used to obtain a comprehensive view of the clarity of XAI explanations and to identify areas for improvement. The most important factor is to ensure that the explanations are accessible and understandable to the target audience, regardless of their technical background or expertise.

4.2. Transparency

Transparency is a key factor in determining the interpretability of XAI explanations (Y.-S. Lin et al., 2021). It refers to the degree to which the inner workings of the AI model are visible and understandable to the user. A transparent XAI system provides detailed information about how decisions are made, including the input variables, weights, and algorithms used. The following are some ways in which transparency can be incorporated into XAI systems:

1. Feature attribution: This involves providing information about which input features are most influential in making a decision. This can help users understand how the model is using specific input variables to make predictions.
2. Model visualization: This involves visualizing the structure and parameters of the AI model, such as decision trees, neural networks, or rule sets. This can help users understand the underlying mechanisms and reasoning behind the AI's decisions.
3. Explanation generation methods: This involves using techniques such as saliency maps, counterfactual reasoning, or prototype-based explanations to provide detailed and understandable explanations of AI decisions.

Having a transparent XAI system can help increase users' trust and confidence in the AI's decisions, as well as provide valuable insights into the strengths and limitations of the model. It can also facilitate debugging and improvement of the model over time. However, transparency also has limitations, such as the potential for overloading users with information or providing information that is too complex to understand. It is important to strike a balance between providing enough information to be transparent and not overwhelming users with too much detail.

Measuring the transparency of an AI explanation can be challenging, as it is a subjective concept that can vary depending on the user and the context. Therefore, transparency is considered a human-centric factor. Consequently, there are some metrics and techniques that can be used to quantify the transparency of an XAI explanation:

1. User studies: This involves conducting surveys or user testing to gauge users' perceptions of the transparency of the XAI explanation. This can provide valuable insights into how well the XAI system is meeting users' needs and expectations.
2. Model complexity metrics: This involves measuring the complexity of the XAI model, such as the number of parameters, layers, or nodes (Liao et al., 2020). A more complex model may, to some extent, increase transparency. However, increasing complexity beyond common sense causes less transparency, as it may be more difficult for users to understand the reasoning behind the AI's decisions.
3. Feature importance metrics: This involves measuring the importance of each input feature in making a decision. Features that have a large impact on the decision may be considered more transparent, as they are easier for users to understand.

While these metrics can provide a rough estimate of the transparency of an XAI explanation, it is important to keep in mind that transparency is a subjective concept that can vary depending on the user and the context. It is also important to consider the trade-off between transparency

and model performance, as increasing transparency may come at the cost of reduced accuracy or efficiency.

4.3. Relevance

Relevance is an important factor in determining the interpretability of XAI explanations (Arrieta et al., 2020; Gilpin et al., 2019; Liao et al., 2020; Miller et al., 2017; Samek & Müller, 2019). In the context of XAI, relevance refers to the extent to which the information provided in the explanation is relevant to the user's goals or understanding. A relevant explanation is one that provides the user with the information they need to understand the AI's decision and make informed decisions based on that information. The relevance of an XAI explanation is influenced by a number of factors, including the user's background knowledge, the task context, and the specific decision being made by the AI model. To ensure that an XAI explanation is relevant, it is important to consider the user's needs and goals and to provide information that is tailored to their level of understanding. One way to measure the relevance of an XAI explanation is through user studies. This involves conducting surveys or user testing to gauge users' perceptions of the relevance of the XAI explanation. Another way is to evaluate the extent to which the XAI explanation provides the user with the information they need to understand the AI's decision and make informed decisions based on that information. It is important to note that relevance is a subjective concept that can vary depending on the user and the context. What may be relevant for one user may not be relevant for another. Therefore, it is important to consider the user's needs and goals when evaluating the relevance of an XAI explanation.

Measuring the relevance of an AI explanation can be challenging, again, because of its subjective nature. Therefore, relevance is also considered to be a human-centric factor. Consequently, several methods that can be used to evaluate the relevance of an AI explanation are the following:

1. User studies: User studies can provide valuable insights into the relevance of an AI explanation by gauging users' perceptions of the explanation. This can be done through surveys, user testing, or other methods.
2. Task completion rate: The task completion rate can be used to measure the relevance of an AI explanation by evaluating the extent to which the explanation helps users achieve their goals. For example, if an AI system is designed to help users make informed decisions, the task completion rate can be measured by evaluating the number of users who successfully make a decision using the AI explanation.
3. Information transfer rate: The information transfer rate (ITR) can be used to measure the relevance of an AI explanation by evaluating the extent to which the information provided in the explanation is transferred to the user. This can be done by measuring the mutual information between the AI's predictions, the annotations provided by human labellers, and the explanations provided by the AI system.
4. User satisfaction: User satisfaction can also be used to measure the relevance of an AI explanation by evaluating the extent to which users feel that the explanation meets their needs and provides the information they need to make informed decisions. This can be done through surveys or other methods.

It is important to note that the relevance of an AI explanation can vary depending on the user and the context. For this reason, it is important to consider the user's needs and goals when evaluating the relevance of an AI explanation.

4.4. Trustworthiness

Trustworthiness is another important factor in determining the interpretability of AI explanations. It refers to the degree to which the explanation provided by an AI system can be trusted to be accurate and reliable. The trustworthiness of an AI explanation is dependent on several factors, including the quality of the training data, the robustness of the model, and the

transparency of the decision-making process (Arrieta et al., 2020; Gilpin et al., 2019; Liao et al., 2020; Schmidt & Biessmann, 2019).

The trustworthiness of an AI explanation is also a human-centric factor, as individuals can perceive the trustworthiness differently amongst one another. Hence, trustworthiness can be measured in several ways, including:

1. User studies: Conducting user studies to gauge how much people trust the explanations provided by an AI system.
2. Independent verification and validation: Verifying the accuracy and reliability of an AI system through independent testing and validation.
3. Model performance evaluation: Evaluating the performance of an AI system on a held-out test dataset to assess its ability to generalize and make accurate predictions.
4. Explanation transparency: Providing a clear and transparent explanation of the decision-making process of an AI system, including the features and data used to make a prediction, can also increase trust in its explanations.
5. Model robustness: Evaluating the robustness of an AI system to changes in the input data and to adversarial attacks and demonstrating its resilience to these challenges can also increase trust in its explanations.

In general, measuring the trustworthiness of an AI explanation involves evaluating its accuracy, reliability, and transparency, as well as gauging the perception of trust among users and stakeholders.

4.5. Overlap with human intuition

The overlap of explanations with human intuitions is a significant factor in determining the interpretability of explanations (Lundberg & Lee, 2017). Explanations that align with the expectations, beliefs, and prior knowledge of the person receiving them are more likely to be understood, accepted, and remembered. This is because these explanations resonate with the person's existing mental models and are therefore easier to process and make sense of. However, it can be difficult to scale this concept as it requires task-specific user studies to understand the human intuitions relevant to each individual explanation (Schmidt & Biessmann, 2019). This is because human intuitions are highly context-specific and can vary greatly between individuals and across different domains. To fully understand how to design explanations that overlap with human intuitions, it is often necessary to conduct in-depth studies of the target audience and the specific task or problem being addressed. Despite these challenges, the overlap of explanations with human intuitions is an important factor in determining the interpretability of explanations. By aligning with the expectations and prior knowledge of the person receiving the explanation, an explanation is more likely to be effective and have a greater impact.

Measuring human intuition as a factor of interpretability is also challenging as it is subjective and varies between individuals (therefore also a human-centric factor). One approach is to conduct task-specific user studies, where participants are asked to provide their understanding and interpretation of the explanation. The results of these studies can be analysed to understand how well the explanations align with human intuition and to identify potential areas for improvement. However, this approach can be time-consuming and difficult to scale, as it requires task-specific user studies.

4.6. Intuitive understandability

Intuitive understandability is a significant factor in determining the interpretability of XAI. It refers to how easily users can comprehend the explanations provided by an AI system, without requiring significant cognitive effort. Human intuition is taken into account when assessing the intuitive understandability of an explanation, and an explanation that is easier to understand and aligns with the user's mental models is more likely to be perceived as intuitive and easier

to interpret. Measuring the intuitive understandability of an XAI explanation can be challenging due to its human-centric nature, but user studies or other forms of human-centric qualitative evaluations can be conducted to assess the ease with which users can understand and interpret the explanations. Metrics such as the information transfer rate or recall response time can be used to measure the speed and accuracy with which users reproduce and recall the information in the explanation (see sections 4.7 and 4.8). Overall, providing explanations that are easy to understand and align with the user's natural intuition can improve the transparency, trustworthiness, and overall effectiveness of XAI systems.

4.7. Information Transfer Rate

The information transfer rate (ITR) is a key factor in determining the interpretability of explanations from AI models (Lakkaraju et al., 2016). It refers to the accuracy with which explainees can reproduce the decisions of a machine learning model after receiving an explanation of how the model works (Huysmans et al., 2011). The higher the ITR, the better the interpretability of the explanation. In other words, the interpretability of an explanation is closely tied to the ability of a person to understand, apply, and utilize the information contained within it. If the explanation is clear, concise, and accurately reflects the underlying ML model, it is more likely to be easily interpreted and its information effectively transferred. On the other hand, if the explanation is complex, incomplete, or unclear, it may be more difficult for annotators to reproduce the decisions of the ML model. The information transfer rate is therefore a crucial metric for evaluating the interpretability of explanations, as it measures the effectiveness of an explanation in facilitating accurate and efficient information transfer. The higher the information transfer rate, the more interpretable an explanation is considered to be, and the greater its potential to support human decision-making and understanding of ML models. The information transfer rate can be quantified using bits per second, which is calculated as follows (Schmidt & Biessmann, 2019):

$$ITR = \frac{I(y_{xai}, y_{explainee})}{t}$$

Here, the numerator represents the mutual information between the XAI explanation and the explainee (the annotations made by the explainee – the explainees understanding). The variable t represents the average response time in an explaining task. Since ITR is based on mathematical entities, it is considered to be a more objective factor compared to the previous six.

A concept that is interesting to use with regard to the ITR, is the Shannon entropy. Shannon entropy is a measure of the uncertainty or randomness in a set of data (J. Lin, 1991). It is often used to quantify the amount of information contained in a message or signal. In the context of the ITR, Shannon entropy can be used to determine the ITR by measuring the amount of information that is transferred from the AI system to the user through an explanation. Shannon entropy can be used to calculate mutual information by measuring the degree of uncertainty in both the explanation and the (annotations made by the) explainee. The greater the degree of uncertainty, the greater the amount of mutual information between the annotations and the predictions. This can also be expressed mathematically as:

$$I(A;B) = H(A) - H(A|B)$$

where I(A;B) is the mutual information between the explainee and the XAI, H(A) is the entropy of the explainee, and H(A|B) is the conditional entropy of the explainee given the XAI. By calculating the mutual information between the annotations and the predictions using Shannon entropy, the ITR can be measured and used to evaluate the quality of the explanations provided by the AI system. A higher ITR indicates that more information is being transferred from the AI system to the user through the explanation, while a lower ITR may indicate that the explanation is not effectively conveying the relevant information.

4.8. Recall Response Time

Closely aligned with the ITR is the recall response time (RRT). The RRT is a metric used to quantify the interpretability of an explanation. It refers to the amount of time it takes for an explainee to recall and reproduce the explanation provided by an XAI system. The idea behind recall response time is that the faster an explainee can recall the explanation, the more likely it is that they have understood and internalized the information, which indicates a higher level of interpretability. Recall response time can be measured through user studies, amongst other evaluation methods. In a typical evaluation, participants are provided with an explanation and asked to recall and reproduce the information provided in the explanation. The time it takes for them to recall and reproduce the information is measured and used directly as RRT. Since the time taken can simply be measured by an objective observer, RRT is considered an objective factor.

While recall response time can be a useful metric for evaluating the interpretability of an XAI system, it has limitations. For example, it is important to take into account individual differences in cognitive ability or prior knowledge. Additionally, recall response time may be influenced by factors such as the complexity of the task or the quality of the explanation. Therefore, it is important to interpret the results of recall response time evaluations with caution and to use it in conjunction with other factors for a more comprehensive evaluation of the interpretability of XAI explanations.

Chapter 5. Interpretability of any explanation for laypeople

When aiming to design XAI that is truly able to provide an interpretable explanation to people, it is fair to say that models of humans explaining decisions to other humans is a good way to start the analysis. Literature has proven that links can be made between social science and AI research. For example, Miller has presented models and results in his article that pertain to the behaviour of humans. However, he states that it is reasonably clear that all these models and results have a distinct place in explainable AI (Miller, 2019). Therefore, in this section, sub-question 2 will be addressed:

"What are the assessment factors that measure to what extent any explanation is interpretable to laypeople?"

5.1. Principles from Thagard

Thagard's principles of explanatory coherence were selected as the foundational structure to address sub-question 2 for a number of reasons. Firstly, Thagard is a recognized expert in cognitive science and philosophy, with his work extensively cited and relied upon in the fields of AI and cognitive psychology. His contributions have enriched our understanding of how humans create and evaluate explanations, making his principles highly relevant to this research. Secondly, Thagard's seven principles of explanatory coherence provide a comprehensive framework that aligns closely with the requirements of Explainable AI (XAI). These principles are not only comprehensive, covering various aspects from simplicity to coherence, but they are also robust, having been refined through empirical testing. This makes them uniquely suited for evaluating the interpretability of AI explanations, as they encompass many facets that are key to effective communication of complex ideas. Finally, these principles were chosen because they possess a certain universal applicability. While originally devised to assess human explanations, they can also be extended to the field of AI. They encompass various aspects of human cognition and communication that remain relevant in the evaluation of AI systems' explanations, regardless of the recipients' prior understanding of AI. This aspect makes them particularly well-suited to the task of ensuring that AI explanations are interpretable to laypeople (Thagard, 1989).

Given this reasoning, Thagard's seven principles provide a highly appropriate and effective method to establish the relations of an explanation's coherence, enabling us to assess the global coherence of the explanation. The seven principles are the following (Thagard, 1989):

1. Symmetry: The degree to which the same principles and processes are used to explain a variety of phenomena.
2. Explanation: The degree to which the explanation accounts for the data in a simple, parsimonious and coherent way.
3. Analogy: The degree to which the explanation uses analogies and models to facilitate understanding.
4. Data Priority: The degree to which the explanation gives priority to empirical data over preconceived ideas and beliefs.
5. Contradiction: The degree to which the explanation does not lead to logical contradictions or inconsistencies.
6. Acceptability: The degree to which the explanation is acceptable to relevant experts and is consistent with accepted scientific principles and theories.
7. System Coherence: The degree to which the explanation fits into a larger, more comprehensive system of knowledge and understanding.

For more information on the individual meaning of the principles, please consult Thagard's article. The most important conclusions that can be drawn based on the principles and the core of the article from Thagard are listed in the bullet points below.

- People are more likely to accept a simple and generalizable explanation: not too specific.
- The fewer causes that are being cited in the explanation; the more overview is being created with the explainee. Therefore, simplicity is important.
- The more that is being explained, the more coherent and thus acceptable the explanation is (comprehensive explanation).
- People are more likely to accept an explanation if it is coherent with their prior beliefs.

The fourth bullet point (regarding coherence to prior beliefs) is rather self-explanatory: explainees agree with the explanation because the reasoning behind the explanation corresponds to what the explainee already thought. As a summary of the first three bullet points: people prefer simple generalizable explanations, with fewer causes being presented (instead of complex causes) that explain more events. Several tests were performed on this hypothesis. In one example (Read & Marcus-Newhall, 1993), participants were asked to evaluate explanations about given symptoms. The symptoms that the imaginary patient was presented with, were weight gain, tiredness, and nausea. Multiple explanations were given, as presented in the list below:

- The patient stopped exercising (hence the weight gain), or;
- The patient has mononucleosis (hence the tiredness), or;
- The patient has a stomach virus (hence the nausea), or;
- The patient is pregnant (explaining all three symptoms), or;
- The patient stopped exercising (hence the weight gain), the patient has mononucleosis (hence the tiredness), and the patient has a stomach virus (hence the nausea) at the same time.

In line with the hypothesis presented by Thagard, participants preferred the simplest generalizable explanation that explained the most events whilst naming the fewest causes: the pregnancy in the example above.

Furthermore, this model of seven principles has been proven to align with the core evaluation factors that humans adopt to decide on the value of an explanation (Ranney & Thagard, 1988). Due to the human-centric nature of this model, these factors can consequently be used to determine the interpretability of an explanation.

5.2. Probability

A misconception about using probability in an explanation is that it is beneficial for the explanation. However, people tend to accept the explanation less when the statistical relationship is provided as an explanation (Josephson & Josephson, 1996). Consider for example a container filled with apples of a single variety (for example ‘Pink Lady’ apples). When choosing an apple randomly from the container, it will be a Pink Lady, and one might ask: "Why is this apple a Pink Lady?" The answer that uses the statistical generalization "Because all the apples in the container are all Pink Lady" is not a satisfactory explanation, as it does not explain why that particular apple is of that variety. A better explanation would be that the apple was grown on a Pink Lady farm. However, for the question: "Why did we observe an apple of the same variety being chosen from the container", the statistical generalization is a good explanation, as having only apples of the same variety in the container does result in one being chosen (Josephson & Josephson, 1996).

5.3. Model fidelity

The likelihood of an explanation being true is considered an important factor for determining the quality of an explanation: therefore, model fidelity will be included. Furthermore, explanations are more interpretable when they accurately reflect reality and can be verified through empirical evidence. Truthful explanations are in general more credible and trustworthy,

and thus more likely to be accepted and understood by those who receive them. In contrast, explanations that are false or misleading can lead to misunderstandings and misinterpretations and may ultimately harm credibility and trust in the source of the explanation. It is important to note that the concept of truth can be complex and multi-faceted. There can be different degrees of truthfulness and multiple perspectives on what constitutes truth in each context. In some cases, what is considered true may be a matter of personal belief or interpretation, while in others it may be based on objective evidence or consensus. Additionally, according to Hilton (1996), the most likely or "true" cause is not always the best explanation. Truthfulness is necessary, but on its own not enough for a good explanation. The truth or probability of a cause is only one aspect of a good explanation, and it is incorrect to assume that the most probable cause is always the best.

Despite these complexities, the importance of truth as a factor of interpretability for explanations cannot be overstated. Whether an explanation is perceived as interpretable often depends on how closely it aligns with the truth, and how effectively it can be verified and supported by evidence. When it comes to understanding and accepting explanations, truth is a fundamental and indispensable aspect.

5.4. Abnormality

The concept of abnormality can play a significant role in determining the interpretability of explanations. Abnormality refers to a deviation from what is considered typical, normal or expected in a given situation. When explaining a phenomenon or event, it is often easier to understand and interpret an explanation that takes into account the presence of abnormalities or deviations from the norm. This is because abnormality can often serve as a marker for why something has occurred differently from what would be expected and can provide a more compelling or illuminating explanation. For example, when trying to explain why a particular person is experiencing health problems, it is often more helpful to identify any abnormalities or deviations from typical health patterns, such as the presence of a specific condition or lifestyle choices, than to simply provide a general explanation. The concept of abnormality can therefore play a key role in shaping our understanding and interpretation of explanations (Miller, 2019).

5.5. Intentionality (and functionality)

A concept regarding interpretability of explanations that is often mentioned together with abnormality, is intentionality (Hart & Honoré, 1985). Intentionality is a concept that refers to the deliberate or purposeful nature of an action or event. It is an important factor when it comes to explaining causal chains, as intentional actions are often seen as more significant than non-intentional actions or natural events in determining the causes of a particular outcome. This is because intentional actions are typically seen as having more agency and impact than other types of events.

Studies found that intentional action takes priority over non-intentional action in opportunity chains (Hilton et al., 2005). The authors note that there are two important contrasts in explanation selection: normal vs. abnormal and intentional vs. non-intentional. Causes will be "traced through" a proximal abnormal condition if there is a more distal event that is intentional. Think of the following example: suppose a car accident occurs and a pedestrian is injured. If it is discovered that the driver was intoxicated at the time of the accident, the intentionality of their decision to drive while impaired would receive priority in the explanation over other factors, such as the weather or road conditions (unintentional). The intentional decision to drive while impaired is seen as more significant in causing the accident and the resulting injuries than other possible contributing factors may have been.

In their experiments, Hilton et al. gave participants different opportunity chains in which a proximal abnormal cause was an intentional human action, an unintentional human action, or a natural event, depending on the condition to which they were assigned. Participants were

asked to rate the explanations, and the results showed that intentional action was rated as a better explanation than the other two causes, and non-intentional action was seen as better than natural cases. In addition, the study found that there is little preference for proximal over distal events if two events are of the same type (Hilton et al., 2005).

It is further argued that this holds for functional explanations in general, as opposed to just intentional action (Lombrozo, 2010). For instance, citing the functional reason that an object exists is preferred to mechanistic explanations. Overall, intentionality is an important concept when it comes to explaining causal chains and determining the significance of different events in those chains.

Chapter 6. Creation of framework

In the previous two chapters, the most important assessment factors have been laid out considering the first two sub-questions. In this chapter, the framework will be put together and explained accordingly, in line with the third sub-question:

"What are the requirements for the framework prototype as concluded from comparing the results from sub-question 1 and sub-question 2 regarding the main research question?"

First of all, the results from sub-question 1 and 2 (from chapters 4 and 5 respectively) will be analysed and conclusions will be drawn in section 6.1. Furthermore, specific principles and requirements for the framework will be set up. In section 6.2, the first framework prototype will be created. When revisiting the evaluation process progress figure introduced in section 2.2

6.1. Conclusion on sub-question 1 and 2

The assessment factors that were laid out in chapters 4 and 5 are summarized in Table 3.

XAI interpretability experts	Explanation interpretability laypeople
Clarity	Simplicity
Transparency	Generalizability
Relevance to user's goals	Number of causes
Trustworthiness	Comprehensiveness
Overlap with human understanding	Coherence with prior beliefs
Intuitive understandability	Probability
Information transfer rate (ITR)	Model fidelity
Recall response time (RRT)	Abnormality
	Intentionality

Table 3. Summarized assessment factors

The interpretability of XAI is a crucial aspect that needs to be taken into account, especially when it comes to the interpretation of results by different stakeholders such as experts and laypeople. In order to achieve a comprehensive and thorough understanding of XAI, it is necessary to combine the factors that determine interpretability for experts and laypeople.

The important factors for XAI interpretability to experts emphasize the importance of XAI's ability to communicate its findings in a clear and transparent manner that aligns with the user's goals and can be trusted to deliver accurate and reliable results. Additionally, the overlap with human understanding is critical, as XAI must be intuitive and familiar to users, so they can understand and interpret the results quickly and easily. ITR and RRT are also crucial, as XAI must be capable of providing relevant and useful information quickly, which can be recalled accurately and efficiently.

The important factors for any explanation's interpretability to laypeople are critical in making XAI accessible and understandable to laypeople. Simplicity and generalizability are important for making XAI findings easy to understand and applicable to a variety of situations. A low number of causes are important for providing a clear and concise explanation, while comprehensiveness is crucial for providing a thorough understanding of the findings. Coherence with prior beliefs ensures that XAI's explanations align with existing knowledge and beliefs. Probability and model fidelity are important in providing a factual and reliable explanation, while abnormality ensures that XAI's findings are relevant and useful. Intentionality is essential in ensuring that XAI's findings are in line with the user's goals and needs.

Several common themes and patterns can be identified from the two columns in Table 3. For example, both lists highlight the importance of clarity and transparency in XAI interpretability. The interpretability for laypeople list of factors does not specifically mention clarity and transparency as factors for interpretation by laypeople. However, that list does highlight the importance of simplicity and comprehensiveness, which are related to the idea of clarity and transparency. Simplicity is important for laypeople because it makes the explanation easier to understand and interpret. In the same way, comprehensiveness is important because it ensures that all relevant information is included in the explanation, which can aid in understanding. Therefore, while the second list of factors does not specifically mention clarity and transparency, these concepts are indirectly represented through the emphasis on simplicity, comprehensiveness, and coherence with prior beliefs. This suggests that explanations that are simple, comprehensive, and coherent can be considered clear and transparent, making them more interpretable for laypeople. This suggests that XAI systems need to provide clear and transparent explanations that can be easily understood by both experts and laypeople.

Another common theme is the importance of relevance to the user's goals. The right-hand list of factors in Table 3 does not specifically mention relevance as a factor for laypeople. However, the factors of generalizability and abnormality in the second list are related to the idea of relevance. Generalizability is important for laypeople because it makes the explanation applicable to a wide range of situations, indicating that it has broader relevance beyond just the specific context or case at hand. Abnormality, on the other hand, is important because it helps laypeople identify the most important or interesting aspects of the explanation that may be relevant to their needs and interests. While the second list of factors does not specifically mention relevance and usefulness, these concepts are indirectly represented through the emphasis on generalizability and abnormality. This suggests that explanations that are generalizable and highlight what is abnormal can be considered relevant and useful, making them more interpretable for laypeople. Therefore, both lists do prioritize the importance of providing relevant and useful information to the user, albeit in slightly different ways. This highlights the need for XAI systems to consider the unique needs and perspectives of different user groups when designing and presenting their explanations. This suggests that XAI systems should be designed to provide information that is tailored to the specific needs and goals of the user.

Trustworthiness is also highlighted in both lists as an important factor for interpretability. For instance, the factor model fidelity is closely related to trustworthiness because it ensures that the information being provided is accurate and reliable. Similarly, the factor of coherence with prior beliefs can be seen as important for trustworthiness because it ensures that the new information being provided is consistent with what the user already knows or believes to be true. This indicates that XAI systems need to be accurate and reliable, and that users need to be able to trust the information that is provided to them. Additionally, both lists highlight the importance of coherence with prior beliefs, which suggests that XAI systems need to take the user's existing knowledge and beliefs into consideration.

Finally, both lists prioritize the importance of providing comprehensive explanations that cover all relevant factors. For example, in the list of factors for XAI interpretability for experts, the intuitive understandability is directly influenced by the comprehensiveness of the explanation These factors suggest that the explanation needs to cover all relevant factors to ensure that the user can apply the results to their specific goals and objectives. Therefore, both lists indirectly emphasize the importance of comprehensive explanations. This highlights the need for XAI systems to consider the amount of information that is being provided and ensure that all relevant information is included in the explanation. This suggests that XAI systems need to provide detailed and complete explanations that address all relevant aspects of the problem or task at hand.

Overall, several common themes and patterns have emerged from the two lists. By considering these themes and patterns, a more comprehensive framework can be developed specifically for assessing XAI interpretability that takes into account the needs of laypeople. Consequently, XAI can be designed and presented in a way that makes XAI accessible, understandable, and useful to a wider range of users.

6.2. First framework prototype

The first prototype of the framework is a crucial component of this research and encapsulates the assessment factors presented in Table 3. To provide a detailed description of its development, this section will introduce the framework and shed light on the underlying methodology that led to the framework as it is. The development of the framework followed a robust and systematic process designed to ensure the inclusion and consideration of all the essential elements of interpretability as evidenced by the analysis of sub-question 1 and 2.

6.2.1. Incorporating factors

Having considered all of the factors from section 6.1, the first prototype of the framework has been developed. The decision has been made to include all factors presented in Table 3, as at this point, it is believed that not only do all the XAI interpretability factors on experts also apply to laypeople, furthermore, the general explanation interpretability factors on laypeople apply to XAI explanations on laypeople as well. Combining these two groups of assessment factors provides a much broader overview of what interpretability is and how it can be assessed for laypeople. Furthermore, including all factors will make sure that the framework almost presents a 'weighted average' including the relevant factors of the two groups of assessment factors. Overall, this will give a good overview of XAI interpretability factors for laypeople.

6.2.2. Factor examination and analysis

Each individual factor was meticulously examined to discern its implications and roles in enhancing interpretability. The literature reviews from chapters 4 and 5 were used to explore each factor's theoretical underpinnings, followed by brainstorming sessions to refine the understanding and context of these factors further.

6.2.3. Thematic grouping and prioritization

To structure the framework logically and cohesively, factors were placed close to other factors based on their thematic areas. For example, factors such as clarity and simplicity, which both addressed the ease of understanding, were placed in proximity.

6.2.4. Identifying overlaps and bridging gaps

A critical part of the methodology was recognizing the overlaps and potential gaps between the two groups of factors. A holistic approach was adopted to ensure that no aspect of interpretability was overlooked, and the gaps were adequately bridged, ensuring a comprehensive understanding of interpretability.

6.2.5. Iterative Refinement

After the initial development, the framework underwent several iterations of refinement to improve its clarity, flow, and effectiveness. Feedback was taken into consideration, and necessary adjustments were made to enhance the utility of the framework.

The framework, as illustrated in Figure 12, is a culmination of this methodology. It has been designed to include all factors from Table 4, thereby offering a comprehensive perspective on XAI interpretability for laypeople. In this regard, it serves as a valuable tool for assessing interpretability in XAI, laying the groundwork for an accessible, understandable, and user-centric approach to XAI interpretation.

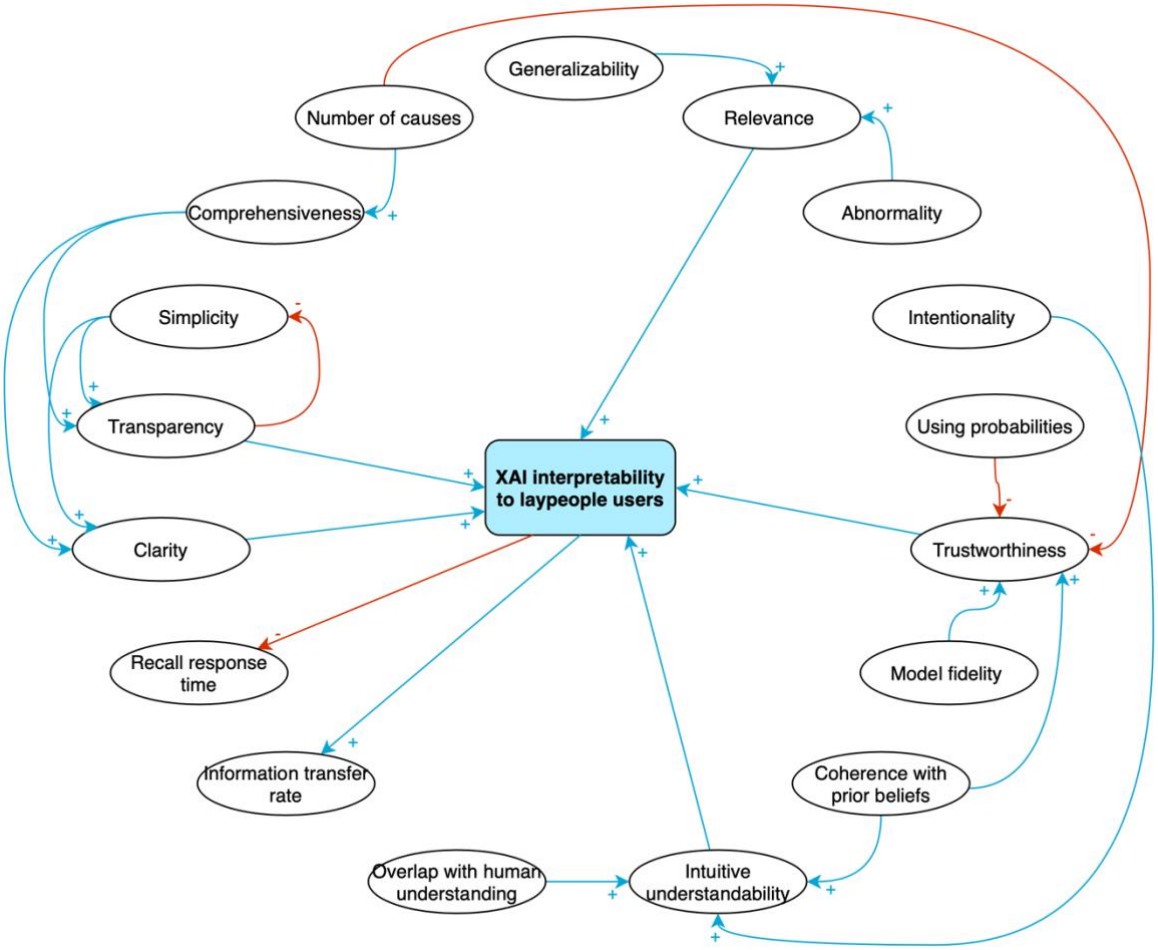

Figure 12. First framework prototype. Red arrows represent negative relationships whereas blue arrows represent positive relationships.

The first thing that should be mentioned is that a total of 17 factors are included in this framework prototype. However, only 7 out of the 17 factors are believed to directly influence (or be influenced by) XAI interpretability to laypeople users. These direct factors include intuitive understandability, trustworthiness, relevance, transparency, clarity, recall response time (RRT), and information transfer rate (ITR). The last two factors are rather stand-alone. In fact, these two factors are included as actual measurements of XAI interpretability. Since a model that is more interpretable causes for a lower RRT and higher ITR. In contrast to the RRT and ITR, the other five direct factors are tightly interlinked with the following ten remaining factors: simplicity, generalizability, number of causes, comprehensiveness, overlap with human understanding, coherence with prior beliefs, using probabilities, model fidelity, abnormality, and intentionality. Let's explore all of the variables and their relationships in more detail:

- Abnormality: This variable refers to the degree to which the model's output is unusual or deviates from expected norms. It positively influences relevance. This means that a model's output that deviates from expected norms may be more relevant to the user's needs or goals.
- Clarity: This variable refers to the quality of being easy to understand or free from confusion. It is positively influenced by simplicity and comprehensiveness, and it positively influences XAI interpretability to laypeople users. This means that the clearer and easier to understand the model's decision-making process is, the more easily it can be interpreted by laypeople users.
- Coherence with prior beliefs: This variable refers to the degree to which a model's decision-making process aligns with the user's existing beliefs or assumptions. It is

positively influenced by intuitive understandability and trustworthiness. This means that a model that aligns with the user's existing beliefs or assumptions is more likely to be easily understood and trusted.

- Comprehensiveness: This variable contains the degree to which a model's decision explains what needs to be explained. It is positively influenced by the number of causes, and it positively influences transparency and clarity. This means that a more comprehensive model, with more factors that contribute to its output, may be easier to understand and interpret by laypeople users.
- Generalizability: This variable refers to the degree to which the explanation can be applied to new, unseen instances. A very generalizable model is not too specific. It positively influences relevance. This means that the more generalizable a model is, the more relevant its output will be to the user's needs or goals.
- Information transfer rate (ITR): This variable refers to the amount of information that is successfully transferred from the model to the user. It is positively influenced by XAI interpretability to laypeople users. This means that the more easily the model can be interpreted by laypeople users, the more information can successfully be transferred to the user. This can be seen as an actual metric of interpretability.
- Intentionality: Intentionality, as it relates to interpretability of explanations, is a concept that refers to the deliberate or purposeful nature of an action or event. It is a critical factor in explaining causal chains, as intentional actions are often seen as more important than non-intentional actions or natural events in determining the causes of a particular outcome. Intentional actions are typically perceived to have more agency and impact than other types of events. It positively influences intuitive understandability, because intentional actions are often seen as more significant in determining the causes of a particular outcome compared to non-intentional actions or natural events.
- Intuitive understandability: This variable refers to the degree to which a model's decision-making process can be easily understood without the need for additional explanation or knowledge. It is positively influenced by overlap with human understanding, intentionality, and coherence with prior beliefs, and it positively influences XAI interpretability to laypeople users. This means that a model that is easily understood without the need for additional explanation or knowledge is more likely to be easily interpreted by laypeople users.
- Model fidelity: This variable refers to the degree to which a model's output aligns with objective reality. It positively influences XAI interpretability to laypeople users. This means that the more closely the model's output aligns with objective reality, the more easily it can be interpreted by laypeople users.
- Number of causes: This variable refers to the number of factors that are being used in the explanation to support the model outcome. It positively influences comprehensiveness, but it negatively influences trustworthiness. This means that while a model with more causes may be more comprehensive, it may be less trustworthy to users.
- Overlap with human understanding: This variable refers to the degree to which a model's decision-making process, or even the final outcome, aligns with human intuition or knowledge. It is positively influenced by intuitive understandability. This means that a model that aligns with human intuition or knowledge is more likely to be easily understood by laypeople users.
- Recall response time (RRT): This variable refers to the time it takes for the user to recall relevant information needed to understand the model's output. It is negatively influenced by XAI interpretability to laypeople users. This means that the longer it takes for the user to recall relevant information needed to understand the model's output, the less easily it can be interpreted by laypeople users.
- Relevance: This variable refers to the degree to which the model's output is applicable to the user's goals or needs. It is positively influenced by abnormality and generalizability, and it positively influences XAI interpretability to laypeople users. This

means that the more relevant a model's output is to the user's needs or goals, the more easily it can be interpreted by laypeople users.

- Simplicity: This variable refers to the quality of being easy to understand or do. It is negatively influenced by transparency, but it positively influences transparency and clarity. This means that while a simpler model may be easier to understand, it may be less transparent and more opaque to users.
- Transparency: This variable refers to the degree to which a machine learning model's decision-making process can be understood, and its logic can be traced. It is positively influenced by simplicity and comprehensiveness, and it positively influences XAI interpretability to laypeople users, but it negatively influences simplicity. This means that the more transparent a model's decision-making process is, the easier it is for laypeople users to interpret, but this may come at the expense of simplicity. In conclusion, there is an interesting loop between simplicity and transparency: the simpler a model is, the more transparent it is most likely. However, increasing transparency in a model means having to explain more, therefore, simplicity will be at stake.
- Trustworthiness: This variable refers to the degree to which a model's output can be trusted. It is positively influenced by model fidelity and coherence with prior beliefs, but it is negatively influenced by using probabilities and number of causes, and it positively influences XAI interpretability to laypeople users. This means that a model that is faithful to its training data and aligns with the user's prior beliefs is more likely to be trusted, while using probabilities and having a large number of causes may make the model appear less trustworthy.
- Using probabilities: This variable refers to the degree to which a model uses probabilities to support their outcome. It negatively influences trustworthiness. This means that a probability-based outcome may appear less trustworthy to users.

6.3. Key principles of the first framework prototype

The key principles and takeaways from analysing the above framework are the following:

1. Simplicity and comprehensiveness must be balanced. A model that is too simple may be less transparent and more opaque to users, while a model that is too complex may be more difficult to understand. Therefore, it's important to strike a balance between simplicity and comprehensiveness to achieve optimal XAI interpretability.
2. Clarity and transparency are crucial for XAI interpretability to laypeople users. A model that is easy to understand, free from confusion, and has a clear and transparent decision-making process is more likely to be easily interpreted by laypeople users.
3. Trustworthiness is key to XAI interpretability. A model that is faithful to its training data aligns with the user's prior beliefs, and does not rely too heavily on probabilities or a large number of causes is more likely to be trusted and therefore more easily interpreted by laypeople users.
4. Relevance and alignment with user goals are critical for XAI interpretability. A model that is relevant to the user's needs or goals and aligns with their existing beliefs or assumptions is more likely to be easily understood and trusted.
5. Intuitive understandability is essential for XAI interpretability. A model that aligns with human intuition or knowledge, appears deliberate or intentional, and can be easily understood without the need for additional explanation or knowledge is more likely to be easily interpreted by laypeople users.
6. Recall response time and information transfer rate are important metrics for XAI interpretability. The faster the model can transfer information to the user and the less time it takes for the user to recall relevant information needed to understand the model's output, the higher the interpretability is.
7. Finally, abnormality can be a positive factor for XAI interpretability. A model's output that deviates from expected norms may be more relevant to the user's needs or goals and therefore more easily interpreted by laypeople users.

In summary, the key principles from the framework suggest that XAI interpretability for laypeople users can be improved by focusing on clarity, transparency, simplicity, comprehensiveness, trustworthiness, relevance, alignment with user goals, intuitive understandability, and abnormality. By considering these factors, researchers and developers can improve the interpretability of their machine learning models for laypeople users, leading to more effective and trustworthy decision-making processes.

Chapter 7. Evaluation of framework part I

In this chapter, the effectiveness of the interpretability framework developed in chapter 6 will be evaluated by conducting the first of two separate evaluations. Therefore, XAI developers and experts will evaluate the core components of the framework and their interconnectivity. This is aligned with the fourth sub-question of this thesis:

"How is the first framework prototype evaluated by experts and XAI developers in order to determine its strengths, limitations, and potential areas for improvement?"

Thereafter, the second evaluation will use the framework to assess the interpretability of an XAI system with laypeople through a use case-based survey, in order to test its effectiveness in real-world applications. This second evaluation is shown in chapter 8. This is based on Figure 8 in the methodology chapter.

By conducting these two evaluations, the aim is to assess the effectiveness of the interpretability framework in two key areas. Expected is that the feedback collected from XAI experts/developers and laypeople will help in improving the framework, making it more effective in guiding the design and development of XAI systems and more accessible to laypeople. Ultimately, it is believed that this framework will help promote transparency, trust, and understanding in the development and application of XAI systems.

7.1. Nature of first evaluation round

To evaluate the framework's effectiveness for XAI developers, a series of interviews with XAI experts and developers was conducted to gauge their opinions on the framework's usefulness, ease of use, and applicability to their work. Participants were presented with the framework, including the target audience, goals of the XAI system, key factors and key principles. They were asked to provide feedback on the framework and to suggest any areas for improvement, based on an interview template. The full interview template that was used can be found in appendix D. Since the interviews were both done in English as well as in Dutch, both the English and Dutch template are shown there. The interview can be divided into two parts. It starts with a thorough review of the individual components of the framework. The following five questions are asked:

1. Do you think that the factors included in the framework are relevant to evaluating the interpretability of XAI systems?
2. Are there any factors that you believe should be included in the framework that are currently missing?
3. Are there any factors that you believe should be excluded in the framework that are currently unjustifiably there?
4. Could you please rank the five most important factors for XAI interpretability, from most important to least important?
5. Are there any improvements that you would suggest to the framework to make it better in practice?

The second part of the interview covers the implementation and overall limitations of the framework. This section will be covered through the following four questions:

6. In your experience, are there any challenges or limitations to implementing the framework in practice?
7. Additionally, are there any challenges that might arise when trying to get stakeholders to agree on the interpretability of an XAI system using the framework?
8. How might these challenges be overcome (practically)?

9. How relevant do you see the framework to your current work in XAI and how useful do you estimate the framework to be when developing XAI for non-expert users?

The aggregated feedback collected from XAI experts and developers will be analysed to identify the strengths and weaknesses of the framework, as well as any areas that may require further clarification or refinement. This analysis will be used to improve the framework and make it more effective in guiding the design and development of XAI systems.

7.2. Nature of experts and developers

Whilst maintaining the necessary level of privacy for the interviewees, in this section, some information about the interviewees will be presented. All interviewees are considered highly valuable experts within the XAI field. In total, this interview has been conducted with twelve XAI developers/experts gathered from the following organizations/institutions (presented in alphabetical order):

- Deeploy
- Fraunhofer Fokus
- LMU Munich
- Microsoft
- Erasmus University - RSM
- TNO
- TU Delft
- TU Dublin
- University of Hildesheim
- University of Melbourne

The twelve interviewees are further categorized using Table 5 below. This table first of all distinguishes interviewees between academical and practical/industrial experts regarding XAI. After that, a classification is made based on application domains, and experience level. Please note that the interview numbers correspond to the interviews as presented in appendix E.

Interviewees	Expertise	Domain	Area
Interviewee 1	XAI	Human-AI interaction	Academia & Industry
Interviewee 2			
Interviewee 3			
Interviewee 4		Human-computer interaction	Academia
Interviewee 5			
Interviewee 6			
Interviewee 7		Interactive Intelligence	
Interviewee 8		Organizational XAI	
Interviewee 9		XAI development	Industry
Interviewee 10			
Interviewee 11			
Interviewee 12			

Table 4. Categorization of interviewees based on expertise, domain, and area

7.3. Evaluation of first framework prototype

The initial prototype presented in chapter 6, as shown in Figure 12, was developed with the intention of creating an objective and transparent approach for assessing XAI interpretability for non-experts. The framework was refined through feedback from experts who were consulted regarding their experience with XAI, as well as their opinions on the framework's components and overall structure. The experts' valuable feedback helped identify areas for improvement, which were subsequently incorporated into an updated version of the framework. In this section, the aggregated and anonymized conclusions of the expert

interviews will be presented, including their feedback and suggestions, as well as their thoughts on the framework's implementation and limitations. The feedback from the experts was generally positive, with most explicitly stating that the framework was a useful tool for evaluating XAI interpretability for laypeople. However, some key suggestions were made for improving the framework even further, which were taken into account in the updated version (referred to as framework prototype 2). Framework prototype 2 is presented in section 7.4.

After analysing the twelve interviews, the following aggregated feedback and common themes emerge regarding framework prototype 1. First of all, individual factors will be discussed. Secondly, operationalization and practicality, thirdly, context, fourthly, measurability and fifthly, the level of abstraction will be mentioned. Lastly, specific and isolated comments will be discussed. In summary, there are interesting conclusions regarding the individual factors, which are presented in section 7.3.1. After which there are four major categories of practical conclusions. These four categories are also visually represented in Figure 13. Based on this figure, it can be concluded that the majority of interviewees mentioned operationalisation, context and measurability as practical improvements to the framework.

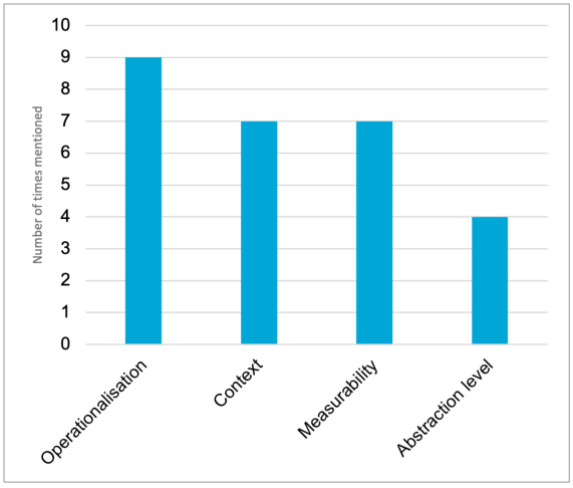

Figure 13. Main categories of practical conclusions from the expert interviews

7.3.1. Individual factors

The first point of discussion for every interview was about the individual factors in the framework. Several experts pointed out that certain factors might be missing from the framework, and they advised revisiting the framework to incorporate these factors. Other points were raised to merge or exclude factors from the framework. These notes are summarized below:

- Overlap with human understanding can be combined with coherence with prior beliefs (interview 2 and 10).
- Intuitive understandability might need to be changed to affordance (interview 4, 6, and 9).
- Actionability and/or fit-for-purpose factors can be added (interview 3, 4, and 8).
- Number of causes can be captured into complexity, which may benefit the level of abstraction (interview 1 and 3).
- Trustworthiness should focus more on correctness and can be split into trustworthiness of the explanation and trust from the user (interview 3).
- Explanation fidelity is not yet in the framework. However, it may be even more important than model fidelity (interview 1).
- Some other factors are very similar, insofar that they are almost the opposite. For example simplicity and comprehensiveness, coherence with prior beliefs and abnormality (interview 8).

- Explanation complexity should be balanced - too simple or too complex explanations may not be convincing (interview 3, 4, 5, and 7).
- Trustworthiness needs to be balanced with transparency (interview 5).
- Using probabilities factor needs more research: causal information is more convincing than probabilistic information (interview 3).
- Abnormality may not positively influence relevance to the user's goals, as it influences the desire for understanding rather than goals/needs (interview 3).
- Transparency may be less relevant for laypeople, as it could overcomplicate the explanation (interview 4).

Furthermore, interviewees were also asked to shed light on the factors that they deemed most important. A commonly received answer was that this is highly context-specific (see section 7.3.3). However, some interesting conclusions can be drawn from the results of these twelve interviews. This is visualized in Figure 14.

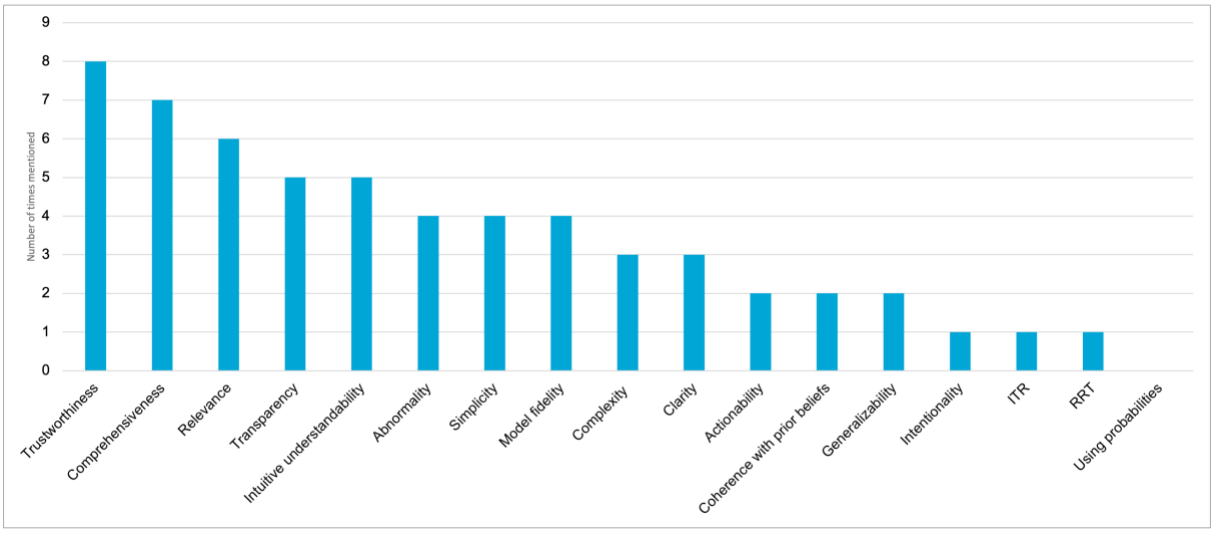

Figure 14. Importance of individual factors based on expert interviews

Based on the graph in Figure 14, the min-max scaling method (or: normalisation method) was applied. Min-max scaling is a simple and common technique for transforming numerical data into a common range, which can be helpful for comparing values or when feeding data into machine learning algorithms that are sensitive to feature scales (Jain et al., 1999).

The decision has been made to scale all factors from 1 (less important) to 5 (most important). The choice of this range was initially inspired by the commonality of a 5-star scale, as it is frequently employed in rating systems and is intuitively understood by a broad audience. Nevertheless, as the evaluation progressed, it became evident that the 1-5 scale, although intuitive, might not be adequately granular to capture the nuanced trade-offs and competing priorities inherent in assessing XAI systems. Therefore, while the scale remained anchored at 1 and 5 for the least and most important factors, respectively, it was adjusted to allow for decimal points. This provided the necessary granularity and flexibility to more accurately represent the relative importance of different factors in the framework. The weights, as determined through this enhanced scaling process, are presented in Table 5.

Factor	Weight
Trustworthiness	5.0
Comprehensiveness	4.5
Relevance	4.0
Transparency	3.5
Intuitive understandability	3.5
Abnormality	3.0
Simplicity	3.0
Model fidelity	3.0
Complexity	2.5
Clarity	2.5
Actionability	2.0
Coherence with prior beliefs	2.0
Generalizability	2.0
Intentionality	1.5
ITR	1.5
RRT	1.5
Using probabilities	1.0

Table 5. Relative importance of individual factors used for weight attribution

7.3.2. Operationalization and practicality

The feedback regarding operationalization and practicality primarily revolves around the need to transform the current abstract, conceptual framework into a more actionable and user-friendly tool that XAI developers and researchers can apply in real-world scenarios. Here's a more comprehensive look at the concerns and suggestions raised by the experts considering this topic of evaluation:

- Actionable guidelines: Several experts suggested that the framework should provide clear, step-by-step guidelines that XAI developers can follow during the design and development process. This would make the framework more practically useful, as developers would have a structured approach to incorporating the key principles and factors into their XAI systems. This suggestion might result in focusing more on the key principles as opposed to the framework itself being the key deliverable of this research (interview 1, 2, 8, 10 and 11).
- Tools and methods: Some experts recommended creating specific tools or methods based on the framework, such as questionnaires, forms, or evaluation metrics. These tools would help developers assess the effectiveness of their explanations, identify areas for improvement, and iteratively refine their XAI systems. Moreover, tools like questionnaires could facilitate communication with stakeholders and help elicit their preferences and requirements regarding XAI interpretability (interview 5, 7, 11).
- A/B testing and evaluation: One expert mentioned the possibility of using A/B testing or other evaluation techniques to compare different XAI models or explanations. By doing so, developers can better understand the impact of specific factors on interpretability and identify the most effective explanations for their target audience (interview 2, 3, 5, and 11).
- Bridging the gap between theory and practice: Experts emphasized the importance of making the framework more practical and accessible for developers with varying levels of expertise in XAI. This could involve providing examples, case studies, or templates that demonstrate how to apply the framework in different contexts or domains (interview 1, 2, 6, 10, 11).
- Tailoring the framework to different stakeholder needs: As various stakeholders have different requirements and expectations, the experts suggested that the framework should be adaptable to address these differences. This might involve creating extended forms, customizable templates, or modular components that can be adjusted according to the specific needs of a given stakeholder group or application (interview 1, 4, 10, 11).

- Tracking success and progress: Experts stressed the need for the framework to include methods for evaluating its own effectiveness, as well as the impact of individual factors on XAI interpretability. By incorporating such evaluation mechanisms, the framework could help developers track their progress and ensure that their XAI systems align with the key principles and desired outcomes (interview 5, 7, 11).
- Categorizing factors: Since the framework is perceived as rather overwhelming, it may be a good idea to categorize some of the factors (interview 8, 10).

In summary, enhancing the operationalization and practicality of the framework involves providing actionable guidelines, developing tools and methods, implementing evaluation techniques, bridging the gap between theory and practice, tailoring the framework to different stakeholder needs, and incorporating mechanisms for measuring success and progress. Addressing these concerns will make the framework more useful and accessible to XAI developers and researchers across a range of contexts and applications.

7.3.3. Context

The feedback related to contextualization highlights the importance of considering the specific context in which an XAI system operates when evaluating interpretability factors. Experts emphasized that context dependence can influence the effectiveness and relevance of the framework and its factors in various ways. Here's a more comprehensive look at the points raised by the experts regarding context dependence:

- Addressing context in the framework: To account for context dependence, experts advised to incorporate context into the framework, either directly or indirectly. This could involve discussing how each factor might be influenced by context or providing examples of how the factor's relevance might change based on the situation. In this way, the framework can remain adaptable and relevant across different contexts. This may be one of the most important takeaways from the entire interview evaluation round. Context is an influential factor for all constructs mentioned in the framework. Therefore, context should be taken into serious account (interview 1, 2, 3, 5, and 8).
- Variable importance of factors: Experts noted that the importance of specific factors might vary depending on the context. For instance, the significance of trustworthiness, model fidelity, or relevance could differ for patients compared to insurance agencies. Similarly, the relevance of factors such as transparency or simplicity might depend on the target audience (e.g., experts or laypeople) (interview 2, 3, 6, and 9).
- Tailoring explanations to stakeholders: Several experts suggested that the framework should take into account the different needs and requirements of various stakeholder groups. By considering the specific expectations and informational needs of diverse stakeholders, developers can create explanations that are more relevant and useful in the given context (interview 3, 4, 8, and 11).
- Application and domain influence: The context dependence of factors can also be affected by the type of application or domain in which the XAI system is being used. For example, decision support systems might require a different set of factors to be considered most important compared to autonomous or robotic systems. Recognizing these distinctions can help developers tailor their explanations to the specific use case (interview 1, 5, and 8).
- Context-specific evaluation: Given the context dependence of factors, experts recommended using context-specific evaluation methods to assess the effectiveness and relevance of explanations. This might involve developing tailored evaluation metrics or questionnaires for different stakeholder groups, applications, or domains to ensure that the framework remains useful and applicable across diverse scenarios (interview 5, 6, 8, and 11).
- Stakeholder engagement in the design process: A few experts emphasized the importance of incorporating stakeholder engagement into the XAI system design

process to address context dependence. By involving stakeholders early on and considering their needs and requirements, developers can create explanations that are more relevant, useful, and acceptable in the specific context (interview 1, 3, 4, and 8).

In summary, addressing context dependence in the framework involves recognizing the variable importance of factors, tailoring explanations to stakeholders, considering application and domain influence, incorporating context variables into the framework, using context-specific evaluation methods, and engaging stakeholders in the design process. By acknowledging and accounting for context dependence, developers can create more effective and meaningful explanations that meet the needs of diverse users and scenarios.

7.3.4. Measuring constructs

The feedback on related to measurability of the constructs primarily focuses on the importance of distinguishing between factors that can be easily measured and those that are more abstract or subjective. Experts expressed concerns about the challenge of quantifying certain factors, and they offered suggestions on how to approach these issues. Here's a more comprehensive look at the points raised by the experts:

- Differentiating between measurable and abstract factors: Experts advised categorizing the factors in the framework based on their measurability. They suggested that some factors, like the number of causes or simplicity (indirectly measured by the number of words in the explanation), could be quantified relatively easily. In contrast, other factors, such as transparency or clarity, might be more challenging to measure due to their subjective nature (interview 2 and 9).
- Developing metrics and evaluation methods: To make the framework more useful and actionable, experts recommended developing specific metrics and also combine them with the evaluation methods for the various factors. These metrics would allow developers and researchers to assess the effectiveness of their XAI systems and explanations, as well as track their progress in improving interpretability (interview 2, 3, 5, 6, 7, 8, 11 and 12).
- Balancing factors: Some experts noted that certain factors may need to be balanced against each other to achieve optimal interpretability. For example, simplicity and comprehensiveness might be in tension, requiring developers to find the right balance between them. Similarly, transparency might be essential for experts but less critical or even counterproductive for laypeople. This can be a very important point of interest when designing the key principles (interview 1, 2, 4, and 11).
- Human-centred interpretability: A few experts highlighted the need for a human-centred approach to measuring factors and constructs. They suggested that some factors might be more relevant to the perceived interpretability of the explanation by users, while others could be more focused on technical aspects. Distinguishing between these types of factors could help developers prioritize their efforts and create explanations that are more meaningful and useful to end users (interview 3, 5, and 8).
- Necessity: Some experts also questioned whether it is really necessary for this research already to comprehensively include the measurability of all factors. This can be left open for other researchers (interview 6 and 9).

In summary, addressing the concerns around measuring factors and constructs involves differentiating between measurable and abstract factors, developing metrics and evaluation methods, balancing factors, accounting for context, considering robustness, and adopting a human-centred approach to interpretability. By refining the framework to account for these concerns, developers and researchers can more effectively assess and improve the interpretability of their XAI systems and explanations.

7.3.5. Level of abstraction

Interviewees 1, 3, 6, and 8 mentioned the need to consider the level of abstraction in the XAI interpretability framework. The reason for this is that the factors included in the framework are not all on the same level of abstraction, and this can make the framework more difficult to use in practice. For example, the experts pointed out that recall response time (RRT) is a clear metric that can be measured, while transparency is a vaguer construct that is difficult to measure. This means that these factors are not on the same level of abstraction, which can create confusion when trying to apply the framework in practice. To address this issue, the experts suggested that the framework could be rephrased or categorized to make it more concrete and practical. This would involve organizing the factors based on their level of abstraction and providing more structure to the framework. By doing so, the framework would be easier to understand and use in practice. Another benefit of considering the level of abstraction is that it could help identify trade-offs between different factors. For example, a factor that is more concrete and measurable may be prioritized over a more abstract factor that is difficult to measure. This could help XAI developers make more informed decisions when designing XAI systems and explanations.

7.3.6. Other interesting comments

Apart from the five main categories of comments, which consistently appeared in almost every interview, several other interesting comments were made. Other than that, interesting contradictions between individual interviewees are also briefly discussed. These are listed below:

- The linguistic perspective of XAI explanations can be considered, and the framework should be shaped to be more of an action plan (interview 4).
- It is important to draw a clear line between experts and laypeople when using the framework (interview 1).
- The key principles and the framework are both important, and they should be used together to interpret the framework.
- Expert interviews are a great starting point for reviewing the framework. However, it is believed that laypeople (the target audience of the framework) should be involved in the evaluation process. This is what will be done in chapter 8 (interview 2)
- For example interviewee 2, 3 and 6 suggested transparency as a key feature of the framework, but interviewee 4 and 11 questioned whether transparency is always relevant, particularly for laypeople who might find it more confusing than helpful.
- Interviewee 7 noted that the current framework is too abstract and conceptual, making it less useful in practice. However, interviewee 9 suggested that even though many aspects of the framework are difficult to measure and appear abstract, this isn't necessarily a problem for the framework.
- Some interviewees, like interviewee 9 and interviewee 11, mentioned the challenge of measuring factors and suggested that not everything needs to be measurable for the framework to be valuable. However, Interview 12 suggested the inclusion of metrics for evaluation purposes, indicating a desire for more concrete measures of success.
- Overall, the experts acknowledged the relevance and usefulness of the framework for both novice and experienced XAI researchers and developers. They noted that it could serve as a starting point for discussions, encourage a more systematic approach to XAI interpretability, and potentially improve the design and development of XAI systems.

7.4. Second framework prototype

Responding to all feedback provided by experts, this framework has been shaped by revisiting certain factors, adopting a more context-aware perspective, pursuing practical and measurable approaches, reconsidering the level of abstraction within the framework, and introducing

weights to the factors. Below, we delve into the comprehensive methodological approach used to construct this enhanced prototype.

The process began with a thorough review and categorization of the expert feedback. Each suggestion, critique, and remark was classified into one of five categories as presented in sections 7.3.1 to 7.3.6. This step allowed for the identification of common themes and areas of improvement. Next, a deep dive into each category was conducted, examining the detailed feedback from the experts, understanding their implications, and strategizing on how best to integrate this feedback into the framework. In essence, this stage allowed for a transition from qualitative feedback to actionable steps.

7.4.1. Revisiting and revising factors

Each factor in the original framework was revisited, using the aggregated feedback as a guiding tool. The purpose of this step was to understand whether the original factors were comprehensive enough, or whether they needed refinement or replacement. Each factor was evaluated in terms of its clarity, relevancy, measurability, and its level of abstraction. In cases where factors needed revision, definitions, their implications on interpretability, and their relationships with other factors were reassessed. This allowed for a refinement of the factors, ensuring they are well-defined, significant, measurable, and compatible with the rest of the framework.

7.4.2. Context-awareness and adaptability

The third stage involved a comprehensive analysis of the framework's ability to adapt to different contexts. Given the feedback emphasizing the importance of context dependence in XAI, we implemented a context-aware approach in the framework. This involved considering how different factors may be influenced by the context, and how their relevance or importance may change depending on the situation. This led to the development of a more flexible and adaptable framework that can be tailored according to different stakeholder needs and applications.

7.4.3. Practicality, measurability, and abstraction levels

Practicality and measurability of the factors, as well as their level of abstraction, were the next crucial areas of focus. We sought to ensure that the framework not only provided theoretical insights but also offered practical, actionable guidelines for XAI developers and researchers. To this end, we differentiated the factors based on their measurability, developing specific metrics for each factor, and ensured they could be balanced against one another when needed. We also reassessed the levels of abstraction of the factors, rephrasing or categorizing them to make the framework more concrete, user-friendly, and useful in practical settings.

7.4.4. Factor weights

Lastly, based on the feedback suggesting the inclusion of weights to factors, we introduced a weight mechanism to the framework. The weights represent the relative importance of each factor in contributing to XAI interpretability, giving developers a clearer view of which factors to prioritize. These weights were determined based on a combination of expert feedback, current literature on XAI interpretability, and our understanding of the framework and its goals. The weights are flexible and can be adjusted based on the specific needs of different contexts and applications.

7.4.5. Presentation of the second framework prototype

After incorporating these extensive methodological steps, we present the revised second framework prototype in Figure 15. This prototype represents an evolved version of the original framework, shaped and refined based on expert feedback, more flexible and adaptable to context, practical and measurable, and includes a refined set of weighted factors. By making these changes, we aim to have created a framework that better aids developers and

researchers in improving XAI interpretability and in turn, fostering trust and understanding among users.

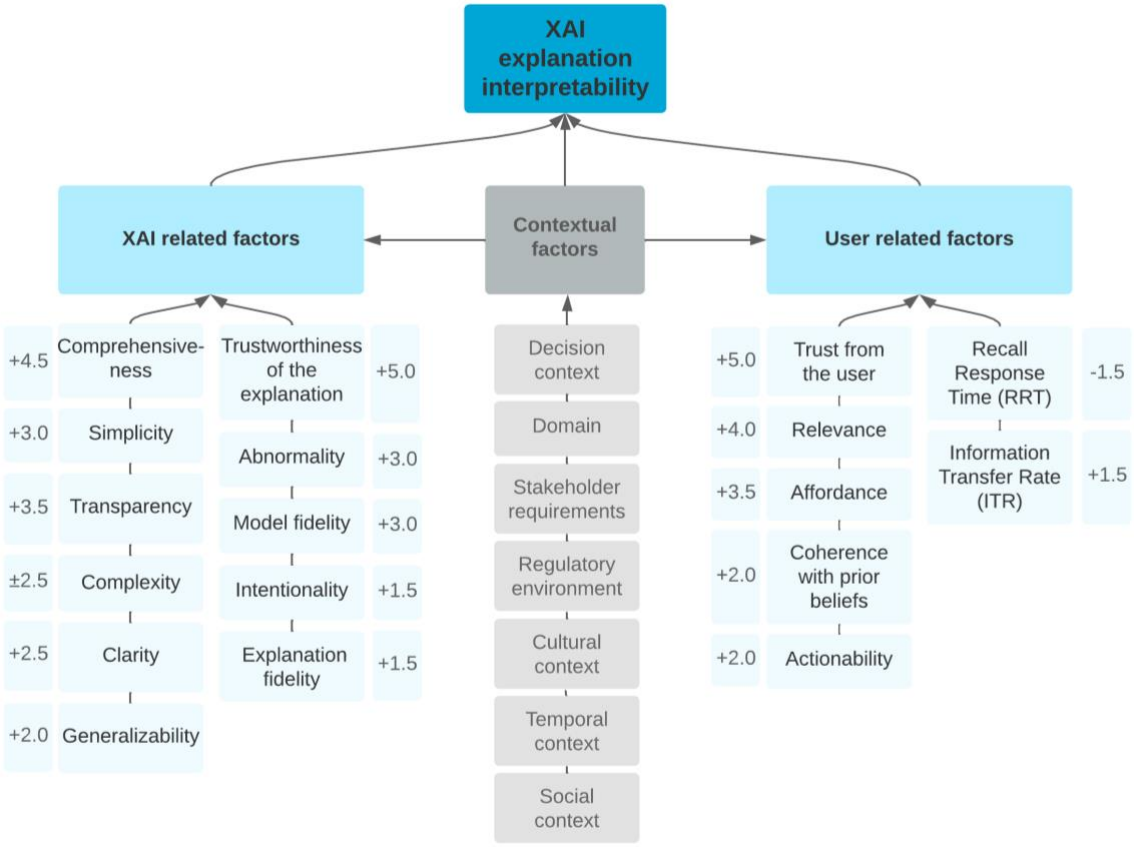

Figure 15. Second framework prototype

Upon reviewing the framework, it's evident that it has undergone a significant restructuring compared to the initial framework prototype. This modification has led to a trade-off, namely a decrease in overall comprehensiveness but a corresponding increase in simplicity. The decision to emphasize simplicity over comprehensiveness in the framework was not made lightly (Todd & Gigerenzer, 2000). The goal of this framework is to be accessible and usable in practical contexts. The previous version, while comprehensive, risked being perceived as overwhelming, especially considering the additional factors and relationships inferred from the interviews incorporated. In practical applications, a framework that is too intricate may lose potential users, reducing its overall effectiveness and accessibility. This is consistent with principles discussed by Berthoz (2012), emphasizing the need for simplicity in managing complex tasks efficiently. Simplicity improves workability and efficiency (Duignan, 2023), making the framework easier to understand and apply. XAI developers can glean insights quickly and efficiently without being overwhelmed by complexity. Additionally, simplicity aids in the acceptance and adoption of the framework by practitioners, who often need quick, straightforward solutions to complex problems. Workability and efficiency are not just a 'nice-to-have' characteristic of the framework; it is a necessity for its successful application (Todd & Gigerenzer, 2000). By boosting its workability and efficiency, the drive for simplicity in the framework indirectly enhances its overall utility. Therefore, it was concluded that the benefits of increasing the framework's simplicity and consequently workability and efficiency, while losing some degree of comprehensiveness, would result in a more effective and applicable tool in practical scenarios.

Besides the overall presentation of the framework, several alterations were made based on the input from the interviews. First of all, context plays a significant role in determining the interpretability of XAI for laypeople. By splitting the factors into three categories: XAI-related

factors, user-related factors, and contextual factors, the importance of context in XAI interpretability is stressed. Here's why and how the categorization was made:

1. XAI-related factors: These factors are specific to the explanation generated by the XAI system. They include aspects like simplicity, transparency, and model fidelity. These factors are crucial because they directly affect how well a user can understand and trust the explanation provided by the system. By categorizing these factors, you acknowledge the technical aspects of XAI interpretability that need to be considered by developers and researchers.
2. User-related factors: These factors are specific to the individual users or laypeople who interact with the XAI system. They include aspects like cognitive abilities, prior knowledge, and trust in the system. These factors are important because they determine how well a user can process and make sense of the provided explanations. By categorizing these factors, you acknowledge that the interpretability of an XAI system is also dependent on the user's unique characteristics.
3. Contextual factors: These factors are specific to the situation or environment in which the XAI system operates. They include aspects like the decision context, the domain of the problem, and the stakeholder requirements. These factors are essential because they shape the relevance and appropriateness of the explanations for different use cases. By categorizing these factors, you acknowledge that the effectiveness of an XAI system in providing interpretable explanations depends on the context in which it is applied.

By creating these categories, practitioners can better account for the varying influences on XAI interpretability and address the complex interplay between the XAI system, the user, and the context. This categorization can help guide researchers and developers in tailoring their explanations to better suit the needs of laypeople and specific use cases.

Furthermore, as context is considered to be a very important factor, a total of seven individual context factors were added (Anjomshoae, 2022; Arrieta et al., 2020; Miller, 2019; Tomsett et al., 2018):

1. Decision context: The nature of the decision that the XAI system is assisting with (e.g., high-stakes decisions like medical diagnoses or low-stakes decisions like movie recommendations) can impact the level of interpretability required.
2. Domain: The specific field or industry in which the XAI system is operating (for example, healthcare, finance, criminal justice) can influence the interpretability requirements, as different domains have varying levels of complexity and regulations being standard.
3. Stakeholder requirements: Different stakeholders (e.g., end-users, regulators, decision-makers) might have different expectations and needs from the XAI system, affecting the type of explanation that is considered interpretable.
4. Regulatory and legal environment: The legal and regulatory landscape in which the XAI system operates may impose specific interpretability requirements, such as the need for transparency and accountability.
5. Cultural context: The cultural background and values of the users can influence their interpretation of the explanations provided by the XAI system. For example, some people have naturally low trust levels, others very high. Some people are sceptical towards computers, others are not.
6. Temporal context: The time-sensitive nature of the decision or problem at hand may impact the level of interpretability required, as urgent situations might necessitate more straightforward explanations.
7. Social context: The social dynamics and relationships between the users and other stakeholders can influence the interpretability requirements, as trust and credibility may play a role in how explanations are perceived.

First of all, please note that this list is not necessarily exhaustive, and other contextual factors may also be relevant (depending on the specific application of the XAI system). However, by considering these seven contextual factors, researchers and developers can better tailor their explanations to improve interpretability for various scenarios and user groups. Second of all, it should be noted that the context variables do not only influence the final dependent variable: XAI interpretability. They moreover influence the other two categories of factors: XAI factors and user factors. This is because, depending on the context, the weight that can be accounted to the variables will most likely change.

7.5. Key principles of the second framework prototype

Considering the feedback from the interviews and the new framework, here's a revised version of the key principles:

1. Trustworthiness in general is of great importance for laypeople in interpreting XAI explanations. Trust can be interpreted in two different ways, and both are valid when designing an XAI. On the one hand, you require a high trustworthiness of the explanation, and on the other hand, you want a high level of trust from the user. A trade-off may be experienced between trustworthiness and relevance. Finding a balance between the two is necessary.
2. Comprehensiveness and complexity must be balanced with simplicity. A model that provides sufficient information to address the user's needs or goals without overwhelming them with complexity is more likely to be easily understood and trusted. Striking the right balance between simplicity on the one hand, and comprehensiveness and complexity on the other, is essential for optimal XAI interpretability.
3. Affordance and user relevance are essential for XAI interpretability. A model that aligns with human intuition, knowledge, and the user's specific context, appears deliberate or intentional, and can be easily understood without the need for additional explanation or knowledge is more likely to be easily interpreted by laypeople users. Explanations that focus on abnormalities can provide more compelling insights, but they may contradict users' prior beliefs. Striking a balance between highlighting abnormalities and maintaining coherence with prior beliefs can help users accept and understand explanations better.
4. Explanations that focus on intentionality provide insights into why events occurred, while actionable explanations offer practical advice for addressing a problem. Balancing these factors can help users understand the causes behind a situation and identify effective solutions.
5. High model fidelity means that explanations accurately represent the underlying model, while high explanation fidelity means that explanations are faithful to the real-world context. Ensuring both model and explanation fidelity is difficult, yet important to provide accurate and useful explanations.
6. Contextual factors play a significant role in XAI interpretability. The interpretability of an explanation can be influenced by various contextual factors, such as the user's domain knowledge, their goals, the specific application, and the cultural and social context. Taking these factors into account when designing explanations is crucial for achieving optimal interpretability.
7. Consider your target audience with great detail. Explanations should be coherent with users' prior beliefs and knowledge, as well as relevant to their goals and needs. This can help users understand the explanation better and make it more likely that they accept it. Also, be aware of the potential trade-offs between factors, such as simplicity vs. comprehensiveness or transparency vs. complexity, and aim to strike a balance that optimizes the interpretability of the explanation for the target audience.
8. Explanations should take into account abnormal or unusual factors and intentional actions, as these can provide a better understanding of the causal chains and contribute to the interpretability of the explanation.

9. Explanations should be presented in an intuitively understandable manner with clear language and structure, allowing users to easily grasp the information provided.
10. Evaluating the interpretability of an explanation in general is essential. The use of both measurable and perceived interpretability factors is important for evaluating how well an explanation meets the needs of laypeople users. This evaluation should consider factors such as recall response time, information transfer rate, and user feedback in general.
11. Stakeholder engagement and ethical considerations are important in XAI interpretability. Engaging with stakeholders, understanding their needs and expectations, and addressing potential ethical issues (such as bias, over-reliance on AI, and transparency trade-offs) are crucial for designing and implementing XAI systems that are both interpretable and responsible.
12. Recall response time and information transfer rate are metrics for XAI interpretability. The faster the model can transfer information to the user and the less time it takes for the user to recall relevant information needed to understand the model's output, the higher the interpretability is.

Concluding this section, the key principles of the revised framework encapsulate essential aspects and considerations for improving XAI interpretability. They illustrate an intricate balance among numerous factors that influence the perception and usability of XAI explanations, from trustworthiness and simplicity to user relevance and context.

The principles underscore a user-centric approach in XAI design, emphasizing the importance of aligning explanations with users' prior beliefs, their intuitive understanding, and the specific context of their interaction with the XAI system. Paying careful attention to these aspects and consistently evaluating the interpretability of explanations could lead to a framework that is not only theoretically robust but also practically valuable and ethically grounded. The ultimate vision for this framework is to inform and inspire the development of more transparent, accountable, and user-friendly XAI systems that truly meet the needs and expectations of laypeople users.

7.6. Conclusion and reflection on the second framework prototype

The development of the second framework prototype entailed critical refinement steps, including iterative evaluations, expert inputs, and a focus on the pragmatic requirements of XAI applications. This process transformed the framework from a purely comprehensive model into a more accessible and context-sensitive tool, striving for simplicity without losing sight of its primary purpose of enhancing XAI interpretability. While the shift towards a less complex model might seem to compromise its comprehensiveness, it was a strategic response to meet the demand for a more user-friendly and widely accepted framework. The revised version emphasizes the importance of context in XAI interpretability, leading to the categorization of three major influencing factors: XAI-related, user-related, and contextual.

Seven contextual factors were added to further address interpretability issues, emphasizing the multifaceted nature of interpretability, shaped by the system, user, and significantly, the operating environment. While not exhaustive, this robust set of factors forms a foundation adaptable to various scenarios and user groups.

Incorporating the revised key principles and contextual factors, the second framework prototype aims to serve as a highly effective tool for enhancing general XAI interpretability. It offers an evolved approach that caters to the diverse needs of XAI developers, researchers, and stakeholders, providing guidance to navigate the complex interplay between the XAI system, the user, and the context.

As this prototype represents the most advanced version for general (non-use-case-specific) XAI interpretability in this study, it is an invaluable resource for creating accessible, user-

friendly, and practical interpretability solutions in the XAI field. It underscores the importance of trade-offs, user relevance, audience targeting, and ethical considerations in enhancing XAI interpretability.

However, acknowledging that there is always room for further refinement, the framework will be subjected to additional scrutiny in a medical use case, in a survey format with laypeople as respondents. Detailed in the next chapter (chapter 8), this step will offer insights from a specific real-world scenario and contribute to potential enhancements of the framework.

Chapter 8. Evaluation of framework part II

Chapter 8 will concern the evaluation of the framework on laypeople through a use-case-based survey, in order to test its effectiveness on the target audience. This is in line with the final sub-question, sub-question 5:

"How can the second framework prototype be applied in practice and thus evaluated?"

The use-case-based survey is based entirely on the trade-offs as presented in the key principles of the last section of the previous chapter. There are numerous trade-offs that XAI developers should make when designing an XAI system. For example, explanations that are more comprehensive (are automatically more transparent) tend to cover more aspects of a problem, while simpler and more generalizable explanations are easier to interpret. Striking a balance between these two combinations of factors is believed to be important for overall interpretability. However, according to the expert interviews, when considering laypeople, it may be more important for explanations to be simple and generalizable. This is what will be tested in the survey on laypeople. Moreover, the results of the survey will be analysed to identify any further areas for improvement with regard to the individual factors and their weights used in the framework.

8.1. Nature of second evaluation round

The second evaluation round will validate the framework by focusing on the influential factors on XAI interpretability for laypeople via a survey. The survey presents a medical use case. Therefore, the survey has been validated by two medical professionals, both affiliated with the LUMC. LUMC stands for Leiden University Medical Centre, which is a Dutch academic medical centre located in the city of Leiden. The LUMC is a large teaching hospital that offers a wide range of patient care services in various medical fields, including oncology, cardiology, neurology, paediatrics, and transplantation. The medical centre also has a strong focus on medical research and education, and therefore hosts various research institutes and departments. Furthermore, the survey made use of commonly accepted guidelines and frameworks for developing similar questionnaires (Aithal & Aithal, 2020).

The full, reviewed, template of the survey can be found in appendix F. A general overview of the survey is presented below.

The first section of the survey presents a simple welcome message and an explanation of the context of the survey. In short, the survey puts participants in a place where they experience numerous symptoms and use a self-diagnosis app (SymptomSolver) to determine their illness/condition.

In the second part of the survey, a total of six sets of explanations are provided in separate sections. Each section starts by setting the context. It does so by providing the symptoms that the participant is supposedly feeling. After which the explanations that are provided by the app SymptomSolver are presented. The respondent will be asked to rank those explanations from better to worse. First, in terms of natural understandability. Second, in terms of general satisfaction.

These two factors were selected as they collectively provide a comprehensive measure of interpretability in the context of AI explanations. While they may not encompass every facet of interpretability, they offer a practical and accessible approach to understanding how well the participants can comprehend the explanations. Asking participants directly about the interpretability of an explanation was deemed less effective, as this concept can be somewhat technical and abstract. Therefore, by using "natural understandability" and "general satisfaction", the aim is to capture the foremost essential aspects of interpretability.

These six sets of explanations are, as previously mentioned, based on the trade-offs that are addressed in the key principles from section 7.5. The sets of explanations are briefly shown in Table 6, whilst the full specifics and explanations themselves are shown in appendix F.

Explanation set	Associated factors	Individual explanation characteristics
Set 1	Comprehensiveness, transparency, simplicity, and generalizability	High comprehensiveness, high transparency, little simplicity, little generalizability
		Little comprehensiveness, little transparency, high simplicity, high generalizability,
		Balanced comprehensiveness, transparency, simplicity, and generalizability
Set 2	Complexity, transparency, simplicity, and clarity	High complexity, high transparency, little simplicity, little clarity
		Little complexity, little transparency, high simplicity, high clarity
		Balanced complexity, transparency, simplicity, and clarity
Set 3	Abnormality, coherence with prior beliefs and affordance	High abnormality, little coherence with prior beliefs and little affordance
		Little abnormality, high coherence with prior beliefs and high affordance
		Balanced abnormality, coherence with prior beliefs, and affordance
Set 4	Intentionality and actionability	High intentionality, little actionability
		Little intentionality, high actionability
		Balanced intentionality and actionability
Set 5	Model fidelity and explanation fidelity	High model fidelity, little explanation fidelity
		Little model fidelity, high explanation fidelity
		Balanced model and explanation fidelity
Set 6	Trustworthiness and relevance	High trustworthiness, little relevance
		Little trustworthiness, high relevance
		Balanced trustworthiness and relevance

Table 6. Explanation sets specifics for the survey

Finally, the third section provides us with information about the participant. The important takeaway from this section is to assess if the user is in fact a layperson considering the XAI at hand. Since the framework is intended to be used on laypeople, and not on experts, it must be ensured that answers given by medical experts (however interesting they may be) are not taken into account during analysis of final results.

When designing the explanations that formed the basis for the survey, numerous important elements from the second framework prototype that required attention already came to light. The first thing that became apparent, was that several factors are still very similar to each other. As one of the expert interviewees pointed out (see appendix E, interview 8), when thinking of examples for each factor, you will truly find out if two factors represent the same thing.

8.1.1. Transparency, complexity, comprehensiveness, simplicity, generalizability and clarity

Consider transparency, complexity, and comprehensiveness. Or simplicity, clarity, and generalizability. One cannot make an explanation more comprehensive without making it more complex or transparent. Similarly, one cannot make an explanation simpler without making it clearer or more generalizable. The only thing is that one can, for example, make an explanation more comprehensive without making it more complex, as long as the information is presented in a non-complex way. That is the reason why there is a slight difference in explanation set 1 and 2 (see Table 6). These two groups of factors are most relevant and can be categorized under two distinct approaches: the holistic approach and the parsimonious approach. The holistic approach encompasses transparency, complexity, and comprehensiveness, while the parsimony approach focuses on simplicity, clarity, and generalizability. These groups of factors and their relationships have been visualized in Figure 16.

Accordingly, the first key principle can be altered.

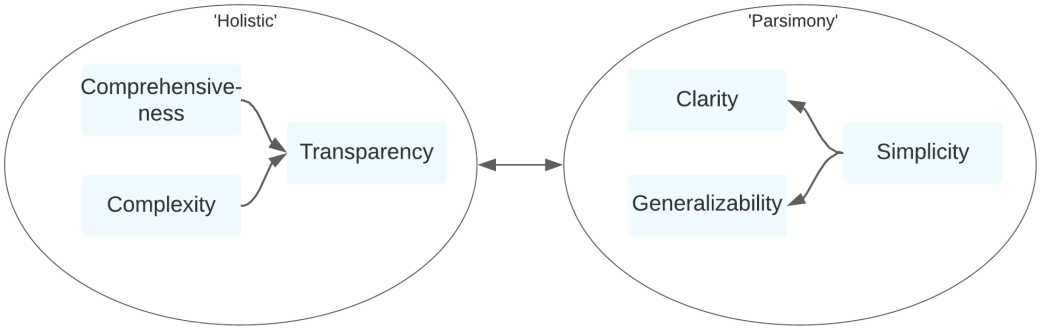

Figure 16. Holistic versus reductionist factors

8.1.2. Coherence with prior beliefs and affordance

Coherence with prior beliefs and affordance are often highly related. When increasing the coherence with prior beliefs, the level of affordance is consequently automatically increased. This is another example of great overlap that came to light when designing explanations for the survey. This overlap is dealt with in explanation set 3.

8.1.3. ITR and RRT

Secondly, Information Transfer Rate (ITR) and Recall Response Time (RRT) are metrics that can be used to assess the interpretability of an explanation provided by an XAI system, rather than factors that influence the interpretability of the explanation itself. While these metrics are valuable for evaluating the interpretability of XAI explanations, they are not factors that directly influence interpretability. Instead, they are the outcomes of the interaction between the factors influencing interpretability and the user's cognitive abilities. By measuring ITR and RRT, researchers can gauge the effectiveness of an XAI system's explanations and make adjustments to improve their interpretability based on the identified factors. However, it is essential to understand that these metrics serve as indicators of interpretability performance rather than factors that shape interpretability itself. This became very apparent when creating the sets of explanations as discussed in section 8.1. Since it is not at all possible to create an explanation with a certain level of ITR and/or RRT. Therefore, these two factors will not be discussed in the survey and removed from the framework in their current form as of now.

8.2. Survey practicalities

The entire survey is estimated to take approximately 8 minutes. To ensure that there is enough support for the survey, participants are collected via the platform Prolific, as briefly touched upon in the methodology chapter. The primary goal is to get a minimum of 200 responses. The only selection criterion was that English is considered a primary language by the participant. Furthermore, the decision has been made to make use of a balanced sample, therefore distributing the study evenly to male and female participants.

In the end, 204 people filled in the survey. Of those 204 people, it was interesting to find that 8 people are considered experts on the topic on which the XAI provides an explanation. In this case, medical experts. Since the aim of this study is to gain insights into the interpretability of laypeople, as opposed to experts, these responses were not taken into account during analysis. Furthermore, after excluding medical experts from the study, participants that filled in the survey too quickly and failed to complete attention checks were also excluded from the survey. This is according to the methodology chapter. This resulted in a final number of 157 responses for the survey. For the details on this analysis, see appendix G.

8.3. Survey results

In this section, the results of the survey conducted with laypeople will be presented. To this extent, Table 6 was used as a basis and the results were added to that table, consequently

creating Table 7. Table 7, therefore, presents the characteristics of each individual explanation. For example, high comprehensiveness, high transparency, little simplicity, and little generalizability, from explanation set 1. Since participants had to present their top 3 explanations per explanation set, for each specific explanation, we can provide the number of people that presented that specific explanation as the best one of that set. Thereafter, we can see how many people presented that specific explanation as the middle one, and as the worst one, within that explanation set. Not only does this allow a comparison of how well one explanation did, we can see how well it did in comparison to the other two specific explanations within that set. Therefore, allowing us to see which explanation was chosen by the public as being most understandable/satisfying.

Secondly, per explanation set, participants were asked to present two rankings. One based on natural understandability (presented as 'U' in the table) and one based on general satisfaction (presented as 'S'). This is highly interesting to compare when analysing the results.

Finally, all results are presented as percentages. Note that the percentages shown add up to 100% when considering one individual explanation, and one metric. So, for example, the first explanation of explanation set 1, and considering natural understandability (U). In that case, 25,5% + 15,0% + 59,5% = 100%. But also: 25,5% + 35,3% + 39,2% = 100%. That 25,5% therefore means that out of all people that voted, 25,5% saw explanation 1 as the best. And, out of all three explanations within that set, 25,5% of the votes for the best explanation were awarded to explanation number 1. Logically, it works both ways.

Explanation set	Associated factors	Individual explanations and distribution			
		Individual explanation characteristics	Best score	Mid score	Worst score
Set 1	Comprehensiveness, transparency, simplicity, and generalizability	High comprehensiveness, high transparency, little simplicity, little generalizability	U: 25,5%	U: 15,0%	U: 59,5%
			S: 49,0%	S: 15,0%	S: 36,0%
		Little comprehensiveness, little transparency, high simplicity, high generalizability	U: 35,3%	U: 28,6%	U: 26,1%
			S: 20,4%	S: 35,4%	S: 44,2%
		Balanced comprehensiveness, transparency, simplicity, and generalizability	U: 39,2%	U: 46,4%	U: 14,4%
			S: 30,6%	S: 49,7%	S: 19,7%
Set 2	Complexity, transparency, simplicity, and clarity	High complexity, high transparency, little simplicity, little clarity	U: 11,0%	U: 11,7%	U: 77,3%
			S: 25,5%	S: 24,8%	S: 49,7%
		Little complexity, little transparency, high simplicity, high clarity	U: 44,8%	U: 34,4%	U: 20,8%
			S: 20,3%	S: 35,3%	S: 44,4%
		Balanced complexity, transparency, simplicity, and clarity	U: 44,2%	U: 53,9%	U: 01,9%
			S: 54,2%	S: 39,9%	S: 05,9%
Set 3	Abnormality, coherence with prior beliefs and affordance	High abnormality, little coherence with prior beliefs and little affordance	U: 18,5%	U: 50,0%	U: 31,5%
			S: 10,7%	S: 47,0%	S: 42,3%
		Little abnormality, high coherence with prior beliefs and high affordance	U: 66,4%	U: 28,1%	U: 05,5%
			S: 60,4%	S: 29,5%	S: 10,1%
		Balanced abnormality, coherence with prior beliefs, and affordance	U: 15,1%	U: 21,9%	U: 63,0%
			S: 28,9%	S: 23,5%	S: 47,7%
Set 4	Intentionality and actionability	High intentionality, little actionability	U: 30,2%	U: 41,6%	U: 28,2%
			S: 15,8%	S: 42,8%	S: 41,4%
		Little intentionality, high actionability	U: 31,5%	U: 22,8%	U: 45,6%
			S: 27,0%	S: 30,9%	S: 42,1%
		Balanced intentionality and actionability	U: 38,3%	U: 35,6%	U: 26,2%
			S: 57,2%	S: 26,3%	S: 16,4%
Set 5	Model fidelity and explanation fidelity	High model fidelity, little explanation fidelity	U: 17,4%	U: 51,4%	U: 31,3%
			S: 12,8%	S: 37,6%	S: 49,7%
		Little model fidelity, high explanation fidelity	U: 67,4%	U: 22,9%	U: 09,7%
			S: 59,1%	S: 28,9%	S: 12,1%
		Balanced model and explanation fidelity	U: 15,3%	U: 25,7%	U: 59,0%
			S: 28,2%	S: 33,6%	S: 38,3%
Set 6	Trustworthiness and relevance	High trustworthiness, little relevance	U: 12,9%	U: 49,0%	U: 38,1%
			S: 05,4%	S: 45,3%	S: 49,3%
		Little trustworthiness, high relevance	U: 70,1%	U: 23,8%	U: 06,1%
			S: 70,3%	S: 19,6%	S: 10,1%
		Balanced trustworthiness and relevance	U: 17,0%	U: 27,2%	U: 55,8%
			S: 24,3%	S: 35,1%	S: 40,5%

Table 7. Core survey results. U = natural understandability and S = general satisfaction of the explanation

The results from Table 7 will be discussed in the following section.

8.4. Survey conclusions

After analysing the survey responses, the following aggregated feedback and common themes emerge regarding the second framework prototype. This feedback will be used to revise and refine the framework, ultimately leading to the development of the final version, which will be presented in section 8.5.

8.4.1. Explanation set 1

For natural understandability, the balanced approach (explanation 3) has the highest percentage of best rankings, indicating that participants found this explanation to be the most effective in terms of interpretability, as it strikes a balance between comprehensiveness/ transparency, and simplicity/generalizability.

However, for general satisfaction, the first explanation, which emphasizes high comprehensiveness and high transparency at the expense of simplicity and generalizability,

receives the highest percentage of best rankings. This suggests that participants might value these two factors more when evaluating the overall satisfaction with the explanation, despite the potential drawbacks in terms of complexity and specialization.

Finally, when comparing the results of explanation set 1 with the other sets, it is visible that the responses are rather divided. There is (for example compared to explanation set 3, 5, or 6) no one individual explanation that performs superior. Some people prefer explanations that are more comprehensive and transparent. Others prefer explanations that are simpler and more generalizable. Whilst another (large) group prefers the balanced approach. This observation itself is a clear indication to prefer the balanced approach, since, on average, balancing all factors will result in average satisfaction for all respondents.

Comparing the survey results with the expert interview weights, it is apparent that comprehensiveness (4.5) and transparency (3.5) are both highly weighted factors. This could explain why the first explanation performs better in terms of general satisfaction. Simplicity (3.0) and generalizability (2.0) have relatively lower weights, but they are still considered important. In terms of natural understandability, the balanced approach (explanation 3) aligns well with expert opinions. However, according to experts, comprehensiveness and transparency should have been superior compared to simplicity and generalizability, which is not the case. Therefore, a revision of these weights may be required.

In conclusion, the balanced approach seems to be the most effective in terms of natural understandability, whereas the first explanation, which prioritizes comprehensiveness and transparency, is preferred when considering general satisfaction. This suggests that a balance between the factors is important for interpretability, but the higher-weighted factors of comprehensiveness and transparency may hold a slight edge when evaluating overall satisfaction with the explanations. In the end, due to the great division amongst participants regarding this explanation set, a balanced approach is the approach that will satisfy most people.

8.4.2. Explanation set 2

For natural understandability, the little complexity, little transparency, high simplicity, high clarity approach and the balanced approach have around the same percentage of best rankings, suggesting that participants found it most effective when considering interpretability. This explanation strikes a balance between complexity, transparency, simplicity, and clarity, however, leaning more towards high simplicity and clarity, as opposed to high complexity and transparency. This becomes highly apparent when looking at which explanation is considered the worst explanation (the final column). 77% of the people consider the highly complex and transparent explanation the worst explanation.

When examining general satisfaction, the balanced approach (explanation 3) holds the highest percentage of best rankings, which indicates that participants found this explanation to be the most satisfying overall. It appears that a balance between the factors is important not only for interpretability but also for general satisfaction in this case. For general satisfaction, the little complexity, little transparency, high simplicity, and high clarity approach loses traction compared to the balanced approach.

Comparing these results with the expert interview weights, it is apparent that transparency (3.5) and simplicity (3.0) are both assigned significant weights. Complexity (2.5) and clarity (2.5) have slightly lower weights, but they are still deemed important. The strong performance of the balanced approach (explanation 3) in both natural understandability and general satisfaction seems to be in line with the expert opinions, as it addresses all these factors.

In conclusion, the balanced approach appears to be rather effective concerning explanation set 2. This suggests that striking a balance between the factors of complexity, transparency,

simplicity, and clarity is important for XAI interpretability and aligns well with expert opinions. However, when an XAI developer finds him- or herself in a situation where a decision needs to be made, leaning towards high simplicity and clarity, as opposed to high complexity and transparency appears to be the best option.

8.4.3. Explanation set 3

When examining explanation set 3, it is rather straightforward to create a top 3. The majority of respondents selected the little abnormality, high coherence with prior beliefs and high affordance explanation as the best explanation for both natural understandability and general satisfaction. Therefore, when a trade-off occurs between these three factors, XAI developers should aim for high coherence with prior beliefs and high affordance. Secondly, it is surprising to see that the balanced explanation received the worst score from the majority of participants. Leaving the high abnormality explanation in second place.

Comparing these results with the expert interview weights, it is apparent that the weights for affordance (3.5), abnormality (3.0) and coherence with prior beliefs (2.0) do not reflect the same relationship. Therefore, a revision of these weights may be wise given this context.

In conclusion, for explanation set 3, the explanation with little abnormality and high coherence with prior beliefs and high affordance appears to be the most effective in terms of both natural understandability and general satisfaction. This suggests that coherence with prior beliefs and affordance, along with low abnormality, are important factors for XAI interpretability.

8.4.4. Explanation set 4

For natural understandability and general satisfaction, the balanced approach to intentionality and actionability (explanation 3) has the highest percentage of best rankings. This suggests that participants found explanations that maintain a balance between intentionality and actionability to be more understandable and generally satisfying. Furthermore, the other two individual explanations performed only slightly worse in terms of natural understandability. The balanced approach got 38,3% of the votes, whilst the other two received 31,5% and 30,2%. Also, when comparing the results of explanation set 4, along the same lines as explanation set 3, with the other sets, it is visible that the responses are rather divided. There is (for example compared to explanation set 3, 5, or 6) no one individual explanation that performs superior. Some people prefer explanations that are more comprehensive and transparent. Others prefer explanations that are simpler and more generalizable. Whilst another (large) group prefers the balanced approach. This observation itself is a clear indication to prefer the balanced approach, since on average, balancing all factors will result in average satisfaction for all respondents.

When comparing these results with the expert interview weights, it should be noted that intentionality (1.5) and actionability (2.0) have been assigned relatively similar weights. Therefore, the strong performance of the balanced approach in both natural understandability and general satisfaction suggests that balancing these factors accordingly is indeed preferred.

In conclusion, for explanation set 4, the balanced approach to intentionality and actionability appears to be the most effective in terms of both natural understandability and general satisfaction. Furthermore, due to great division among participants regarding this explanation set, a balanced approach is the approach that will satisfy most people. This suggests that achieving a balance between intentionality and actionability is important for XAI interpretability, even though these factors have been assigned relatively lower weights by the experts.

8.4.5. Explanation set 5

For both natural understandability and general satisfaction, the explanation with little model fidelity and high explanation fidelity (explanation 2) has the highest percentage of best rankings. This suggests that participants found explanations with less focus on model fidelity

and more emphasis on explanation fidelity to be more understandable and satisfying. Secondly, it is surprising to see that the balanced explanation received the worst score from the majority of participants. Leaving the high model fidelity explanation in second place.

When comparing these results with the expert interview weights, it is apparent that model fidelity (3.0) has been assigned a moderate weight, while the weight for explanation fidelity is lower (1.5). The strong performance of the explanation with little model fidelity and high explanation fidelity suggests that explanation fidelity may however be more important compared to model fidelity. A revision of the weights is in order, which will be presented in section 8.5.

In conclusion, an explanation with little model fidelity and high explanation fidelity appears to be the most effective in terms of both natural understandability and general satisfaction. This suggests that prioritizing explanation fidelity over model fidelity is important for XAI interpretability, even though model fidelity has been assigned a moderate weight by the experts.

8.4.6. Explanation set 6

For both natural understandability and general satisfaction, the explanation with little trustworthiness and high relevance (explanation 2) has the highest percentage of the best rankings. This suggests that participants found explanations with less focus on trustworthiness and more emphasis on relevance to be more understandable and satisfying. Secondly, it is surprising to see that the balanced explanation received the worst score from the majority of participants. Leaving the high trustworthiness in second place.

It is apparent that trustworthiness (5.0) has been assigned the highest weight, while relevance (4.0) has been assigned a slightly lower weight, when comparing these results with the expert interview weights. The strong performance of the explanation with little trustworthiness and high relevance in both natural understandability and general satisfaction suggests that relevance might be more important than trustworthiness for effective XAI interpretability.

In conclusion, the explanation type with little trustworthiness and high relevance appears to be the most effective in terms of both natural understandability and general satisfaction. This suggests that prioritizing relevance over trustworthiness is important for XAI interpretability, even though trustworthiness has been assigned a higher weight by the experts.

8.4.7. Aggregated conclusions

This section will provide one overview with the aggregated conclusions for each explanation set, in the form of a table.

Explana-tion set	Associated factors	Preferred manner of handling the trade-off
Set 1	Comprehensiveness, transparency, simplicity, and generalizability	A balance between comprehensiveness and transparency on one side and simplicity and generalizability on the other side is the preferred alternative.
Set 2	Complexity, transparency, simplicity, and clarity	A balance between complexity and transparency on one side and simplicity and clarity on the other side is the preferred alternative. However, slightly leaning towards high simplicity and clarity, as opposed to high complexity and transparency (if necessary) is the best option.
Set 3	Abnormality, coherence with prior beliefs and affordance	Focusing on high coherence with prior beliefs and high affordance as opposed to high abnormality is the preferred alternative.
Set 4	Intentionality and actionability	A balance between intentionality and actionability is the preferred alternative.
Set 5	Model fidelity and explanation fidelity	Focusing on high explanation fidelity as opposed to high model fidelity is the preferred alternative.
Set 6	Trustworthiness and relevance	Focusing on high relevance as opposed to high trustworthiness is the preferred alternative.

Table 8. Survey conclusions

The weights for the final framework have been altered under the assumption that the overall weights presented by the experts were correct. However, the trade-offs were not interpreted correctly (considering the medical context). For instance, in the case of the trustworthiness and relevance trade-off, trustworthiness initially had a weight of 5, while relevance had a weight of 4. The survey results revealed that, for laypeople, relevance was actually more important than trustworthiness. Consequently, the weights in the updated framework were switched. This is further discussed in the discussion in chapter 9. The following weight changes have taken place:

- In the second framework prototype, comprehensiveness was weighed at 4.5, while simplicity was weighed at 3.0. Model fidelity and explanation fidelity have changed weights, in order to accommodate the findings with regard to explanation set 1 and 2. This transition ensures that the average weight of comprehensiveness and transparency (3.25) is balanced with the average weight of simplicity and generalizability (3.5). Furthermore, it ensures that the average weight of complexity and transparency (3.0) is balanced with, whilst slightly less than the average weight of simplicity and clarity (3.5).
- In the second framework prototype, abnormality was weighed at 3.0 whilst coherence with prior beliefs was weighed at 2.0. Affordance was assigned the weight 3.5. Abnormality and coherence with prior beliefs have switched weights, in order to accommodate the findings with regard to explanation set 3.
- In the second framework prototype, model fidelity was weighed at 3.0 whilst explanation fidelity was weighed at 1.5. Model fidelity and explanation fidelity have changed weights, to accommodate the findings with regard to explanation set 4.
- In the second framework prototype, trustworthiness was considered the most important factor with a weight of 5.0 whilst relevance was weighed at 4.0. Trustworthiness and relevance have changed weights, in order to accommodate the findings with regard to explanation set 6.

8.5. Final framework prototype

In conclusion, the aggregated and analysed responses from the survey amongst laypeople highlight the need for addressing trade-offs. What does this imply for the final framework prototype? The first thing that should be noted is that the entire survey was in the medical context. Therefore, the middle section of the second framework prototype (Figure 15) comes into place. The conclusions that can be drawn from the survey, will in the end only directly be valid for XAI explanations in the medical context and domain.

That being stated, it is of course possible to base general conclusions upon the results from the specific medical use case. As has been stated by several interviewees, providing weights to the individual factors is highly dependent on context. Therefore, although weights have been altered specifically for the medical context, this does not imply that these weights are applicable to all contexts.

Having taken that into account, the final framework is presented below in Figure 17. This version of the framework can be presented as the framework that can assist in assessing the interpretability of an explanation for laypeople in the medical context.

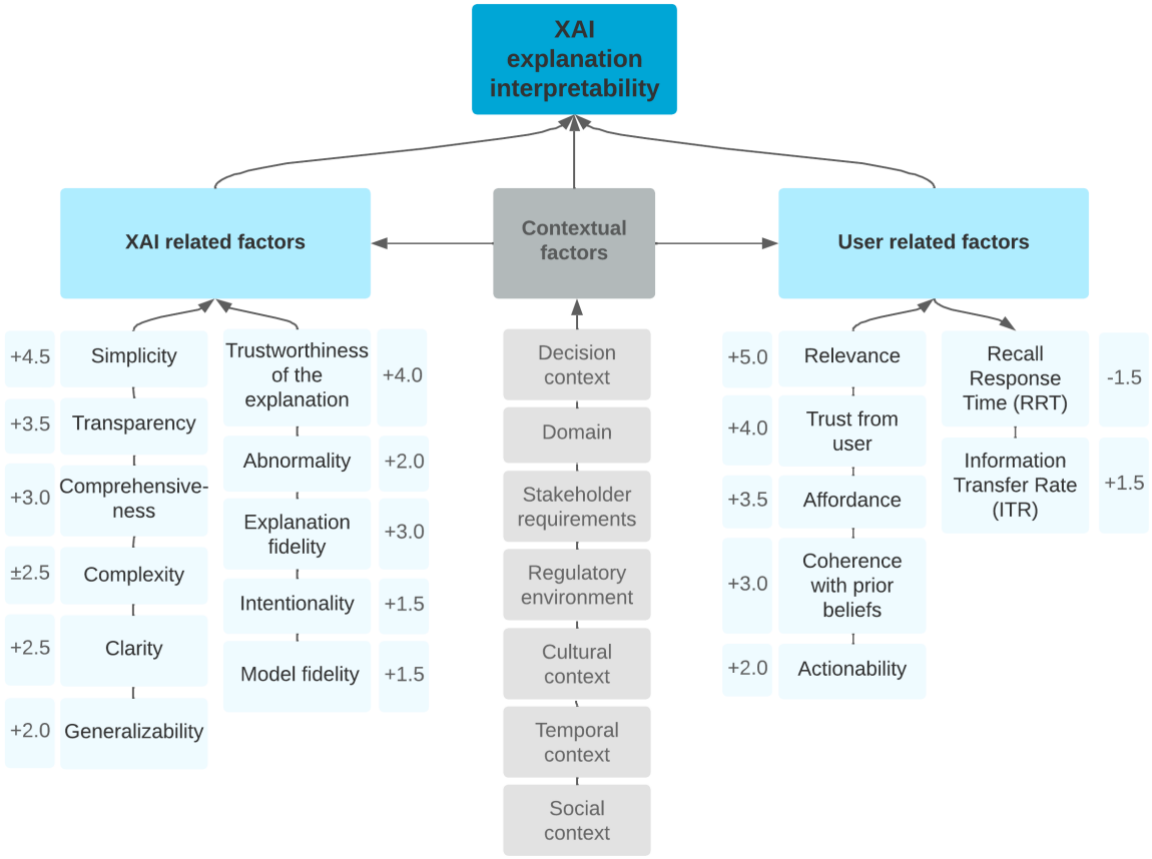

Figure 17. Final framework specifically for medical contexts

8.6. Key principles of the final framework

The final set of key principles that accompanies the framework as presented in section 8.5, are shown below.

1. Trustworthiness and relevance are crucial for laypeople in interpreting XAI explanations. Trust can be interpreted in two different ways, both of which are important when designing an XAI system. On the one hand, you need a high trustworthiness of the explanation itself, and on the other hand, you want a high level of trust from the user. Relevance is also of great importance for laypeople, as an explanation must be pertinent

to the explainee's goals and objectives. When a trade-off between trustworthiness and relevance arises, prioritize high relevance over high trustworthiness.

2. Striking the right balance between simplicity, clarity, and generalizability on the one hand, and comprehensiveness, complexity, and transparency on the other is essential for optimal XAI interpretability. A model that provides comprehensive information to address the user's needs or goals transparently without overwhelming them with complexity, therefore remaining clear, simple, and generalizable, is more likely to be easily interpreted. When a trade-off is necessary, lean towards high simplicity and clarity over high complexity and transparency.
3. Affordance, coherence with prior beliefs, and addressing abnormalities are essential for XAI interpretability. A model that aligns with human intuition, knowledge, and the user's specific context, and can be easily understood without the need for additional explanation or knowledge is more likely to be easily interpreted by laypeople users. While explanations that focus on abnormalities can provide more compelling insights, they may contradict users' prior beliefs. Prioritize high coherence with prior beliefs and high affordance over emphasizing abnormalities when facing trade-offs.
4. Balancing explanations that focus on intentionality (why events occurred) and actionability (practical advice for addressing a problem) is essential for helping users understand causes and identify effective solutions. A model that conveys both the underlying reasons for the events and offers actionable steps to address the situation is more likely to be easily interpreted by laypeople users. This balance helps users make sense of the situation and take appropriate action based on the explanation provided.
5. Ensuring both model and explanation fidelity is difficult, yet important for accurate and useful explanations. Explanation fidelity focuses on the quality of explanations, while XAI model fidelity measures the alignment between the XAI model and the original AI model. When facing trade-offs between these two aspects, prioritizing high explanation fidelity over high model fidelity to provide more contextually relevant and meaningful explanations to laypeople users is most important.
6. The primary goal of XAI developers when designing explanations for laypeople should be to maximize adherence to every single factor in the framework. By striving to optimize each aspect of the explanation, including trustworthiness, relevance, simplicity, clarity, coherence, intentionality, actionability, fidelity, contextualization, and ethical considerations, developers can create more effective and interpretable explanations. Recognizing that trade-offs may be necessary for certain situations, the overarching objective remains to achieve the highest level of interpretability and user satisfaction possible by considering all factors and continuously refining the design based on evaluations and user feedback.
7. Contextual factors significantly influence XAI interpretability. The interpretability of an explanation can be influenced by various contextual factors, such as the user's domain knowledge, their goals, the specific application, and the cultural and social context. Taking these factors into account when designing explanations is crucial for achieving optimal interpretability. By considering the unique context of each user, designers can create more tailored and effective explanations that resonate with the user's understanding and experiences.
8. Evaluate interpretability using both measurable and perceived interpretability factors, such as recall response time, information transfer rate, and user feedback. This evaluation should assess how well an explanation meets laypeople users' needs. By gathering quantitative and qualitative data on the effectiveness of the explanation, designers can identify areas for improvement and make informed decisions on how to enhance the overall interpretability of their XAI systems. Regularly conducting evaluations and iterating on the design can help ensure that explanations continue to evolve and remain relevant to the needs of the users. Recall response time and information transfer rate are prime examples of metrics for XAI interpretability. The faster the model can transfer information to the user and the less time it takes for the user to recall relevant information needed to understand the model's output, the higher

the interpretability is. These metrics can help designers quantitatively evaluate the effectiveness of their explanations and identify areas for improvement.

9. Stakeholder engagement and ethical considerations are vital in XAI interpretability. Engage with stakeholders, understand their needs and expectations, and address potential ethical issues (e.g., bias, over-reliance on AI, transparency trade-offs) to design and implement interpretable and responsible XAI systems. By involving stakeholders throughout the development process, designers can ensure that the resulting explanations are not only interpretable but also ethically sound and aligned with the values and concerns of the users and other relevant parties. This collaborative approach can help foster trust in the XAI system and promote responsible AI practices.

In summary, the final key principles underscore the importance of addressing and managing the complex trade-offs inherent in the design of XAI systems, particularly in a medical context where decisions can carry high stakes. These principles articulate that trustworthiness, relevance, simplicity, clarity, coherence, intentionality, actionability, fidelity, contextualization, and ethical considerations are paramount for XAI interpretability. Balancing these elements, while challenging, is essential to provide explanations that are meaningful and satisfactory to the users, considering their unique context and understanding.

Moreover, the principles emphasize the role of continuous evaluation and refinement of the XAI systems. By employing both measurable and perceived interpretability factors, developers can iteratively improve the interpretability of their systems. Also, by placing significant value on stakeholder engagement and ethical considerations, the principles ensure the development of XAI systems that are not just technically sound but also ethically responsible. These key principles serve as a compass guiding the development of interpretable explanations, grounding the XAI design process in a deep understanding of user needs and ethical standards.

Chapter 9. Discussion

This chapter embarks on a critical appraisal of this thesis, examining the potential shortcomings and limitations while also paving the way for future research directions. It aims to provide a holistic understanding of the methodological choices made during the study, their implications, and the potential areas that could be improved or further developed in future studies.

The discussion begins by scrutinizing the key aspects of the research design and implementation, from factor selection to the assumptions made, and the potential biases introduced by the survey's demographics. Each subsection offers a balanced perspective, acknowledging the strength of the current approach while being open to its inherent limitations. Moreover, this chapter also addresses the importance of the broader socio-cultural and ethical considerations surrounding XAI. It underscores the need for a nuanced understanding of XAI's potential impact on various stakeholders, including laypeople and experts, and society as a whole. Finally, the chapter concludes with a thoughtful reflection on the ethical dimensions of XAI. This final section delves into the potential pitfalls of overreliance on XAI, the need for critical thinking, and the importance of maintaining a healthy balance of trust between human decision-makers and AI systems.

9.1. Shortcomings and recommendations for future research

All shortcomings and any possibly derived recommendations for future research are presented one by one in this section.

9.1.1. Factor selection

Firstly, the factors used provide a comprehensive overview and fair assessment of interpretability. Nevertheless, it should be noted that there are more valid factors that determine the interpretability of XAI. However, one has to draw a line at some point to find the right balance between comprehensiveness and complexity on the one hand and simplicity and clarity on the other hand. Adding numerous more factors to the final framework would most likely give a more comprehensive overview of XAI interpretability, however, that would not be advantageous to the framework's simplicity and clarity.

9.1.2. First framework prototype and first set of key principles

For the first framework prototype in chapter 6, the individual factors were not yet prioritized, and little focus was laid on the trade-offs inherent within the key principles. In retrospect, a more nuanced understanding of these factors might for example have led to more specific and insightful questions during the expert interviews. The absence of an initial prioritization of these factors potentially missed out on highlighting which factors hold more weight in interpretability and in which context. The interpretability of an XAI system is a multifaceted concept and not all factors hold equal importance in every situation. A prioritized list could have helped to better guide the development of the framework, emphasizing the factors of greatest importance based on specific user requirements or contexts. Furthermore, the trade-offs within the first framework prototype were not exhaustively explored in the first set of key principles. For instance, literature had also shed light on the fact that achieving simplicity might sometimes conflict with the need for comprehensiveness. Discussing these trade-offs could have helped to better illustrate the complex nature of the interpretability problem, which is not about maximizing every factor, but rather about finding a balance that best suits the particular needs of the user.

In future research, it would be beneficial to prioritize the factors and thoroughly discuss potential trade-offs at an early stage. This would allow the development of more refined interview questions for experts, further enhancing the quality and applicability of the

interpretability framework. The implications of this oversight in the current study underline the importance of iterative refinement and continuous learning in this rapidly evolving field of XAI.

9.1.3. Laypeople and experts

The created framework is intended for assessing interpretability specifically for laypeople. Theoretically, it could also be used for experts. However, it is important to note that it is not possible to use the framework to compare an expert's interpretability to a layperson's interpretability. Logically, it would appear as if the explanation to the expert is much more interpretable in comparison to the explanation to the layperson. It may be interesting to come up with measures that are different for experts on the one hand and laypeople on the other hand. In that case, the framework's results can be compared. However, this claim is largely coupled with the next two claims.

9.1.4. Theoretical nature

It is essential to acknowledge that the current framework is primarily theoretical in nature. While it serves as a valuable guide for researchers and practitioners in considering the crucial factors that influence XAI interpretability, it does not yet offer a practical tool for direct application. To address this, we have proposed actionable guidelines as a starting point for transforming the theoretical framework into a more pragmatic instrument. These guidelines are designed to streamline the process of XAI development and simplify the decision-making for various stakeholders in the XAI space. Future research should focus on refining and expanding these actionable guidelines, ensuring their applicability and effectiveness in real-world scenarios. This development would not only bridge the gap between XAI theory and practice but also contribute to more accurate and efficient AI in general, ultimately benefiting the wider AI space. By focusing on turning theoretical insights into practical tools, researchers can foster the responsible and widespread adoption of AI technologies across various industries and sectors.

A rather similar, but also important aspect of this research is the theoretical nature of the individual factors identified within the framework. The measurability of these factors presents a challenge in making the entire framework more practically applicable, as quantifying their respective influence on XAI interpretability may not be straightforward. Consequently, future research should address this issue by developing methodologies to effectively measure and quantify the impact of these factors on XAI interpretability. Such advancements would enable a more accurate and reliable understanding of the intricate interplay between the factors and their influence on pricing decisions. Additionally, improving the measurability of these factors would further contribute to the development of a practical tool, facilitating the implementation of informed XAI.

These shortcomings were also frequently mentioned by interviewees (interview 1, 3, 4, 5, 6, 7, 8, 9 and 11). The ultimate measure of the framework's success will therefore be its impact in real-world settings, including its ability to guide practitioners towards creating meaningful, interpretable explanations across diverse scenarios, users, and contexts. By continuously refining the framework in accordance with its guiding principles and practical application, it makes a significant contribution towards fostering a more transparent, accountable, and inclusive XAI field.

9.1.5. Assumptions

This research, as almost all available XAI research is based on numerous assumptions. The most relevant one is that the XAI interpretability framework assumes that people care at all about what a machine has to say (Miller, 2023). If people tend to dismiss recommendations and any explainability information, the framework becomes virtually useless.

The process of weight adjustment was based on the assumption that the initial weights derived from the expert interviews were accurate. However, it seemed that the trade-offs were not interpreted correctly, specifically within the medical context. For instance, in the case of the

trustworthiness and relevance trade-off, trustworthiness initially had a weight of 5, while relevance had a weight of 4. The survey results revealed that, for laypeople, relevance was actually more important than trustworthiness. Consequently, the weights in the updated framework were switched. The misinterpretation of trade-offs might stem from the focus of this research, which was specifically aimed at laypeople rather than subject matter experts, even though the latter is the most common application for XAI. This could have led to a subconscious inclination during the discussions with interviewees to prioritize explanations suitable for experts. In such a context, it is reasonable to argue that trustworthiness may indeed be more important than relevance.

9.1.6. Interview shortcomings

Conducting interviews as a primary source of data has proven to be a valuable method in many studies; however, it is not without potential limitations (Creswell & Poth, 2017; Hofisi et al., 2014; Nunkoosing, 2005; Potter & Hepburn, 2005). The chosen sample size for example, though consisting of 12 XAI experts, may present some limitations in terms of representativeness. With such a limited sample, it might not fully represent the diversity of experiences, backgrounds, and perspectives among XAI experts. This lack of representation could lead to selection bias and limit the generalizability of the results. The interpretative nature of interviews also presents some inherent challenges. Interviewer bias could potentially influence the process, as the interviewer's preconceptions could affect the manner in which questions are asked and responses are interpreted. Similarly, interviewees may also adjust their responses due to social desirability bias, seeking to present themselves favourably rather than being entirely truthful. These factors could consequently skew the overall results.

In addition, the dependence on memory in interviews may contribute to inaccuracies in the collected data. When asked to recall past experiences or decisions, interviewees might unintentionally provide inaccurate or biased information. Furthermore, the process relies heavily on correct understanding and interpretation of the questions and responses, which can potentially lead to misunderstandings or misinterpretations. Additionally, depending on the degree of anonymity maintained during the interviews, the candidness of interviewees' responses might be affected, leading to further potential bias in the results. This element of disclosure could impact the openness and honesty of their responses (Creswell & Poth, 2018).

Given these potential shortcomings, it is important to interpret the results from these interviews with an awareness of these potential biases and limitations. However, to address the potential limitations in the interview process, several strategic steps were taken. First, to minimize the issue of recall bias and to ensure the accuracy of responses, all interviews were recorded with the permission of the interviewees. This allowed for subsequent detailed analysis and minimized the chance of misinterpretation or misunderstanding of the responses (Creswell & Poth, 2018). In order to alleviate selection bias and improve the representativeness of the data, the recruitment strategy was aimed at securing as many interviewees as possible from various backgrounds within the field of XAI. Despite the limited sample size of 12, efforts were made to reach out to a diverse group of experts. Moreover, to further mitigate social desirability bias and to enhance the openness of responses, all interviewees were assured of the confidentiality of their responses. It was made clear that their participation was voluntary, and they could withdraw from the process at any time. Furthermore, a few days before each individual interview, the questions were sent to the participants. This approach was adopted to give interviewees adequate time to reflect on their experiences and formulate their thoughts, thereby facilitating more thoughtful and comprehensive responses during the actual interview (Creswell & Poth, 2018). In addition, throughout the interview process, conscious efforts were made to limit the influence of the interviewer's biases. The interview questions were designed to be neutral, open-ended, and unbiased to elicit true perspectives and experiences from the participants.

Despite these efforts, it is recognized that no research method is entirely free from limitations. However, these measures have been undertaken to limit the impact of the potential shortcomings, thereby enhancing the reliability and validity of the findings from these interviews. Future research should consider ways to further mitigate the potential issues, such as further increasing the sample size whilst ensuring a diverse sample and using more elaborate methods to reduce bias.

9.1.7. Survey shortcomings

The relationships between comprehensiveness, simplicity, transparency, complexity, clarity and generalizability should be explored in more detail. The survey from chapter 8 only explored two dimensions of this intricate relationship. The assumption was made that switching the weights of comprehensiveness and simplicity would be sufficient to cover all connections between the six factors, simply because it ensured that the average weights would comply with the conclusions from the survey as can be seen in Table 8. However, this methodology has its limitations. It is important to acknowledge that the approach to forming the new weights and the assumptions underlying it may be considered subjective. Other researchers could feasibly adopt different strategies to investigate the relationships between these factors, and their approaches would be equally valid. The inherent subjectivity in this kind of analysis underlines the need for further research. Future studies could benefit from exploring these relationships from different angles, using varied weighting systems, or prioritizing different factors based on their research contexts or objectives. Such a broadened perspective would undoubtedly enrich our understanding of the intricate dynamics between these important elements of interpretability.

Besides the contents of the survey, the demographic distribution of the survey also shows shortcomings (see appendix H). The most notable limitation is the geographical and ethnical distribution of the survey participants. The survey was primarily distributed in the UK, the Netherlands, and the US, resulting in a majority of Caucasian participants (at least 75%). This skewed representation may have influenced the preferences and perspectives of the participants regarding the trade-offs between interpretability factors. Consequently, the derived key principles may not be fully generalizable to other cultural and demographic contexts. Future research should aim to incorporate more diverse and inclusive samples to better understand the role of culture and diversity in shaping interpretability preferences and requirements in XAI systems. By acknowledging and addressing these limitations, researchers can further refine the interpretability framework and key principles, ensuring that they are more inclusive, comprehensive, and applicable to a broader range of users and contexts.

9.1.8. Other recommendations for future research

While the refined interpretability framework and key principles offer valuable guidance for XAI research and practice, several areas warrant further exploration. Interesting future research that comes to mind is for example understanding the role of culture and diversity in interpretability or investigating the impact of emerging AI technologies on the need for explainability. Additionally, time-series studies could explore the evolution of interpretability requirements as users gain experience with AI systems, and the relationship between interpretability and other ethical dimensions of AI, such as fairness and privacy, could be explored.

9.2. Ethical note

As a concluding ethical reflection on this research, it seems appropriate to address the complex and sensitive nature of XAI in general. The first two paragraphs of chapter 1 address the problems of AI and implicate how XAI can solve those problems. However, there are also problems of AI that cannot be solved by making AI explainable. Overreliance on (X)AI is one example of this. It refers to the possibility that users might place excessive trust in AI systems and their explanations, leading to a reduced inclination to engage in critical thinking, scepticism, or human intervention. This overreliance can manifest itself in several ways. The

most logical manifestation of overreliance is inadequate scepticism. Inadequate scepticism presents itself insofar that users may accept the provided explanations without questioning their validity, even when the explanations might be inaccurate or misleading. This can lead to a lack of critical assessment of the AI system's outputs, resulting in suboptimal decision-making. Other problems of XAI revolve around diminished human responsibility, loss of human expertise, and excessive trust. Diminished human responsibility happens when users trust AI systems and their explanations too much, they might absolve themselves of responsibility for decisions made with the help of AI. This can therefore still lead to a lack of accountability and a reduced sense of ownership over the outcomes of those decisions, whilst XAI should ensure a higher degree of accountability and ownership of the outcomes. Loss of human expertise; overreliance on XAI might lead to a decline in human expertise in certain domains, as users become increasingly dependent on AI systems for decision-making. This could make it more difficult for users to recognize when the AI system is making an error or to intervene effectively when necessary. Finally, excessive trust; this is when users may be more inclined to trust AI systems that provide explanations, even when the explanations are not entirely accurate or relevant. This can lead to overconfidence in the system's capabilities and a reduced inclination to seek alternative information sources or human input.

In the paper "Against Explainable AI: The Case for Pragmatic AI" by Tim Miller, the author argues that explainable AI, as currently researched and developed, might not be the best approach to enhancing trust and enabling users to understand AI systems. Instead, Miller proposes a shift towards a more pragmatic approach to AI, focusing on system behaviour and user interaction rather than providing detailed explanations: evaluative AI. Miller's perspective highlights the potential pitfalls of overreliance on XAI, emphasizing that explanations might not be sufficient to ensure responsible AI usage. It is important to consider the user's context, needs, and expertise when designing AI systems, and focus on fostering a meaningful human-AI interaction that encourages critical thinking and appropriate trust, rather than simply providing explanations for AI decision-making (Miller, 2023).

Chapter 10. Conclusion

In this final chapter, a comprehensive conclusion on the influential factors that determine XAI interpretability is provided. Throughout this thesis, this has been done based on the following main research question:

"How can XAI developers assess to what extent XAI is interpretable to laypeople?"

Throughout the research, the intricate relationship between various factors and their respective impacts on XAI interpretability have been examined. By analysing all relevant literature, conducting numerous interesting interviews, evaluating the framework via the survey distributed amongst laypeople, and synthesizing all findings, insights have been uncovered that contribute to a deeper understanding of the complex dynamics governing XAI interpretability. First, exemplary actionable guidelines based on the key principles will be presented. Second, we will look at the areas of impact that this thesis contributes to.

10.1. From key principles to exemplary actionable guidelines

Understanding the factors influencing XAI interpretability is complex. While the framework itself and the key principles offer comprehensive insight into these factors, their direct application in real-world scenarios might not always be clear-cut. This is primarily because principles, in their richness and depth, may not straightforwardly translate into practical steps for XAI developers. Therefore, bridging this gap between theory and practice was crucial, thus leading to the development of actionable guidelines in this final chapter.

The process of transforming theoretical key principles into practical, actionable guidelines was both methodical and careful. First, each principle was thoroughly reviewed and individually analysed. The core elements within each principle, which could translate into practical actions, were identified. The objective was to distil the essence of the principle into simple, actionable steps while preserving its original intent and depth.

Subsequently, these core elements were refined and transformed into specific actions. These proposed actions were designed to be clear, practicable, and adaptable across a variety of contexts and scenarios.

One critical aspect of this methodology was acknowledging the potential trade-offs associated with applying the principles. To navigate this, the process of forming actionable guidelines took these trade-offs into account, offering practical ways to handle them.

The resulting actionable guidelines thus represent a user-friendly version of the original theoretical principles. They maintain the substance of these principles, while ensuring they are accessible and practically applicable. This methodical process of transition enhances the relevance of the key principles, translating them into a set of practical tools that developers can apply directly in the design and implementation of XAI systems.

Furthermore, these guidelines simplify the decision-making process for various stakeholders in the XAI space, enabling them to navigate trade-offs and prioritize factors critical for interpretability. Ultimately, this shift from key principles to actionable guidelines bridges the gap between XAI theory and real-world implementation, ensuring the insights gained from the theoretical framework are effectively utilized in practice.

In the following section, we present the key principles as a foundation to derive these actionable guidelines:

1. Trustworthiness and relevance:
    a. Design explanations that are open about their limitations and uncertainties to increase trustworthiness.
    b. Ensure explanations are relevant by addressing the user's goals and objectives, and customize explanations based on the user's context.

2. Balance simplicity, clarity, and generalizability with comprehensiveness, complexity, and transparency:
    a. If possible, use visualizations and analogies to simplify complex concepts.
    b. Strive to employ plain language and avoid technical jargon.
    c. Provide different levels of detail to accommodate the user's needs and preferences.

3. Affordance, coherence with prior beliefs, and addressing abnormalities:
    a. Utilize familiar concepts, terms, and formats when presenting explanations.
    b. Validate the user's prior beliefs when possible and provide evidence when challenging them.
    c. Highlight abnormalities only when necessary and explain their significance clearly.

4. Intentionality and actionability balance:
    a. Explain the reasons behind the model's decision-making process.
    b. Offer actionable recommendations or next steps users can take based on the explanation.

5. Fidelity prioritization (explanation/model):
    a. Focus on creating explanations that are meaningful and contextually relevant.
    b. Strive for high model fidelity but prioritize explanation fidelity when trade-offs are necessary.

6. Comprehensive optimization:
    a. Consider all factors in the framework when designing explanations to maximize interpretability.
    b. Continuously evaluate and iterate on the explanation design based on user feedback and evaluations.

7. Contextualization
    a. Account for the user's domain knowledge, goals, application, and cultural context when designing explanations.
    b. Consider offering customizable explanations to better fit users' unique contexts.

8. Evaluation:
    a. Make use of quantitative metrics, such as recall response time and information transfer rate, to assess the interpretability of explanations.
    b. Regularly gather qualitative user feedback to identify areas of improvement and inform design decisions.

9. Stakeholder engagement and ethical considerations:
    a. Involve stakeholders throughout the development process (including the design phase) to understand their needs, expectations, and concerns.
    b. Address ethical issues, such as bias and overreliance on AI, during the design process.
    c. Be transparent about any trade-offs made in the pursuit of interpretability.

By following these guidelines, XAI developers can create more interpretable and useful explanations that address the unique needs of different users and contexts. This will ultimately help foster trust in AI systems and promote their responsible and ethical use across various industries and sectors.

10.2. Comparison with literature

This study on Explainable Artificial Intelligence (XAI) brings forth significant contributions to the current body of literature as it is now. All key principles will be discussed in this section and compared to the literature. By emphasizing trustworthiness and relevance in the first key principle, the research aligns with Ribeiro et al.'s assertion on the importance of trust in interpretable machine learning (Ribeiro et al., 2016). However, the distinctive delineation of the balance between trustworthiness of the explanation and the user's trust augments the practical application of XAI in real-world scenarios.

The second key principle underlines the importance of balancing simplicity, clarity, and generalizability with comprehensiveness, complexity, and transparency. This echoes Lipton's argument for simplicity and comprehensibility in model interpretability (Lipton, 2017). The research extends this by emphasizing the nuanced understanding of the trade-offs involved, enriching the dialogue in the process.

The discussion around affordance, coherence with prior beliefs, and addressing abnormalities in the third principle introduces a fresh perspective to the field. Miller discusses the importance of alignment with human intuition and knowledge (Miller, 2019). However, by focusing on the implications of emphasizing abnormalities, this research provides a novel angle, thereby filling a gap in the current literature.

The fourth principle's emphasis on intentionality and actionability resonates with Doshi-Velez and Kim and their assertions on the necessity of reasons and recommendations in explanations (Doshi-Velez & Kim, 2017). The added value here lies in detailing the careful balance between these two components to generate comprehensible explanations for laypeople.

Principle five accentuates the importance of explanation and model fidelity in XAI. This is a critical issue discussed by Gilpin et al. (2019). By highlighting the potential trade-off between these two elements and advocating for explanation fidelity, this research provides more contextually meaningful explanations to laypeople, enriching the current discourse.

The holistic approach highlighted in the sixth principle, which prioritizes optimizing all factors of the explanation, is a progression of the all-encompassing approach recommended by Lundberg and Lee (2017). By emphasizing the highest level of interpretability possible, the application of this principle is broadened to real-world XAI systems.

The significance of contextual factors in XAI interpretability forms the crux of the seventh principle. While it aligns with Holzinger et al.'s argument on the importance of the context in human-AI interaction, this study provides a deeper understanding of these factors by acknowledging the unique contexts of each user (Holzinger et al., 2019).

The eighth principle recommends evaluating interpretability through both measurable and perceived factors, resonating with Murdoch et al. and their emphasis on the value of evaluations in model interpretability (Murdoch et al., 2019). The emphasis on recall response time and information transfer rate as quantitative metrics introduces a novel approach to interpretability evaluation.

Lastly, the ninth principle reinforces the importance of stakeholder engagement and ethical considerations. This echoes the ethical AI guidelines proposed by Floridi et al. (2018). By

directly linking these considerations to interpretability and offering practical ways for designers to incorporate them into their XAI systems, this research offers a new dimension to the field.

Overall, this study enriches the existing body of literature by demonstrating and emphasizing that the identified key principles are not only central for designing interpretable XAI systems, but are also highly applicable to laypeople concerning the domain in which the XAI operates. The research underscores the universal relevance of these principles and reinforces their role in enhancing the interpretability of XAI systems across different domains.

10.3. Scientific implications and relevance

The pivotal outcomes of this research are embodied in the final interpretability framework and the key principles, as outlined in sections 8.6 and 8.7, respectively. The depth of these findings signals a profound understanding of the factors that determine the interpretability of XAI explanations. Their potential impact radiates across the scientific community and society, illuminating diverse aspects of XAI research, system design, and real-world applications. This section will elaborate on the broader implications of this research, underscoring its substantial scientific relevance. The scientific implications of this study are multifold, with the potential to inspire future research, inform system design, and influence algorithm development.

10.3.1. Advancing XAI research and system design

The refined interpretability framework and key principles offer a valuable foundation for both novice and experienced XAI researchers and developers, based on, among other things, the discussions with interviewees 1, 2, 3, 5, 9, 11 and 12. By providing a systematic and structured approach to understanding the various factors influencing interpretability, the framework serves as a starting point for discussions and encourages a more methodical approach to XAI system design. The key principles offer practical guidance for navigating trade-offs and prioritizing factors when designing and evaluating explanations.

10.3.2. Informing comparative studies and benchmarking

The framework can also serve as a tool for comparative studies and the creation of benchmarks for different XAI models and techniques. This facilitates standardized assessment of XAI interpretability, thus contributing to the progression of the field.

10.3.3. Fostering Interdisciplinary Research

The findings from this research furthermore emphasize the importance of context dependence, human-centric approaches, and stakeholder engagement in the design and development of XAI systems. This opens up opportunities for interdisciplinary research and collaboration between experts in artificial intelligence, human-computer interaction, psychology, philosophy, and other fields. By bringing together diverse perspectives, researchers can create more effective, meaningful, and useful explanations that address the unique needs of different users and contexts.

10.4. Societal Relevance

This section will elaborate on the broader implications of this research, underscoring its substantial societal relevance. The societal implications of this research range from enhancing accessibility and inclusion to influencing policy and regulation.

10.4.1. Enhancing trust and adoption of AI systems

The implications of this research extend beyond academia, influencing the adoption and acceptance of AI systems in various industries and sectors. By developing explanations that are more understandable, satisfying, and useful to users, XAI systems can foster trust and accountability, ultimately promoting the responsible and ethical use of AI technologies. This could have a significant impact on the adoption of AI systems in various sectors such as healthcare, finance, transportation, and other domains where trust is crucial for successful implementation.

10.4.2. Empowering consumers and encouraging informed decision making
The principles and guidelines developed in this research could empower individuals to make more informed decisions when interacting with AI systems, particularly in critical sectors like healthcare or finance.

10.4.3. Promoting accessibility and inclusion
By emphasizing interpretability for laypeople, this research supports the development of AI systems that are accessible to a wide range of individuals, promoting equal access to AI technologies regardless of their technical knowledge, domain knowledge, or background.

10.4.4. Informing policy and regulation
Lastly, the insights gained from this research can also inform the development of policies and regulations governing AI technologies. As governments and organizations worldwide grapple with the ethical and societal implications of AI, the interpretability framework and key principles offer a foundation for understanding and evaluating explainable AI systems. This can support the creation of more informed and effective policies, guidelines, and standards that promote responsible AI practices.

In highlighting both the scientific and societal relevance of this study, the far-reaching potential of this research is emphasized. It not only deepens academic understanding of XAI interpretability but also seeks to bridge the gap between AI technology and its real-world applications, hence encouraging more responsible AI usage in society.

10.5. Relevance to Complex Systems Engineering and Management
The thesis is intrinsically aligned with the ethos of the master's program Complex Systems Engineering and Management at TU Delft, which champions an interdisciplinary approach to understanding and managing intricate systems. This alignment is most evidently reflected in the multi-disciplinary nature of the research, which beautifully unites principles from computer science, cognitive psychology, and philosophy to tackle the complex challenge of enhancing XAI interpretability.

At its core, this research is about managing the increasing complexity in the realm of AI, specifically XAI. The key principles and guidelines developed through this research offer a structured approach to navigating the complexities of XAI. By harnessing these insights, we can design XAI systems that are not just more interpretable and user-friendly, but also better integrated within the social fabric. Moreover, the research echoes the program's emphasis on system design. The crafting of an interpretability framework and the subsequent translation of theoretical principles into actionable guidelines showcase a comprehensive approach to system design, resonating with the program's focus on engineering complex systems. In keeping with the program's commitment to inclusive and participatory approaches to managing intricate systems, this research also lays considerable emphasis on stakeholder engagement. It underscores the importance of understanding user needs and preferences, and advocates for active stakeholder involvement in the design and evaluation process. Mirroring the program's dedication to the responsible management of complex systems, the research also foregrounds the ethical dimensions of (X)AI. It acknowledges potential ethical dilemmas surrounding AI and presents strategies to ensure the responsible and trustworthy application of AI technologies. Finally, like the program's orientation towards preparing students for managing future complex systems, this research is inherently forward-looking. It tackles a pivotal challenge in the contemporary AI landscape and, in doing so, lays the foundation for making AI more transparent, understandable, and trustworthy, better-equipping society to navigate the AI-driven future.

Thus, this research embodies the objectives and values of the Complex Systems Engineering and Management program. It exemplifies an interdisciplinary approach to complex problems,

delivers valuable insights for system design and stakeholder engagement, upholds ethical and responsible AI practices, and provides a steppingstone towards future challenges in XAI.

10.6. Reflection on Design Science Research Methodology

In reflecting on the scientific approach of this study, the application of the Design Science Research Methodology (DSRM) has demonstrated its value and relevance. The study followed the iterative, problem-solving paradigm intrinsic to DSRM: identifying a problem, designing and developing an artefact (in this case, a framework with key principles for XAI interpretability), and then evaluating this artefact.

The problem identification phase was crucial in this study to understand the challenges with current XAI interpretability, especially regarding laypeople's needs. Rigorous literature reviews have helped to delineate these challenges and set the stage for the development of the XAI interpretability framework and key principles. The design and development phase involved a synthesis of the framework and key principles for XAI interpretability based on extant literature. This process demonstrated the iterative and adaptive nature of DSRM, as the framework continually evolved to integrate new insights and feedback. The first evaluation phase consisted of twelve XAI expert interviews, greatly reshaping the deliverable. The second evaluation phase saw the application of the framework to the specific domain of medical XAI in a use-case-based survey distributed over around 200 people, gauging its effectiveness in enhancing interpretability. It included both objective and perceived interpretability metrics, underscoring the multi-dimensional approach to evaluation advocated by DSRM. The communication of results, the final stage of DSRM, was fulfilled through the drafting of this thesis, contributing to the academic discourse around XAI and its interpretability.

In conclusion, the DSRM provided a robust and flexible methodological foundation for this research. Its problem-oriented and iterative nature guided the development of a meaningful and applicable artefact - the XAI interpretability framework and its key principles - demonstrating its value in the realm of AI research. Future research can take this framework further, refining and expanding it in different contexts following the iterative spirit of DSRM.

10.7. Closing remarks

In conclusion, this thesis has sought to shed light on the factors that influence the interpretability of XAI for laypeople, and the findings point to a multi-faceted and complex interplay of these factors. The proposed framework and the derived key principles have wide-ranging implications, from advancing XAI research and system design, fostering interdisciplinary research, and enhancing trust and adoption of AI systems, to informing policy and regulation. The actionable guidelines derived from these principles offer a pragmatic approach for XAI developers to enhance interpretability, thereby fostering greater trust and responsible use of AI technologies. As we continue to navigate the AI-driven era, the dynamics of XAI interpretability will evolve, warranting ongoing research and discourse. Therefore, this thesis should be seen as a stepping stone towards a more comprehensive understanding of XAI interpretability, and I hope it serves as a catalyst for further exploration and discussion in this growing field.

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aithal, A., & Aithal, P. S. (2020). *Development and Validation of Survey Questionnaire & Experimental Data – A Systematical Review-based Statistical Approach*. https://doi.org/10.5281/ZENODO.4179499

Anjomshoae, S. (2022). *Context-based explanations for machine learning predictions*.

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & 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), Article 11. https://doi.org/10.3390/app11115088

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bader, V., & Kaiser, S. (2019). Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence. *Organization*, *26*(5), 655–672. https://doi.org/10.1177/1350508419855714

Bellman, R. (1978). *An introduction to artificial intelligence: Can computers think?* Boyd & Fraser Pub. Co.

Berthoz, A. (2012). *Simplexity: Simplifying Principles for a Complex World* (G. Weiss, Trans.). Yale University Press.

Bibal, A., & Frénay, B. (2016, April 27). *Interpretability of Machine Learning Models and Representations: An Introduction*.

Biran, O., & Cotton, C. V. (2017). *Explanation and Justification in Machine Learning: A Survey*. https://www.semanticscholar.org/paper/Explanation-and-Justification-in-Machine-Learning-%3A-Biran-Cotton/02e2e79a77d8aabc1af1900ac80ceebac20abde4

Bohanec, M., Kljajić Borštnar, M., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications*, *71*, 416–428. https://doi.org/10.1016/j.eswa.2016.11.010

Bohanec, M., Robnik-Šikonja, M., & Kljajić Borštnar, M. (2017). Decision-making framework with double-loop learning through interpretable black-box machine learning models. *Industrial Management & Data Systems*, *117*(7), 1389–1406. https://doi.org/10.1108/IMDS-09-2016-0409

Bolam, B., Gleeson, K., & Murphy, S. (2003). 'Lay Person' or 'Health Expert'? Exploring Theoretical and Practical Aspects of Reflexivity in Qualitative Health Research. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *Vol 4*, No 2 (2003): Subjectivity and Reflexivity in Qualitative Research II. https://doi.org/10.17169/FQS-4.2.699

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20. https://doi.org/10.1038/538020a

Charniak, E., & McDermott, D. (1985). *Introduction to artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc.

Chromik, M. (2021). Making SHAP Rap: Bridging Local and Global Insights Through Interaction and Narratives. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, & K. Inkpen (Eds.), *Human-Computer Interaction – INTERACT 2021* (Vol. 12933, pp. 641–651). Springer International Publishing. https://doi.org/10.1007/978-3-030-85616-8_37

Cooper, A. F., Laufer, B., Moss, E., & Nissenbaum, H. (2022). Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. *ArXiv:2202.05338 [Cs]*. http://arxiv.org/abs/2202.05338

Creswell, J. W., & Poth, C. N. (2017). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*.

DARPA. (2016). *Explainable Artificial Intelligence (XAI)* (Broad Agency Announcement DARPA-BAA-16-53; p. 52). DARPA. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

Deeploy. (2023). *Deeploy | Making Machine Learning Explainable*. Deeploy. https://www.deeploy.ml/

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* (arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

Duignan, B. (2023, March 31). *Occam's razor | Origin, Examples, & Facts | Britannica*. https://www.britannica.com/topic/Occams-razor

European Commission. (2021). *Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*.

Ferrara, D. (2016). Self-Driving Cars: Whose Fault Is It? *Georgetown Law Technology Review*, *1*, 182.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. Scopus. https://doi.org/10.1037/h0057532

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fraunhofer FOKUS. (2023). *Fraunhofer FOKUS | About us – We connect everything*. https://www.fokus.fraunhofer.de/en/fokus/about_fokus

Ganzach, Y. (1994). Theory and configurality in expert and layperson judgment. *Journal of Applied Psychology*, *79*(3), 439–448. https://doi.org/10.1037/0021-9010.79.3.439

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). *Explaining Explanations: An Overview of Interpretability of Machine Learning* (arXiv:1806.00069). arXiv. http://arxiv.org/abs/1806.00069

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 93:1-93:42. https://doi.org/10.1145/3236009

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Gunning, R. (1969). The fog index after twenty years. *Journal of Business Communication*, *6*(2), 3–13. Scopus. https://doi.org/10.1177/002194366900600202

Hart, H. L. A., & Honoré, T. (1985). *Causation in the Law*. Oxford University Press UK.

Hasa. (2020, March 2). *What is the Difference Between Positivism and Constructivism*. Pediaa.Com. https://pediaa.com/what-is-the-difference-between-positivism-and-constructivism/

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge: MIT Press.

Hilton, D. J. (1996). Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance. *Thinking & Reasoning*, *2*(4), 273–308. https://doi.org/10.1080/135467896394447

Hilton, D. J., McClure, J. L., & Slugoski, B. R. (2005). The Course of Events: Counterfactuals, Causal Sequences and Explanation. In *The Psychology of Counterfactual Thinking*. Routledge.

Hofisi, C., Hofisi, M., & Mago, S. (2014). Critiquing Interviewing as a Data Collection Method. *Mediterranean Journal of Social Sciences*. https://doi.org/10.5901/mjss.2014.v5n16p60

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, *9*(4), e1312. https://doi.org/10.1002/widm.1312

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, *51*, 141–154. https://doi.org/10.1016/j.dss.2010.12.003

IBM. (2020, June 3). *What is Artificial Intelligence (AI)?* https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How can I choose an explainer? An Application-grounded Evaluation of Post-hoc

Explanations. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 805–815. https://doi.org/10.1145/3442188.3445941

Jin, W., Fan, J., Gromala, D., Pasquier, P., & Hamarneh, G. (2022). *EUCA: The End-User-Centered Explainable AI Framework* (arXiv:2102.02437). arXiv. https://doi.org/10.48550/arXiv.2102.02437

Josephson, J. R., & Josephson, S. G. (1996). *Abductive Inference: Computation, Philosophy, Technology*. Cambridge University Press.

Kurzweil, R. (1992, January 30). The Age of Intelligent Machines. *MIT Press*. https://mitpress.mit.edu/9780262610797/the-age-of-intelligent-machines/

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable Decision Sets: A Joint Framework for Description and Prediction. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684. https://doi.org/10.1145/2939672.2939874

Lewis, D. (1986). Causal Explanation. In D. Lewis (Ed.), *Philosophical Papers Vol. Ii* (pp. 214–240). Oxford University Press.

Liao, Q. V., Singh, M., Zhang, Y., & Bellamy, R. K. E. (2020). Introduction to Explainable AI. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–4. https://doi.org/10.1145/3334480.3375044

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. Scopus. https://doi.org/10.1109/18.61115

Lin, Y.-S., Lee, W.-C., & Celik, Z. B. (2021). What Do You See?: Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1027–1035. https://doi.org/10.1145/3447548.3467213

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Lipton, Z. C. (2017). *The Mythos of Model Interpretability* (arXiv:1606.03490). arXiv. https://doi.org/10.48550/arXiv.1606.03490

LMU Munich. (2023). *Home—LMU Munich*. https://www.lmu.de/en/

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*, 10. https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Mackenzie, J. (2011). Positivism and Constructivism, Truth and 'Truth'. *Educational Philosophy and Theory*, *43*(5), 534–546. https://doi.org/10.1111/j.1469-5812.2010.00676.x

Microsoft. (2023). About Microsoft Research. *Microsoft Research*. https://www.microsoft.com/en-us/research/about-microsoft-research/

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T. (2023). *Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support* (arXiv:2302.12389). arXiv. http://arxiv.org/abs/2302.12389

Miller, T., Howe, P., & Sonenberg, L. (2017). *Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*.

Mohseni, S., Zarei, N., & Ragan, E. D. (2020). *A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems* (arXiv:1811.11839). arXiv. http://arxiv.org/abs/1811.11839

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-Wise Relevance Propagation: An Overview. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 193–209). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_10

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Newman, D. (2014, April 22). *Experts May Have Influence, But What Makes An Expert?* Forbes. https://www.forbes.com/sites/danielnewman/2014/04/22/experts-may-have-influence-but-what-makes-an-expert/

Nilsson, N. J. (1998). *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann Publishers Inc.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, *2*(1), 25–42. https://doi.org/10.1007/BF02639315

Nunkoosing, K. (2005). The Problems With Interviews. *Qualitative Health Research*, *15*(5), 698–706. https://doi.org/10.1177/1049732304273903

Overton, J. (2012). Explanation in Science. *Electronic Thesis and Dissertation Repository*. https://ir.lib.uwo.ca/etd/594

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect* (1st ed.). Basic Books, Inc.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Poole, D., Mackworth, A. K., & Goebel, R. (1998). Computational intelligence: A logical approach. *Choice Reviews Online*, *35*(10), 35-5701-35–5701. https://doi.org/10.5860/CHOICE.35-5701

Potter, J., & Hepburn, A. (2005). Qualitative interviews in psychology: Problems and possibilities. *Qualitative Research in Psychology*, *2*(4), 281–307. https://doi.org/10.1191/1478088705qp045oa

Ranney, M., & Thagard, P. (1988, July 1). *Explanatory Coherence and Belief Revision in Naive Physics:* https://doi.org/10.21236/ADA201093

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*(3), 429–447. https://doi.org/10.1037/0022-3514.65.3.429

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *'Why Should I Trust You?': Explaining the Predictions of Any Classifier* (arXiv:1602.04938). arXiv. http://arxiv.org/abs/1602.04938

Rich, E., & Knight, K. (1991). *Artificial Intelligence* (2nd edition). Mcgraw Hill Higher Education.

Robnik-Šikonja, M., & Kononenko, I. (2008). Explaining Classifications For Individual Instances. *IEEE Transactions on Knowledge and Data Engineering*, *20*(5), 589–600. https://doi.org/10.1109/TKDE.2007.190734

RSM. (2023). *About RSM - Rotterdam School of Management, Erasmus University*. https://www.rsm.nl/about-rsm/

Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (Fourth edition. Global edition). Pearson.

Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700, pp. 5–22). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_1

Schmidt, P., & Biessmann, F. (2019). *Quantifying Interpretability and Trust in Machine Learning Systems* (arXiv:1901.08558). arXiv. https://doi.org/10.48550/arXiv.1901.08558

Shevskaya, N. V. (2021). Explainable Artificial Intelligence Approaches: Challenges and Perspectives. *2021 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS)*, 540–543. https://doi.org/10.1109/ITQMIS53292.2021.9642869

Symptomate. (2022). *Check your symptoms online*. https://symptomate.com/

Tesla. (2022). *Autopilot and Full Self-Driving Capability*. https://www.tesla.com/support/autopilot

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467. https://doi.org/10.1017/S0140525X00057046

TNO. (2023). *About us | TNO*. Tno.Nl/En. https://www.tno.nl/en/about-tno/

Todd, P. M., & Gigerenzer, G. (2000). Précis of Simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*(5), 727–741. https://doi.org/10.1017/S0140525X00003447

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). *Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems* (arXiv:1806.07552). arXiv. https://doi.org/10.48550/arXiv.1806.07552

TU Delft. (2023). *About TU Delft*. TU Delft. https://www.tudelft.nl/en/about-tu-delft

TU Dublin. (2023). *About the University | TU Dublin*. https://tudublin.ie/explore/about-the-university/

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, New Series*, *59*(236), 433–460.

University of Hildesheim. (2023). *University of Hildesheim* [Text]. Universität Hildesheim. https://www.uni-hildesheim.de/en/

University of Melbourne. (2023). *The University of Melbourne, Australia—Australia's best university and one of the world's finest*. The University of Melbourne; The University of Melbourne. https://www.unimelb.edu.au/

Vilone, G., & Longo, L. (2020). *Explainable Artificial Intelligence: A Systematic Review* (arXiv:2006.00093). arXiv. http://arxiv.org/abs/2006.00093

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

Wee, B. V., & Banister, D. (2016). How to Write a Literature Review Paper? *Transport Reviews*, *36*(2), 278–288. https://doi.org/10.1080/01441647.2015.1065456

Winston, P. H. (1992). *Artificial intelligence* (3rd ed). Addison-Wesley.

Appendix A. Research Flow Diagram

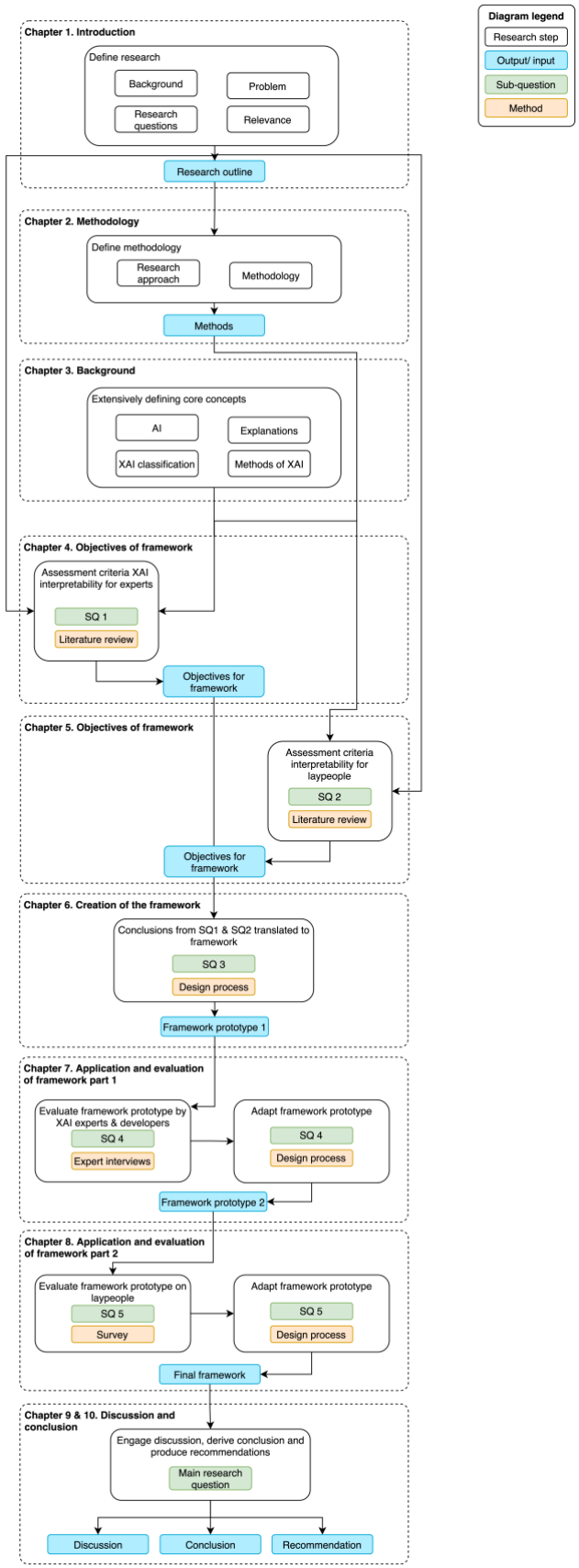

Figure 18. Research Flow Diagram

Appendix B. Gantt chart

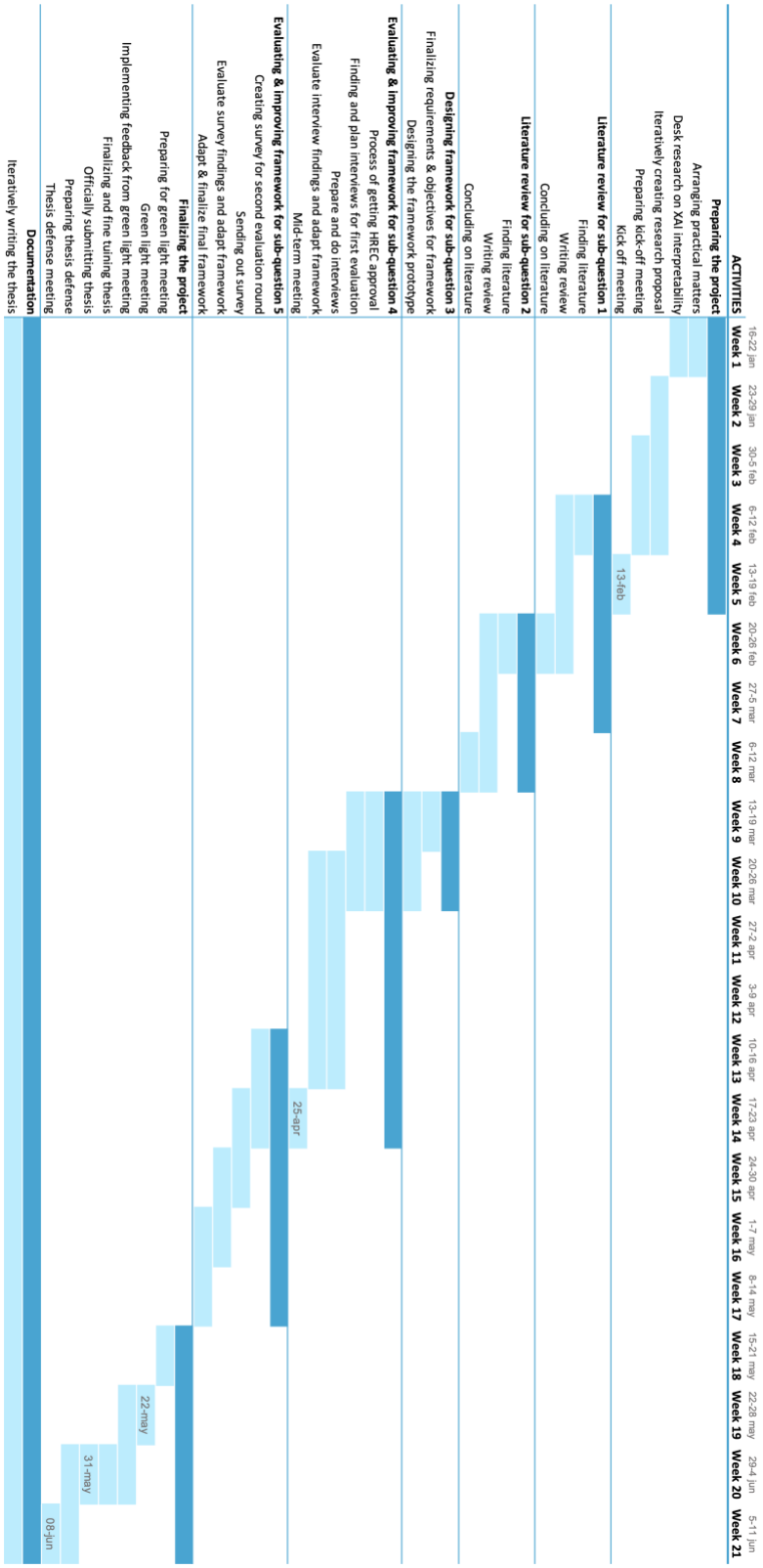

Figure 19. Gantt chart

Appendix C. Interviewee institutions

In this appendix, more extensive descriptions of the institutions from which the interviewed experts and developers in the XAI field are affiliated can be found. The information presented here has been retrieved from the respective institutions' official websites. While these descriptions offer valuable insights into the backgrounds and expertise of the interviewees, it is important to note that they may contain a certain degree of marketing bias or promotional language. However, they serve as a useful reference to gain a deeper understanding of the institutions and their contributions to the field of XAI.

- **Deeploy**. Deeploy is a Dutch software company based in Utrecht that provides a cloud-based infrastructure automation platform for developers and IT teams. Deeploy creates software to enable interaction between humans and machine learning models. Their software makes machine learning deployments manageable, accountable, and efficient. As the 'punchline' of Deeploy is as follows: *making machine learning explainable,* this seems like an excellent fit for the purpose of evaluating the first framework prototype (Deeploy, 2023).
- **Fraunhofer FOKUS**. Fraunhofer FOKUS is a German research organization that focuses on the impact of digital transformation on our economy, technology, and society. The organization has been providing research services to commercial enterprises and public administrations since 1988 to support them in shaping and implementing digital transformation. Fraunhofer FOKUS offers a wide range of research services, including requirements analysis, consulting, feasibility studies, technology development, prototypes, and pilots. These research services are provided in various business segments, such as Digital Public Services, Future Applications and Media, Quality Engineering, Smart Mobility, Software-based Networks, Networked Security, Visual Computing, and Analytics (Fraunhofer FOKUS, 2023).
- **LMU Munich.** Ludwig Maximilian University of Munich is a public research university located in Munich, Germany. It is one of the oldest and most prestigious universities in Germany, founded in 1472 by Duke Ludwig IX of Bavaria-Landshut. The university offers a wide range of academic programs, including over 150 undergraduate and graduate degree programs across 18 faculties. LMU Munich is known for its excellence in research and teaching, and it is ranked among the top universities in Europe and the world. The university has a diverse and international student body, with over 50,000 students from around the world (LMU Munich, 2023).
- **Microsoft.** Microsoft Corporation is a multinational technology company that develops, licenses, and sells computer software, consumer electronics, and personal computers. Microsoft's best-known software products are the Windows operating system, Microsoft Office Suite, and the Internet Explorer and Microsoft Edge web browsers. In addition to its software products, Microsoft also produces hardware devices, including the Xbox video game console, the Surface tablet and laptop, and other devices such as the Microsoft Band and the HoloLens mixed reality headset. Microsoft offers a wide range of services for consumers and businesses, including the Microsoft Azure cloud computing platform, Bing search engine, Microsoft Dynamics business solutions, and the LinkedIn social network for professionals. The company has also made significant investments in AI and machine learning technologies, which are integrated into many of its products and services. Overall, Microsoft is a diverse technology company with a strong focus on innovation and pushing the boundaries of what's possible with technology (Microsoft, 2023).
- **RSM**. RSM stands for Rotterdam School of Management, which is the business school of Erasmus University Rotterdam (EUR) in the Netherlands. RSM offers a wide range of undergraduate, graduate, and executive education programs in business and management. It is known for its high-quality education, research, and international orientation, with a focus on developing leaders who can make a positive impact on

society. RSM is accredited by the AACSB (Association to Advance Collegiate Schools of Business), EQUIS (European Quality Improvement System), and AMBA (Association of MBAs), which are prestigious international accreditations for business schools (RSM, 2023).

- **TNO**. TNO stands for 'Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek', which translates to Dutch Organization for Applied Scientific Research in English. It is an independent research organization based in the Netherlands that focuses on applied research in various areas, including technology, sustainability, defence and security, energy, and health. TNO conducts research in collaboration with government agencies, research institutes, and private sector organizations, with a goal to create innovative solutions and contribute to the development of a sustainable and competitive society. The organization has a wide range of expertise, including engineering, physics, chemistry, social sciences, and humanities. Overall, TNO aims to apply scientific knowledge and research to address societal challenges and create value for the Dutch economy and society (TNO, 2023).
- **TU Delft**. TU Delft stands for Delft University of Technology, which is a leading Dutch public technical university located in Delft, Netherlands. It was founded in 1842 and currently offers a wide range of Bachelor's, Master's, and doctoral programs in various fields of engineering, technology, and sciences. TU Delft is known for its cutting-edge research in areas such as aerospace engineering, robotics, water management, sustainable energy, and many others. It is a member of several prestigious international networks, including the EuroTech Universities Alliance and the League of European Research Universities (LERU) (TU Delft, 2023).
- **TU Dublin**. TU Dublin is also known as Technological University Dublin, is one of the larger universities in Ireland, formed by the amalgamation of three former institutes of technology: Dublin Institute of Technology, Institute of Technology Blanchardstown and Institute of Technology Tallaght. TU Dublin was established on January 1, 2019, and is Ireland's first technological university. TU Dublin offers a wide range of undergraduate and postgraduate courses across several faculties, including engineering and built environment, science, business, arts, humanities and social sciences, and health. TU Dublin is known for its focus on practical and industry-focused education, with strong links to local and international businesses and industries. The university has a reputation for innovation and entrepreneurship, and its research centres are at the forefront of scientific and technological developments in Ireland (TU Dublin, 2023).
- **University of Hildesheim**. The University of Hildesheim is a public research university located in the city of Hildesheim, Lower Saxony, Germany. The university was founded in 1978 and has since become known for its strong focus on the humanities, social sciences, and cultural studies. Hildesheim University offers a range of undergraduate and postgraduate programs across four faculties: Education and Social Sciences, Cultural Studies, Linguistics and Information Science, and Mathematics, Natural Sciences, Economics, and Computer Science. The university is well-regarded for its research, with a focus on interdisciplinary and intercultural studies. Hildesheim University is known for its supportive and inclusive community, with a focus on creating a welcoming and diverse environment for students and staff. The university has a strong international reputation and attracts students from around the world (University of Hildesheim, 2023).
- **University of Melbourne**. The University of Melbourne is a public research university located in Melbourne, Australia. It was founded in 1853 and is the second-oldest university in Australia. The university offers a wide range of undergraduate, graduate, and postgraduate degree programs in various fields of study, including arts, science, engineering, law, medicine, and business. It is known for its high-quality education, research excellence, and strong international reputation. The University of Melbourne is consistently ranked among the top universities in the world and is a member of the

prestigious Group of Eight, which is a coalition of leading Australian universities (University of Melbourne, 2023).

Appendix D. XAI developer and expert interview

As previously detailed in the methodology chapter, this section comprises six distinct elements that serve to maintain coherence in the findings and facilitate the possibility of replication, as required. The components are as follows:

1. Introduction and welcome
2. Background information
3. Overview of the framework
4. Expert feedback on the framework
5. Expert feedback on implementation and limitations
6. Conclusion and next steps

On the following page is the English template that was used for every expert interview. After that, that same template has been translated to Dutch, as a number of interviews were held in Dutch.

Dear [interviewee name],

As you may know, my name is David Lensen, and I am currently in the process of writing my final thesis for the master's program Complex Systems Engineering and Management at TU Delft. For this research, I intend to create a comprehensive framework for assessing the interpretability of XAI explanations, specifically for laypeople. I would like to start by thanking you for agreeing to help me with one of the key elements of this research, namely, expert validation of the first prototype of the framework. The goal of this interview is to gain a better understanding of the strengths and limitations of the framework and to identify potential areas for improvement. As an expert in the field of XAI, your insights and feedback are invaluable, and I am excited to hear your thoughts on the framework. Thank you for taking the time to join me today, and I look forward to a fruitful discussion.

Before we dive into the details of the XAI interpretability framework, I would like to take a moment to get to know you as an XAI developer better. Would you like to introduce yourself and briefly describe your experience with XAI development? This can include any relevant experience you have with developing XAI systems, working with machine learning algorithms, or any other related expertise. Knowing your experience and background will help me to understand your perspective on the framework and how it aligns with your own experiences.

- Room for expert input, potentially followed by follow-up questions and answers.

Thank you for that! Now that we have a better understanding of your background and experience, let's move on to discussing the XAI interpretability framework *(the framework prototype will be shown here)*. As I mentioned earlier, this framework is designed to provide a comprehensive and flexible approach to evaluating the interpretability of XAI systems for laypeople. The framework incorporates a range of factors that are important for XAI interpretability, such as clarity, transparency, relevance to user's goals, trustworthiness, and others. By using this approach, we aim to promote transparency, trust, and understanding in the development and application of XAI systems. Before we discuss your specific feedback on the framework, do you have any initial thoughts or questions about the framework's design or implementation that you'd like to share?

- Room for expert input, potentially followed by follow-up questions and answers.

Now that we have covered the framework's general overview, I would like to hear your thoughts and feedback on the details of the framework. Specifically, I am interested in understanding your opinions on the following:

1. Do you think that the factors included in the framework are relevant to evaluating the interpretability of XAI systems?
2. Are there any factors that you believe should be included in the framework that are currently missing?
3. Are there any factors that you believe should be excluded in the framework that are currently unjustifiably there?
4. Could you please rank the five most important factors for XAI interpretability, from most important to least important? (Useful for applying weights to the factors)
5. Are there any improvements that you would suggest to the framework to make it better in practice?

Please feel free to share any other thoughts or feedback that you may have as well.

- Room for expert input, potentially followed by follow-up questions and answers.

Thank you for sharing your feedback on the framework and potential areas for improvement. I'd now like to discuss the implementation and limitations of the framework. I have some more general questions regarding implementation of the framework:

6. In your experience, are there any challenges or limitations to implementing the framework in practice?
7. Additionally, are there any challenges that might arise when trying to get stakeholders to agree on the interpretability of an XAI system using the framework?
8. How might these challenges be overcome (practically)?
9. How relevant do you see framework to your current work in XAI and how useful do you estimate the framework to be when developing XAI for non-expert users?

- Room for expert input, potentially followed by follow-up questions and answers.

Your insights into the implementation and limitations of the framework will be very valuable in helping me understand how the framework can be best applied in practice. Thank you for your insights and feedback on the XAI interpretability framework. Your contributions are of great value as we continue to refine and improve the framework. In summary, we have discussed the strengths and potential limitations of the framework, explored potential areas for improvement, and identified some challenges that may arise when implementing the framework in practice. Based on your feedback, I will make some adjustments to the framework and incorporate your suggestions. I will also continue to test the framework and evaluate its effectiveness in different scenarios. Finally, I would like to thank you for your time and contributions to this interview. Your expertise and insights have been very valuable, and I appreciate your help in evaluating the framework.

Beste [naam van de geïnterviewde],

Zoals u inmiddels wellicht weet, mijn naam is David Lensen, en ik ben momenteel bezig met het schrijven van mijn master scriptie voor de opleiding Complex Systems Engineering and Management aan de TU Delft. Voor dit onderzoek ben ik begonnen met het maken van een framework voor het beoordelen van de interpretability van XAI-systemen, specifiek voor leken. Ik wil beginnen met u te bedanken voor de medewerking bij een van de belangrijkste elementen van dit onderzoek, namelijk expertvalidatie van het eerste prototype van het framework. Het doel van dit interview is om een beter begrip te krijgen van de sterke en zwakke punten van het framework en om zo eigenlijk verbeterpunten te identificeren. Als expert op het gebied van XAI zijn uw inzichten en feedback van onschatbare waarde, en ik kijk er naar uit om uw gedachten over het framework te horen. Bedankt dat u vandaag de tijd neemt om mij te ontmoeten, en ik kijk uit naar een interessante discussie.

Voordat we ingaan op de details van het XAI-interpretability framework, zou ik graag een moment willen nemen om u als XAI-expert beter te leren kennen. Wilt u zichzelf introduceren en kort uw ervaring met XAI-ontwikkeling beschrijven? Dit kan alle relevante ervaring omvatten die u hebt met het ontwikkelen van XAI-systemen, werken met machine learning algoritmen, of andere gerelateerde expertise. Kennis van uw ervaring en achtergrond zal mij helpen uw perspectief op het framework te begrijpen en hoe het overeenkomt met uw eigen ervaringen.

- Ruimte voor expertinput, eventueel gevolgd door vervolgvragen en -antwoorden

Bedankt daarvoor! Nu we een beter begrip hebben van uw achtergrond en ervaring kunnen we verder gaan met de bespreking van het XAI-interpretability framework. Zoals ik eerder heb vermeld, is dit framework ontworpen om een uitgebreide en flexibele aanpak te bieden voor het evalueren van de interpretability van XAI-systemen voor leken. Het framework omvat een reeks factoren die belangrijk is voor XAI-interpretability, zoals clarity, transparancy, relevance to users goals, trustworthiness en andere factoren. Door deze benadering te gebruiken, streven we naar transparantie, vertrouwen en begrip bij de ontwikkeling en toepassing van XAI-systemen. Voordat we uw specifieke feedback over het framework bespreken, zijn er misschien initiële gedachten of vragen over het ontwerp of de implementatie van het framework die u wilt delen?

- Ruimte voor expertinput, eventueel gevolgd door vervolgvragen en -antwoorden

Nu we de algemene inhoud van het framework hebben behandeld, wil ik graag uw gedachten en feedback horen over de details van het framework. Specifiek ben ik geïnteresseerd in het begrijpen van uw mening over het volgende:

1. Denkt u dat de factoren die in het framework zijn opgenomen relevant zijn voor het evalueren van de interpretability van XAI-systemen?
2. Zijn er factoren die u denkt dat moeten worden opgenomen in het framework die momenteel ontbreken?
3. Zijn er factoren die u denkt dat moeten worden uitgesloten van het framework die momenteel wel aanwezig zijn?
4. Kunt u alstublieft de vijf belangrijkste factoren voor XAI-interpretability rangschikken?
5. Zijn er verbeteringen die u zou voorstellen om het framework beter in de praktijk te maken?

Heeft u verder nog feedback over de inhoud van het framework zelf?

- Ruimte voor expertinput, eventueel gevolgd door vervolgvragen en -antwoorden

Bedankt voor het delen van uw feedback over het framework en de potentiële verbeteringsgebieden. Ik zou nu graag de implementatie en beperkingen van het framework

willen bespreken. Ik heb nog enkele algemene vragen over de implementatie van het framework:

6. Uitgaande van uw ervaring, zijn er uitdagingen of beperkingen bij het gebruiken van dit framework in de praktijk?
7. Misschien iets specifieker: zijn er uitdagingen die zich kunnen voordoen wanneer geprobeerd wordt stakeholders het eens te laten worden over de interpretability van een XAI-systeem met behulp van het framework?
8. Hoe kunnen deze uitdagingen worden overwonnen (praktisch)?
9. Hoe relevant ziet u het framework voor uw huidige werk in XAI en hoe nuttig schat u het framework in bij het ontwikkelen van XAI voor niet-experts?

- Ruimte voor expertinput, eventueel gevolgd door vervolgvragen en -antwoorden

Uw inzichten in de implementatie en beperkingen van het framework zullen zeer waardevol zijn om mij te helpen begrijpen hoe het framework het beste kan worden toegepast in de praktijk. Bedankt voor uw inzichten en feedback over het XAI-interpretability framework. Uw bijdragen zijn van grote waarde terwijl we doorgaan met het verfijnen en verbeteren van het framework. Samenvattend hebben we de sterke punten en potentiële beperkingen van het framework besproken, potentiële verbeteringsgebieden verkend en enkele uitdagingen geïdentificeerd die kunnen ontstaan bij de implementatie van het framework in de praktijk. Op basis van uw feedback zal ik aanpassingen maken aan het framework en uw suggesties opnemen. Ik zal ook doorgaan met het testen van het framework en het evalueren van de effectiviteit ervan in verschillende scenario's. Tot slot wil ik u bedanken voor uw tijd en bijdragen aan dit interview. Uw expertise en inzichten zijn zeer waardevol en ik waardeer uw hulp bij het evalueren van het framework.

Appendix E. Interview summaries

In appendix D, all interviews are summarized and presented in the same order as in Table 4 in chapter 7.

Interview 1. XAI – Human-AI interaction – academia & industry expert

This expert has worked on XAI for the past 7 years and wrote a dissertation on XAI. As a researcher time is divided between fundamental research on XAI and its applications in society in a broad sense. Both looking at the technical side and the human centred side. Often approached as a design process by starting at looking at which explanation needs to be given, consequently thinking about how to generate such an explanation.

First notion is the difference between metrics and constructs. During the interview I talked about metrics when I meant the factors. However, that overcomplicates things, since not all factors can be measured (therefore, they are not metrics). A better notion is constructs. After which the term framework is also discussed. Isn't 'theoretical model' more suited for this deliverable?

The primary concern at first is: how to ensure that this framework is fully comprehensive regarding XAI interpretability for laypeople. The feeling occurred that this can be discussed in a lot of different ways. The validity of the factors was questioned in a sense. After discussing where the factors originated, from multiple literature reviews, therefore giving insights into the process, that concern was managed, however not vanished.

Looking into the details of the framework: there were no immediate red flags or other factors that needed immediate attention, except for one major factor, which is context. A lot of factors (such as relevance) are very context dependent. The advice here is to not include all context variables in the framework itself, but when discussing the factors, include context. For example, for each factor discuss how it could be influenced by context.

Furthermore, RRT as an example is a rather clear metric, whilst transparency is quite a vague construct. Should they be categorized? Rephrased? The framework would benefit from being on the same level of abstraction. This could provide more structure.

Lastly, model fidelity is included, but explanation fidelity is not. Whilst this seems to be an even more important factor considering the scope of this research. This expert has encountered that in some cases, the contents of the explanation are not even that important for the layperson. The fact that there is AN explanation accompanying the decision, is one of the major trust-enhancing factors available (disregarding the contents of the explanation). Especially laypeople are sensitive to that.

It's difficult to pick the five most important factors, since that is highly context dependent. For a patient, it is rather different compared to an insurance agency. Another example, right now we're more looking into decision support systems, whilst autonomous or robotics systems require a different set of factors to be considered most important. They are all very important, but it differs per context.

Practically operationalizing explainability in society: companies for example do not know how to handle explainability in general. They do continuously keep hearing: do something with explainability. But 'what' and 'how' remain unanswered. A lot of data scientists bluntly import a python XAI package and run it and consider the job done. But this is not done! Thankfully, the group that looks back and thinks: 'did I give the right explanation?' is growing. Therefore, this framework could be a real first step into properly explaining. It might not be concrete enough (you will need metrics and so on), but it is a real first step. This will ensure for example that

XAI developers no longer present SHAP explanations to laypeople. In that case, giving no explanation would probably be interpreted in a better way. This might also be a pro for putting more thought and effort into the key principles.

Practical improvements: this framework presents a good medium for researchers to get discussion going, but for developers it should be a bit more concrete. But is this necessarily a step that should already be addressed in this research?

Another 'improvement' – stakeholder wise – is to clearly define laypeople. Short note: a layperson is someone that is not an expert on the domain in which the XAI operates. Stakeholder challenges are very context specific. Include stakeholder engagement into design process could be a possible solution. Basically, the standard stakeholder mitigation tasks.

The opinion is that both for beginning XAI researchers/developers as well as people that have been working on it for years, it's highly important that all factors mentioned in the framework are not solely being considered in the back of their heads. But that they are concretely discussed, using this framework. The entire research field is often quite lax regarding the entire process and the details of XAI interpretability. Whilst in practice, it is much more complex than often regarded. Almost no-one really takes that into account the way it actually should. This framework (considering the necessary adaptations) is therefore very relevant and useful for all XAI researchers and developers.

Interview 2. XAI – Human-AI interaction – academia & industry expert

The interviewee is an expert in human-AI interaction, currently concentrating on explainable AI and responsible AI.

In the initial discussion, the expert agrees that individual factors align well with typical literature. However, they challenge the claim that the framework is specifically for laypeople.

Creating a prioritized list can be complex, they note, given that many factors overlap and vary in scope. They identify the top four elements as comprehensibility, transparency, relevance, and model fidelity. Highlighting the importance of these factors' prioritization, they also argue that their significance should be validated not by experts, but by laypeople through use-case-scenarios. This approach would both validate and rank the factors according to importance. Furthermore, this expert acknowledges that context heavily influences the prioritization, though the focus on laypeople may streamline the context considerably.

The expert proposes that merging overlapping factors could increase the framework's efficiency.

On the subject of operationalization, they suggest that attaching specific criteria to each factor would substantially improve the framework's implementability. Prioritization is again emphasized, not only for its efficiency but also for quicker stakeholder consensus. They also emphasize the need to address potential trade-offs within the framework's fundamental principles.

Clarifying factor definitions is essential for stakeholder consensus, according to the expert.

In conclusion, they stress that the framework serves a crucial role in providing direction for XAI developers, although it's just a starting point. Evaluating each criterion is an essential next phase.

Interview 3. XAI – Human-AI interaction – academical expert

The individual factors of the framework make a lot of sense for this expert. One factor that could be added is: fit-for-purpose. Although it might be a combination of some of the factors that are already there. However, it is highly interesting to think about. They were doing experiments that showed some explanations do indeed make sure that the participants understand the model, however, the explanation does not improve/change their decision-making process. So, in that experiment, the explanation was fit-for-purpose if the purpose was to understand the model better, however it was not fit for the purpose of better decision making. That specific niche factor might not really fit under relevance, since relevance captures a different thing. Perhaps this has some overlap with relevance/actionability. There are no real factors that should be excluded. One thing that may need some more research is the using probabilities factor. This factor is based on the claim that probabilities may be too confusing. However, that may be too strong of a claim. The claim should most likely be that causal information is more convincing than probabilistic information. So, does using probabilities negatively influence interpretability? This expert is sceptical. It might not be a definite negative relationship. Thirdly, number of causes is relevant. However: shouldn't that be captured into complexity? That might also benefit the level of abstraction. Some of the more recent work regarding complexity (number of causes) states that the more complex the event actually is, the more complex an explanation people will accept. There is a sweet spot that influences trust from a person. If you are not 'complex' enough in your explanation, they will not buy it. If you present too complex of an explanation considering the to be explained event, people will also not buy it. Fourth, abnormality influences the desire for understanding rather than goals/needs. Therefore, relevance might not be positively influenced by abnormality. Look into this a bit more.

Lastly, trustworthiness should perhaps focus more on correctness: high model fidelity actually, than people will adopt the explanation and the decision of the XAI. Furthermore, this one could very well be split up into trustworthiness (of the explanation) and trust (from the user, perception (of trustworthiness) of the people).

The five most important factors according to this expert are: trustworthiness, relevance, generalizability, intuitive understanding and comprehensiveness.

Improvements to the framework: context. The context of making decisions is not yet a factor. It should be. For example: how high are the stakes regarding the decision made by the XAI. Another contextual factor is for example the workload of the person (if they have to make a lot of decisions, that might affect how the explanations are being interpreted). One last factor that was discussed is the actual nature of the user of the XAI: some people have natural low trust levels, others very high. Some people are sceptical towards computers, others are not. The task of the user considering the XAI is another context factor. Context could therefore even be another inward factor in the framework. Therefore, context presents a real challenge to implementing the framework in practice.

When there are different stakeholders in a decision-making process, there are always challenges. Expectations and demands differ. However, that's okay. You can go for the average and just say well what we're trying to do is not satisfy everybody, but just the largest number of stakeholders possible. Basically, the standard stakeholder related problems.

This expert generally thinks that the framework makes a lot of sense. The relevance of this framework is also discussed. First of all, for example for new students or postdocs that have limited prior knowledge regarding XAI. This could really benefit them and their work. It does a pretty good job at capturing the complexity of XAI interpretability on a high level. Both from a measurement and a design perspective. The framework even got the expert thinking about some of the conclusions that can be drawn based on the framework. The expert states that he can surely see it being used in their cases. He's sure that a lot of factors are discussed at some

point in time when working on XAI, however, they are not yet presented clearly in such a framework. Right now, there is very limited research on this topic, and we cannot yet compare different studies due to the lack of a structured way. Most people design an XAI in the way that they think is best.

Interview 4. XAI – Human-computer interaction – academical expert

The interview starts off by explaining to perhaps look into the linguistic perspective of XAI explanations. The communications settings basically. There is a specific linguistic theory which the expert usually refers to and that is the systemic function theory. Systemic functional linguistics (SFL) is a linguistic theory that describes language as a social semiotic system. Developed by Michael Halliday in the 1960s and 1970s, SFL is a functional approach to language that emphasizes the social context in which language is used and how it is used to convey meaning. According to SFL, language is a system of resources that speakers use to make meaning in social situations. These resources include grammar, vocabulary, and discourse structures, as well as knowledge of the social and cultural context in which language is used. SFL posits that language is always used to achieve a particular communicative goal, such as giving information, persuading, or expressing feelings. SFL analyses language at three different levels: the textual level, the interpersonal level, and the ideational level. The textual level refers to the grammar and discourse structures that speakers use to convey meaning. The interpersonal level refers to how speakers use language to interact with each other, including how they establish and maintain social relationships. The ideational level refers to how speakers use language to convey their understanding of the world and their experiences. SFL has been applied to a wide range of contexts, including education, media studies, discourse analysis, and language teaching. It is also used in computational linguistics and natural language processing. SFL is known for its ability to provide a comprehensive framework for analysing language use in social contexts, and for its emphasis on the role of language in social interaction and meaning making.

The difference between explanations for experts and novices (laypeople) is rather interesting and this expert advised me to look at research surrounding the Mycin project. It could be interesting to see how that fits into this more modern research (Mao).

The main problem in this field is that there is surprisingly little about how explanations work. How explanatory processes work. So, when you talk to philosophers, linguists, psychologists, and computer scientists, there are still lots of unknowns. It's still an evolving field. That is something that computer scientists do not like. They like implementation ready things. However, this requires close collaboration between all parties. There is often a very big gap:

Background is in AI and Human Computer Interaction (HCI). One of the things you learn early on in HCI is that every system has to be contextualized. You have different notions of factors depending on the context. In a high-risk environment, you will have different users, and different notions of factors, therefore. For example, take a look at the measure of trustworthiness. In a high-risk environment, you might not even need trust. To gain trust, you perhaps do not need to open every black box. Every time you get on a plane, you will feel confident, because the last time you did everything went fine. You do not need to know why it went fine (Wolter Pieters worked on that aspect). So, the first concern is that you cannot have a general model, look at an explanation and say I can assign these values to the factors without looking at the users and the overall context.

Measuring the variables is difficult. Also, very context dependent, but even then, it remains difficult.

Intuitive understandability is difficult. This actually means: how good is the explanation in relation to existing mental models. For example, if you are using smartphones for a while, you will automatically enlarge photos by pinching your fingers. You cannot enlarge a physical piece

of paper like that. You have to learn that it is possible to do so with digital images on smartphones. Now, people call that 'intuitive use', however: there is no intuition/naturalness in that. It is learned. Therefore, the concept of intuitive understandability might work out better when changing it to *affordance*. In HCI, there is a book by Don Norman: the design of everyday things. The idea is that there is something that makes us think: a chair, I can sit down. It has the affordance of being a seat for people. This being a seat Is signified by the shape and colour for example. It's much more in that direction. You have to take the user into account to determine intuitively.

Coherence with prior beliefs is very interesting (cognitive biases). That makes a lot of sense to have in the framework. Trustworthiness in itself is a difficult concept. More information can actually decrease trust (the more you know about planes, the less comfortable you will probably be when you are in one 30.000 feet in the air). It's a complex relationship. Opening up the black box can make people feel insecure. Abnormality is very interesting. Usually, explanations are needed when something is surprising. So, including abnormal cases is very important. The element of surprise is mentioned in a book by David Leake very well. Roger Schank is a cognitive scientist (Explanation patterns – understanding mechanically and creatively). They have the same school of thought regarding that. Next, transparency is opening the black box. That must therefore be balanced. Users are more looking for justification as opposed to explanations. If you are told to eat more fish because it's good for you, that's not an explanation, but it might be enough of a justification without going into the details how omega 3 is beneficial for your cognitive memory.

A factor that could perhaps be added is 'actionability': would the user act based on the explanation?

From design perspective it might be too complex of a diagram. For every factor it needs to be justified why it's there, based on literature. However, for analysis purposes, it's very good to have as many factors included as possible. But, when for example actually designing an XAI explanation, it may be a bit overwhelming. So, it might be better to dam it down if that's the purpose. This may be why it's good to focus more on key principles.

Do not leave any factor out just yet. You need to be able to measure different things. RRT is easier to measure than trustworthiness.

Experiments need to be done with human users (as laypeople). This is an indication that the next evaluation round using a survey is a good idea. However, first of all the issues with the framework need to be addressed, before sending out the survey. After that, larger user studies are necessary.

The moment you involve stakeholders you will have to map out the different roles and find common ground. People will not agree on everything. The field of HCI has a lot of experience with bringing stakeholders together. How can we come up with different definitions.

Interview 5. XAI – Human-computer interaction – academical expert

All factors appear to be very relevant. There is nothing that is not supposed to be there. What is advised is to split all characteristics into what can be measured and other factors that cannot be measured. Number of causes can be very easily measured by counting them. But transparency or clarity might be difficult to measure. There are metrics, but they are less intuitive. Simplicity might seem tricky at first, but this could indirectly be measured by looking at the number of words used in the explanation. Therefore, splitting the factors by measured interpretability and perceived (or human centred?) interpretability. The expert believes that it is almost necessary for a framework to have a method to evaluate itself. Therefore, measuring factors is important.

A missing factor might be 'robustness'. Basically, meaning that if you slightly change the inputs, check the output of the model, and what does the explanation do. Take a good look into that.

The five most important factors (in order) are: simplicity, comprehensiveness, robustness, trustworthiness, and relevance. The two most important factors of course need to be balanced: an explanation needs to be as simple as possible, whilst being comprehensive. Simplicity does come first, because that always is the goal. But whilst staying simple at first, be as comprehensive as is allowed. The other factors are all important, but these five factors can really be worked on by XAI developers. The other factors are also much more dependent on the person and the context.

Understandability and interpretability are by some people interpreted as slightly different compared to explainability. However, this expert thinks it's possible to argue that they are all about the same.

The overall structure of the framework is very nice. The most important and difficult part of implementation is ability to measure the factors.

To be able to talk about challenges when trying to get stakeholders to agree on XAI interpretability, more context is needed: who are the stakeholders? Who is the target audience? What information do they want to see? Therefore, it is very context dependent. In a real-life problem, you have to understand the requirements of every stakeholder. Having evaluation metrics will be advantageous for that as well. The right type of explanation will change accordingly.

The frameworks relevance is also discussed. The expert thinks it's very useful and relevant. There is very little research on the relationship between the different factors used in the framework in combination with XAI interpretability. Therefore, it's very relevant.

Interview 6. XAI – Human-computer interaction – academical expert

Quite a wide range of factors is presented. It is hard to distinguish one factor from another, however, re-reading the definitions would help. It is believed that measurability is a big aspect of this framework. Factors are only relevant if they can be measured in experimental setting. Measuring factors is also necessary to make the framework operationalizable. One example per factor would be rather helpful, since some of the factors can be perceived as quite similar. Affordance is for example highly related to for example complexity and simplicity. How do you distinguish which factor we will look at now.

The top factors as presented by this interviewee are: comprehensiveness, clarity, simplicity, transparency, complexity and trustworthiness.

What is also important, are the factors linear or not? If you enhance multiple factors, do the results increase one another, or continue linearly? Perhaps linear is not the right word, but monotonous is better.

There are a lot of factors included in the diagram. It may be a good idea to go a different level of abstraction and see if you can get fewer factors, that can all be measured. Try to have as little factors as possible in there, which have as little overlap as possible.

Perhaps remove intentionality and add causality. This would also incorporate using probabilities. Considering abnormality: abnormal in comparison to what? Abnormal in comparison to all other explanations ever given? Should the XAI mention that? What is abnormal and what is not, is also related to prior beliefs of the user.

Interview 7. XAI – Interactive intelligence – academical expert

The interviewee starts off by explaining that choosing laypeople as target audience for the framework overcomplicates the situation a lot. Due to the fact that laypeople can simply not be considered a homogeneous group. Each member of that group acts and behaves differently, and therefore anticipates information in a different way.

After having taken that into account, the second point of interest according to this expert was operationalization. As of now, the framework is an abstract conceptual idea. This is much less helpful in practice as opposed to measurable tools. Right now, the contents of the framework are rather generic. We'd have to think about what we want to add to the scientific community. And how to test that. It is believed that the framework is fine, but the potential user of the framework (an XAI developer) will need more help in order for it to actually be helpful.

Number of causes is for example an interesting one. It is believed that there is an optimal balance for the number of causes necessary to implement in the XAI explanation. However, where is that optimal point? Is it with 3 causes for example?

Interview 8. Organizational XAI – academical expert

Organizational and management side of XAI is main research concern of this expert. This is interesting because there are a lot of different types of people that deal with the (X)AI. What are the roles of different people concerning the XAI. Therefore, constant adjustments are necessary. Concerning different types of people: it is very important to clearly define/consider the characteristics expert and laypeople in a given situation. In some cases, it can be very clear. However, there are undoubtedly cases where it can be rather vague. Where will people be placed that have a lot of domain knowledge, however little AI knowledge. So bottom line: who classifies as a layperson and who as an expert? Very context dependent. There are certain matters in our society where everyone knows something from. For example, mowing the lawn. Most people mow their own lawn. A gardener can be considered an expert, but is someone that has been mowing their own lawn for several decades still a layperson? Most likely not. There is a large grey area, therefore.

Context determines how interpretable something is, and how interpretable it needs to be. There are a lot of factors in the framework, but they are most likely to have a different heft to them, depending on the context. There are situations where clarity is very important, but abnormality is not even necessary to be considered. When taking into account the context of an explanation, it might suddenly become necessary to factor in whether or not there is a positive or a negative outcome. A negative outcome may even be less interpretable. But the relative heft of the individual factors changes to say the least.

The entire framework concerns XAI interpretability. However, do we not mean the interpretability of the explanation generated by XAI? Very good to clearly mention that. Furthermore, some of the factors included do not have the same level of abstraction. Some can be devoted to the XAI explanation, others actually to the XAI model, others to the user. Right now, they are all mixed up. Perhaps visualize this in terms of process? The natural order would be model first, XAI explanation second, user third.

Another similar issue is perception versus behaviour. For example, you can see the recall response time as a behaviour. However, trustworthiness will remain a perception of that user. However, you could also try to learn something about trustworthiness from the behaviour of the layperson. You can almost choose what to do for every factor. This also has a major influence on how to view the framework.
Another issue is the abstraction level of the framework. For example, including probabilities in explanations is very much influential. However, on that same level there are more (similar) properties of explanations: text vs video, including graphs, including importance weights, including mathematical functions. These are real contents of the explanations and not how to

evaluate them. Number of causes is a similar factor. There may be a difference in the XAI itself and the evaluation of the XAI. A user might not look at the explanation and think: a probability is used, the model has low fidelity, he experiences that it is incoherent with prior beliefs, and he doesn't trust the explanation.

The expert often encounters that a lot of factors have overlap. Or maybe not even similar, almost opposites of each other. For example: incoherence with prior beliefs is rather similar to abnormality. Comprehensiveness and simplicity also have quite a lot of overlap. A very comprehensive explanation is not simple, and a very simple explanation can almost never be very comprehensive. It could be useful to find measurements/examples for each factor (in the form of a survey for example). That way you notice that the examples and how to measure them are perceived as the same by participants. The expert advises to provide example explanations for each factor and for combinations of factors. When is something generalizable and when is it not? To make it a little more tangible.

What is the difference between direct and indirect factors. The indirect factors may have much more to do with the decision-making context and not necessarily the explanations itself. However, model fidelity concerns the actual XAI model. Again, it could help to differentiate between categories of factors.

The expert doesn't think that it is possible to generally rank the factors from most influential to least influential. Which combination of factors is important in which context. When we're thinking healthcare: transparency and relevance are very important for example. A factor that may be influential in healthcare is again actionability. Does the layperson change his/her behaviour based on the explanation. What does the user do with the explanation. Counterfactual explanations are perceived as very actionable. Actionability is therefore very context dependent.

The first key principle of the framework notes that transparency and clarity are crucial. However, when explaining why they are, understanding, confusion, clear and transparent are mentioned, therefore other factors are included in the explanation and links between factors are explained. This is more a description of the framework and not conclusions drawn from the framework.

Relevance and alignment with user's goals might also have a lot of overlap.

The next plan of the survey to evaluate the framework with laypeople was also discussed briefly. Think carefully about what makes one explanation clear (for example) and the other not. That is what we want to know. A 'nasal issue' might be very much clearer for other people compared to a 'sinus infection'. Categorizing factors may help. Abnormality, intentionality, coherence with prior beliefs and relevance all have to do with what can the subject do with it. Transparency simplicity comprehensiveness clarity all concern how the subject perceives the explanation. RRT and ITR concern measurable factors from the user. This is all very philosophical. If an XAI developer receives the framework, he/she can understand it, but there are not yet clear guidelines or actions that can be undertaken. It is very difficult to give developers handles/hold that allows them to figure out how to program an explanation using concepts like this. It would be very nice if you give the framework to someone and say: start here, follow the guidelines and proceed clockwise. Shape it more in the form of an action plan perhaps.

Something else concerning context, what is the goal of giving an explanation? The goal is not necessarily to be interpretable, but to enhance trust, be transparent, enhance acceptance, or spark actionability, just to give a few examples. The context of a healthcare self-diagnosis app could be to spark actionability: helping someone on their way to take the right action.

The framework is very complete, but maybe too complete. The key principles might make more sense and become more interesting when exploring the relationships between the individual factors a bit more. Perhaps it would be an idea to really focus on the key principles and present that as the key deliverable, derived from the framework. That way, an XAI developer or researcher could look at those principles and based on the principles enhance the explanation of their XAI. It would be very nice to build the framework in a way that you can 'intuitively' spot the key principles. Perhaps work with colours and map them.

Interview 9. XAI development – academical expert

The first thing that is noted is that the XAI developer could also be a layperson in the domain of the XAI. For example, the developer of the SymptomSolver XAI application is not a medical expert. The entire system should therefore be thoroughly validated by medical experts.

This expert is interested in the recall response time. This could be changed or diverted to the time taken by the subject to interpret the explanation, as opposed to amount of time before recall.

All factors included in the framework are properties of different things. For example, RRT is a property belonging to a person. But then, using probabilities belongs to the explanation. And lastly, model fidelity belongs to the XAI model. Those could be the three categories to clearly categorize the factors: XAI model, explanation, and people. This might also help resolve challenges amongst stakeholders, when you are struggling with a factor and you know it's regarding the user, all stakeholders are aware that the factors are differently perceived by different users. *From another interview: you might struggle to find a category for each factor. For example, coherence with prior beliefs. That could fit in each of the three categories.*

Concerning specific factors: relevance and intuitive understandability might be too close. Could they be merged? However, if they can be defined clearly separated from one another: keep both. Also, a pro to change intuitive understandability to affordance. The using probabilities factor: it's important to note that this is most likely something that makes an explanation to an expert on the other hand more interpretable, whilst the explanation is less interpretable to a layperson. Next, the relationship between model fidelity and trustworthiness is believed to be not straight forward. It is believed that something is missing: truthfulness could be added between the two factors.
The most important factors to this expert include: RRT, ITR, intuitive understandability, relevance, model fidelity.

Improvements that can be made to the framework are the three categories mentioned before. Other improvements that can be made include clearly defining more trade-offs in the key principles. That make it interesting. Another trade-off is between simplicity and transparency. Also, the third key principle is not yet fully clear. In generally for the key principles, try not to simply explain the relationships in the framework, try to really draw conclusions from them.

The expert agrees that both the framework and the key principles are important. Without the key principles, it's difficult to interpret the framework. Whilst without the framework, the key principles appear to be coming out of nowhere.

Limitations to the framework right now is that a lot of things in the framework are not measurable. They are all very difficult to measure. A lot of factors are dependent on the person that received the explanation. It is all about perceived interpretability.

The framework was a very interesting and relevant read for this expert. There are currently not a lot of researchers looking into this topic at all and combining factors in a framework is especially interesting. The relevance for XAI developers is stressed, since it provides handles for them, for example after having finished the initial design to go over it once more and think:

okay, what can I do to make the explanation even more interpretable to laypeople. And it should be noted that everyone working in the field of XAI is struggling with measurability. It is definitely something to take into account, however, it is not the end of the world if measurability remains a problem for this framework.

Interview 10. XAI development – industry expert

During the interview with this XAI developer, we discussed the first prototype of the XAI interpretability framework. The developer expressed that the framework was very comprehensive, covering a lot of factors. However, this also made it difficult to use in practice as there were a lot of factors to consider. The developer suggested categorizing the factors or focusing on the framework's key takeaways to make it more practical. A practical alteration was given: to combine ‘coherence with prior beliefs’ and ‘overlap with human understanding’.

The developer also pointed out that the framework is most likely to be rather use-case-dependent.

Considering the stakeholder related question: “are there any challenges that might arise when trying to get stakeholders to agree on the interpretability of an XAI system using the framework?” It was concluded that different stakeholders may have different goals, as in the case of the self-diagnosis app. For example, a health insurer may be happy if people do not utilize ordinary care because of the app, while patients are happy if they are diagnosed correctly. Therefore, all goals need to be taken into consideration when using the framework.

The developer once more emphasized the need to make the framework more practical and usable. Categorizing the factors or creating a checklist based on the framework's principles could help with this. The most important thing at this stage is to focus on making the framework more practical to use.

In conclusion, the XAI developer's feedback highlighted the strengths and weaknesses of the first framework prototype. While the framework was comprehensive, it was also overwhelming and may be difficult to use in practice. Therefore, the developer suggested ways to make the framework more practical and usable, such as categorizing the factors or creating a checklist based on the framework's principles. Additionally, the developer emphasized the importance of considering all stakeholders' goals when using the framework, as they may differ depending on the use case.

Interview 11. XAI development – industry expert

This developer investigates various available XAI methods and trying to develop machine learning models and applications accordingly on several ongoing projects.

Confusion existed regarding the factor abnormality. Most likely, balance needs to be found there as well, since abnormal explanations might decrease interpretability. Depending on application of course. If concrete examples are provided it might provide more clarity. Another thing that should be mentioned according to this expert is that transparency may be very important to experts, however, is transparency actually relevant for laypeople? Making the explanation more transparent, could even overcomplicate things for laypeople.

The five most important factors are trustworthiness, model fidelity, intuitive understandability, clarity and comprehensiveness.

Implementation in practice would probably need one more step, to make the framework better in practice. For example, create XAI explanation design guidelines or create a questionnaire or form to fill. This could allow for higher adoption rates. Implementing it purely as a conceptual framework could be challenging. Also, it’s very difficult to measure the factors and to ‘crunch hard numbers’ according to the framework. However, for system auditing for example, the

framework by itself is already very useful. Perhaps include something to evaluate whether or not you have fulfilled the key principles. This could also be for future work. In conclusion: try to go from philosophical to practical.

Different stakeholders will definitely have different requirements on what they want to have explained by the framework. You can use extended forms for different stakeholders. Based on what you fill into the form for example, you get steered into a different direction.

Relevance is very much focused on being aware of the key principles before designing XAI system. Or actually at the start of the design cycle, so when drafting the requirements. After implementing the first version and being in the process of iteratively altering the XAI, it might be nice to receive feedback based on the framework and implement the necessary changes into the system. In user interface development you have A/B tests and maybe this could be a way to A/B test XAI models.

Interview 12. XAI development – industry expert

This developer investigates various available XAI methods and trying to develop machine learning models and applications accordingly on several ongoing projects.

The expert first of all noted that some factors within the framework do indeed require explanations themselves to be fully understood. This highlights the importance of ensuring that each factor is clear and accessible to users.

The expert raised a question regarding whether the framework is focused on the XAI model or the XAI explanation. This distinction should be clarified to better understand the framework's purpose and application.

The expert suggested considering scoping the framework specifically to textual explanations. This could help streamline the framework by focusing on a specific type of explanation, making it more applicable and practical for users.

According to the expert, the top five factors for assessing XAI interpretability are (in descending order): intuitive understandability, relevance, comprehensiveness, generalizability, and clarity. These factors should be emphasized when refining the framework.

The expert recommended including metrics for evaluation purposes, such as scores for each factor. This would allow users to better understand areas that require improvement and facilitate the development of more interpretable XAI explanations. This enhances operationalization.

The expert acknowledged that the current framework serves as valuable guidance in XAI development. This feedback suggests that the framework is on the right track and can be improved further by incorporating the expert's suggestions.

All interviews summarised

	Individual factors	Improvements overall	Stakeholder related
Interview 1	Model fidelity is included, but explanation fidelity is not included. Considering the scope of the research, explanation fidelity may be even more important.	Since not all factors can be measured, the term 'metrics' should be avoided. Perhaps use the term 'constructs'. Also reconsider the term framework. Stress that it's a theoretical framework/model. There is no context mentioned anywhere! A lot of factors (such as relevance) are very context dependent. The advice here is to not include all context variables in the framework itself, but when discussing the factors, include context. For example, for each factor discuss how it could be influenced by context. RRT as an example is a rather clear metric, whilst transparency is quite a vague construct. Should they be categorized? Rephrased? The framework would benefit from being on the same level of abstraction. This could provide more structure. Try to improve the concreteness of the framework. Make it more practical to use in a sense.	Clearly define laypeople. Stakeholder challenges are very context specific. Include stakeholder engagement into design process could be a possible solution. Basically, the standard stakeholder mitigation tasks.
Interview 2	The individual factors look quite consistent with what is usually presented in literature. However, the question is raised regarding what makes this framework solely for laypeople? The five (actually four in this case) most important factors are presented as: comprehensiveness, transparency, relevance, and model fidelity. Prioritization of these factors is very important. Try to combine overlapping factors. This would enhance efficiency.	Prioritization of factors is context dependent. Operationalization is also discussed. Adding criteria to the individual factors would greatly impact the operationalizability of the framework. Also, prioritization is important for that. Lastly, address trade-offs in the key principles of the framework.	Prioritization is also important to get stakeholders to agree quicker. Also, definitions of factors must be as clear as possible.
Interview 3	Perhaps add fit-for-purpose or include it in relevance. One thing that may need some more research is the using probabilities factor. This factor is based on the claim that probabilities may be too confusing. However, that may be too strong of a claim. Actually, the claim should most likely be that	The context of making decisions is not yet a factor. It should be. For example, height of stakes, workload of the person, nature of the user (trust levels, scepticism levels towards computers e.g.), task of user in relation to XAI.	When there are different stakeholders in a decision-making process, there are always challenges. Expectations and demands differ. However, that's okay. You can go for the average and just say well what we're trying to do is not satisfy

	causal information is more convincing than probabilistic information. Number of causes is relevant. However: shouldn't that be captured into complexity? That might also benefit the level of abstraction. The more complex the event actually is, the more complex an explanation people will accept. There is a sweet spot that influences trust from a person. If you are not 'complex' enough in your explanation, they will not buy it. If you present too complex of an explanation considering the to be explained event, people will also not buy it. Abnormality influences the desire for understanding rather than goals/needs. Therefore, relevance might not be positively influenced by abnormality. Look into this a bit more. Trustworthiness should perhaps focus more on correctness: high model fidelity actually, than people will adopt the explanation and the decision of the XAI. Furthermore, this one could very well be split up into trustworthiness (of the explanation) and trust (from the user, perception (of trustworthiness) of the people). The five most important factors according to this expert are: trustworthiness, relevance, generalizability, intuitive understanding and comprehensiveness.		everybody, but just the largest number of stakeholders possible.
Interview 4	Perhaps look into changing intuitive understandability to affordance.	Look into the linguistic perspective of XAI explanations (SFL). Every system has to be contextualized. There are different notions for each of the factors depending on the context.	The moment you involve stakeholders you will have to map out the different roles and find common

	More information may also decrease trustworthiness. Therefore, it needs to be balanced with transparency. Important factors are the previous two and coherence with prior beliefs, abnormality. A factor that could be added is actionability.	Measuring the variables would be good, but that is very difficult. There are a lot of factors, which is very good from a theoretical perspective. Perhaps dam down by focusing on the key principles.	ground. People will not agree on everything. The field of HCI has a lot of experience with bringing stakeholders together.
Interview 5	A missing factor might be 'robustness'. Basically, meaning that if you slightly change the inputs, check the output of the model, and what does the explanation do. The five most important factors (in order) are: simplicity, comprehensiveness, robustness, trustworthiness, and relevance. These are chosen because they can really be worked on by XAI developers.	Split all characteristics into what can be measured and other factors that cannot be measured. It is necessary for a framework to have a method to evaluate itself. A lot of factors are very context dependent: include context in the framework!	To be able to talk about challenges when trying to get stakeholders to agree on XAI interpretability, more context is needed: who are the stakeholders? Who is the target audience? What information do they want to see? Therefore, it is very context dependent.
Interview 6	Perhaps remove intentionality and add causality. This would also incorporate using probabilities. Considering abnormality: abnormal in comparison to what? Abnormal in comparison to all other explanations ever given? Should the XAI mention that? What is abnormal and what is not, is also related to prior beliefs of the user. The top factors as presented by this interviewee are comprehensiveness, clarity, simplicity, transparency, complexity and trustworthiness.	It is believed that measurability is a big aspect of this framework. Factors are only relevant if they can be measured in experimental setting. Measuring factors is also necessary to make the framework operationalizable. One example per factor would be rather helpful, since some of the factors can be perceived as quite similar. There are a lot of factors included in the diagram. It may be a good idea to go a different level of abstraction and see if you can get fewer factors, that can all be measured. Try to have as little factors as possible in there, which have as little overlap as possible.	
Interview 7	Number of causes is for example an interesting one. It is believed that there is an optimal balance for the number of causes necessary to implement in the XAI explanation. However, where is that optimal point?	As of now, the framework is an abstract conceptual idea. This is much less helpful in practice as opposed to measurable tools. Right now, the contents of the framework are rather generic.	
Interview 8	Incoherence with prior beliefs is rather similar to	Very important to draw a clear line between experts and laypeople.	What are the roles of different people

	abnormality. Comprehensiveness and simplicity also have quite a lot of overlap. Suggests adding actionability factor.	Context determines how interpretable something is, and how interpretable it needs to be. There are a lot of factors in the framework, but they are most likely to have a different heft to them, depending on the context. The entire framework concerns XAI interpretability. However, don't we mean the interpretability of the explanation generated by XAI? Another similar issue is perception versus behaviour. For example, you can see the recall response time as a behaviour. However, trustworthiness will remain a perception of that user: not on the same level of abstraction. It would be very nice to make it more practical, to in the end give the framework to someone and say: start here, follow the guidelines and proceed clockwise. Shape it more in the form of an action plan perhaps.	concerning the XAI. Therefore, constant adjustments are necessary.
Interview 9	The most important factors are: RRT, ITR, intuitive understandability/ relevance, model fidelity. Change intuitive understandability to affordance. the relationship between model fidelity and trustworthiness is believed to be not straight forward. It is believed that something is missing: truthfulness could be added between the two factors.	All factors included in the framework are properties of different things. For example, RRT is a property belonging to a person. But then, using probabilities belongs to the explanation. And lastly, model fidelity belongs to the XAI model. Define more trade-offs in the key principles. The expert agrees that both the framework and the key principles are important. Without the key principles, it's difficult to interpret the framework. Whilst without the framework, the key principles appear to be coming out of nowhere. A lot of things in the framework are not measurable. They are all very difficult to measure. A lot of factors are dependent on the person that received the explanation. It is all about perceived interpretability. However, it is not the end of the world if measurability remains a problem for this framework.	Categorizing into the three groups might also help resolve challenges amongst stakeholders, when you are struggling with a factor and you know it's regarding the user, all stakeholders are aware that the factors are differently perceived by different users
Interview 10	Overlap with human understanding can be combined with coherence with prior beliefs	Lots of factors, perhaps categorize or focus on key takeaways, perhaps create a checklist from them to enhance practical usability: make it more practical and usable. Rather use case (context) dependent.	Many stakeholders will have different goals
Interview 11	Transparency may be very important to experts, however, is transparency actually relevant for laypeople? Making the explanation more transparent, could even overcomplicate things for laypeople. The five most important factors are	Implementation in practice would probably need one more step, to make the framework better in practice. For example, create XAI explanation design guidelines or create a questionnaire or form to fill. It's very difficult to measure the factors and to 'crunch hard numbers' according to the framework. However, for system auditing for example, the framework by itself is already very useful.	Different stakeholders will definitely have different requirements on what they want to have explained by the framework. You can use extended forms for different stakeholders. Based on what you fill into the form for

	trustworthiness, model fidelity, intuitive understandability, clarity and comprehensiveness.	Perhaps include something to evaluate whether or not you have fulfilled the key principles. Think like A/B testing.	example, you get steered into a different direction.
Interview 12	Some factors within the framework do indeed require explanations themselves to be fully understood. The top five factors for assessing XAI interpretability are (in descending order): intuitive understandability, relevance, comprehensiveness, generalizability, and clarity.	Is the framework focused on the XAI model or the XAI explanation? Scoping the framework specifically to textual explanations might be good. Including metrics for evaluation purposes, such as scores for each factor, would enhance operationalization in practice.	Including evaluation metrics would allow for stakeholder agreement.

Appendix F. Survey for evaluation on laypeople

Section 1. Introduction

The following message will be displayed right before the participant enters the survey:

"Dear participant,

My name is David and for my Master Thesis I am conducting a survey to evaluate a new framework I'm creating, specifically designed to assess the interpretability of explanations provided by Explainable Artificial Intelligence (XAI) systems for laypeople on the topic of the XAI.

For this survey, imagine an app on your phone called SymptomSolver. This app can be very useful! Suppose you are feeling unwell. You are however not sure what causes these symptoms. Is it some sort of viral or bacterial infection? Or is something else going on. You do not yet feel like going to a doctor, but you would like to get a better understanding in what it is that you might have. That is where the app SymptomSolver comes in handy. In the app, you can fill in your symptoms and SymptomSolver provides you with the most likely cause for these symptoms. And it provides an explanation to go with that cause.

In this survey, you will be presented with a total of six sets of explanations. Your task is to read the scenario and then evaluate the explanations of that set based on the questions provided. Please note that it may be difficult to perfectly rank the explanations each time. However, do still rank the explanations, no matter how subtle the difference may be to you. After evaluating all sets of explanations, you will be answering three small questions about yourself.

Note that your participation is completely voluntary, and all responses will be kept anonymous and confidential. By clicking next, you informedly agree that your results will be confidentially used for scientific purposes.

Your feedback will be invaluable in helping us refine and improve the framework for future use. Therefore, thank you very much for your time and participation. Please click "Next" to begin the survey."

Section 2. Evaluation of explanations

In section 2, a total of six sets of explanations are provided in separate sections. Each section starts by setting the context. It does so by providing the symptoms that the participant is supposedly feeling. After which the explanations that are (again, supposedly) provided by the app SymptomSolver is presented. The respondent will be asked to rank those explanations from better to worse. First of all, in terms of satisfaction. Second of all, in terms of natural understandability. Both of these terms adequately represent interpretability. However, directly asking participants about the interpretability of an explanation is believed to be less adequate. What the participant will see is the following:

"Explanation set 1: You have recently been experiencing headaches, a stuffy nose, and a sore throat. You decide to use a self-diagnosis app, SymptomSolver, to determine the potential cause of your symptoms.

You could get three different explanations:

- *Explanation 1:* Your symptoms of headache, stuffy nose, and sore throat are indicative of a common cold, which is a viral infection affecting the upper respiratory system.

There are over 200 viruses that can cause the common cold, with rhinoviruses being the most common. Cold symptoms typically appear within 1-3 days after exposure to the virus and last about a week. The immune system usually clears the infection on its own, and treatment focuses on managing symptoms. Rest, staying hydrated, and using over-the-counter medications like pain relievers and decongestants can help alleviate symptoms. Practicing good hygiene and avoiding contact with infected individuals can help prevent the spread of the cold virus.

- Explanation 2: Your symptoms suggest you have a common cold, which is often caused by viruses affecting the respiratory system. Getting plenty of rest and staying hydrated can help you recover. Over-the-counter medications may provide relief for your symptoms.

- Explanation 3: Based on your symptoms of headache, stuffy nose, and sore throat, it's likely that you have a common cold. This condition is caused by various viruses and usually resolves on its own within a week. To manage your symptoms, you can take over-the-counter medications, rest, and drink plenty of fluids.

First of all, please rank the explanations based on how naturally understandable they are to you.

(The participant ranks the explanations)

After having done that, please rank the explanations based on how satisfying they are to you.

(The participant ranks the explanations)"

This will be done for a total of six explanation sets. These sets can be found below. These are the explanations that are presented to the user in the survey, needless to say, without naming any of the factors, as they are presented below. Merely the explanations and the context above is presented in the survey itself. Please note that the example above was used to illustrate how the first set will be asked.

Set 1: Comprehensiveness, transparency, simplicity, and generalizability

Comprehensiveness and transparency vs. simplicity and generalizability: Explanations that are more comprehensive (are automatically more transparent) tend to cover more aspects of a problem, while simpler and more generalizable explanations are easier to interpret. Striking a balance between these two factors is important for interpretability.

Context: You have recently been experiencing headaches, a stuffy nose, and a sore throat. You decide to use a self-diagnosis app, SymptomSolver, to determine the potential cause of your symptoms.

1. High comprehensiveness and high transparency, little simplicity and little generalizability: Your symptoms of headache, stuffy nose, and sore throat are indicative of a common cold, which is a viral infection affecting the upper respiratory system. There are over 200 viruses that can cause the common cold, with rhinoviruses being the most common. Cold symptoms typically appear within 1-3 days after exposure to the virus and last about a week. The immune system usually clears the infection on its own, and treatment focuses on managing symptoms. Rest, staying hydrated, and using over-the-counter medications like pain relievers and decongestants can help alleviate symptoms. Practicing good hygiene and avoiding contact with infected individuals can help prevent the spread of the cold virus.
2. High simplicity and high generalizability, little comprehensiveness and little transparency: Your symptoms suggest you have a common cold, which is often caused

by viruses affecting the respiratory system. Getting plenty of rest and staying hydrated can help you recover. Over-the-counter medications may provide relief for your symptoms.
3. Balanced comprehensiveness, transparency, simplicity, and generalizability: Based on your symptoms of headache, stuffy nose, and sore throat, it's likely that you have a common cold. This condition is caused by various viruses and usually resolves on its own within a week. To manage your symptoms, you can take over-the-counter medications, rest, and drink plenty of fluids.

In this set of explanations, the first one is highly comprehensive and transparent, providing a detailed understanding of the common cold, its causes, and treatment options. The second explanation is simpler and more generalizable, offering a brief and easy-to-understand overview of the common cold. The third explanation strikes a balance between comprehensiveness, transparency, simplicity, and generalizability, providing an accessible yet informative explanation of the common cold and its management.

Set 2: Complexity, transparency, simplicity, and clarity

Complexity and transparency vs. simplicity and clarity: Simple and clear explanations make it understandable how a model arrives at its conclusions, while more complex and transparent explanations might provide deeper insights. However, increased complexity can make it harder for laypeople to understand the reasoning behind the model's decisions.

Context: You have been experiencing fatigue, muscle aches, and a low-grade fever for the past few days. You decide to use a self-diagnosis app, SymptomSolver, to determine the potential cause of your symptoms.

These are the explanations that are presented to the user in the survey, needless to say, without naming the factors. Merely the explanations and the context above is presented in the survey itself:

1. High complexity and high transparency: Your symptoms of fatigue, muscle aches, and low-grade fever suggest a possible viral infection, such as the influenza virus. Influenza is an acute respiratory illness caused by the influenza A or B virus, which targets epithelial cells lining the respiratory tract. The virus enters the cells through endocytosis and replicates, leading to the production of viral proteins and the assembly of new virions. As the immune system responds to the infection, various cytokines are released, causing systemic symptoms like fever, muscle aches, and fatigue. The best course of action is to rest, stay hydrated, and manage symptoms with over-the-counter medications. In some extreme cases, antiviral medications may be prescribed by a healthcare professional within the first 48 hours of symptom onset.
2. High simplicity and high clarity: Your symptoms indicate that you may have the flu, which is a common viral infection. Resting, drinking fluids, and taking over-the-counter medications can help you feel better.
3. Balanced complexity, transparency, simplicity, and clarity: Based on your symptoms of fatigue, muscle aches, and low-grade fever, it's likely that you have the flu. This is a viral infection that affects the respiratory system and is usually managed with rest, hydration, and over-the-counter medications to alleviate symptoms. In some extreme cases, your healthcare professional may prescribe antiviral medication.

In this set of explanations, the first one is highly complex and transparent, providing a detailed understanding of the influenza virus, its mechanism of infection, and the immune response. The second explanation is simple and clear, giving a concise and easy-to-understand overview of the flu and its management. The third explanation balances complexity, transparency, simplicity, and clarity, offering a moderately detailed yet accessible explanation of the flu and its treatment options.

Set 3: Abnormality, coherence with prior beliefs, and affordance

Abnormality vs. coherence with prior beliefs and affordance: Explanations that focus on abnormalities can provide more compelling insights, but they may contradict users' prior beliefs. Striking a balance between highlighting abnormalities and maintaining coherence with prior beliefs can help users accept and understand explanations better.

Context: You have been experiencing sudden episodes of dizziness and lightheadedness, especially when standing up quickly. You decide to use a self-diagnosis app, SymptomSolver, to determine the potential cause of your symptoms.

These are the explanations that are presented to the user in the survey, without naming the factors. Merely the explanation.

1. High abnormality: Your symptoms of dizziness and lightheadedness when standing up quickly could be a rare side effect of a medication you're currently taking. Some medications can cause orthostatic hypotension, which is a sudden drop in blood pressure when standing up, leading to dizziness and lightheadedness.
2. High coherence with prior beliefs and affordance: Your symptoms of dizziness and lightheadedness when standing up quickly are likely due to a common and generally harmless condition called orthostatic hypotension. This occurs when blood pressure drops suddenly upon standing, resulting in temporary dizziness. Drinking more water, standing up slowly, and avoiding prolonged standing can help alleviate these symptoms.
3. Balanced abnormality, coherence with prior beliefs, and affordance: Your symptoms of dizziness and lightheadedness when standing up quickly could be due to orthostatic hypotension, which is a sudden drop in blood pressure upon standing. This condition can be caused by various factors, such as dehydration, prolonged bed rest, or even as a side effect of certain medications. To manage this issue, ensure you stay well-hydrated, rise slowly when standing up, and discuss any medications you're taking with your healthcare professional.

In this set of explanations, the first one focuses on the abnormality, suggesting a rare side effect of medication as the cause. The second explanation emphasizes coherence with prior beliefs and affordance, presenting a common and understandable condition as the cause. The third explanation balances abnormality, coherence with prior beliefs, and affordance, acknowledging various potential causes, including the side effect of medication, while also providing actionable advice for symptom management.

Set 4: Intentionality and actionability

Intentionality vs. actionability: Explanations that focus on intentionality provide insights into why events occurred, while actionable explanations offer practical advice for addressing a problem. Balancing these factors can help users understand the causes behind a situation and identify effective solutions.

Context: You have recently noticed that your hands become shaky and tremble when performing fine motor tasks like writing or picking up small objects. You decide to use a self-diagnosis app, SymptomSolver, to determine the potential cause of your symptoms.

These are the explanations that are presented to the user in the survey, without naming the factors. Merely the explanation.

1. High intentionality: Your symptoms of shaky hands and trembling during fine motor tasks could be due to an underlying neurological condition, such as essential tremor or Parkinson's disease. These conditions affect the communication between your brain and muscles, resulting in involuntary movements and tremors.

2. High actionability: To reduce the shakiness and trembling in your hands, try implementing some practical strategies, such as using heavier pens or utensils, practicing relaxation techniques, and avoiding caffeine, which can exacerbate tremors. If your symptoms persist or worsen, consult a healthcare professional for further evaluation.
3. Balanced intentionality and actionability: The shakiness and trembling in your hands during fine motor tasks could be caused by a neurological condition, such as essential tremor or Parkinson's disease, which affects the communication between your brain and muscles. To manage your symptoms, try using heavier pens or utensils, practicing relaxation techniques, and avoiding caffeine. If your symptoms persist or worsen, it is important to consult a healthcare professional for proper diagnosis and treatment.

In this set of explanations, the first one focuses on intentionality, explaining the possible neurological causes behind the symptoms. The second explanation emphasizes actionability, providing practical advice for symptom management. The third explanation balances intentionality and actionability, offering both insights into potential causes and actionable steps to alleviate the symptoms.

Set 5: Model fidelity and explanation fidelity

Model fidelity vs. explanation fidelity: High model fidelity means that explanations accurately represent the underlying model, while high explanation fidelity means that explanations are faithful to the real-world context. Ensuring both model and explanation fidelity is difficult, yet important to provide accurate and useful explanations.

Context: You have recently experienced a severe, but gradually presenting headache, along with nausea and sensitivity to light. You decide to use a self-diagnosis app, SymptomSolver, to determine the potential cause of your symptoms.

These are the explanations that are presented to the user in the survey, without naming the factors. Merely the explanation.

10. High model fidelity: According to the SymptomSolver's algorithm, your symptoms of a severe headache, nausea, and sensitivity to light are highly correlated with a migraine. The model has identified these symptoms as key indicators of a migraine based on patterns observed in large datasets of patient experiences.
11. High explanation fidelity: Your symptoms of a sudden, severe headache, nausea, and sensitivity to light are consistent with a migraine, a common neurological disorder that can be triggered by various factors, such as stress, hormonal changes, or certain foods. Migraines are usually characterized by intense, throbbing pain on one side of the head, often accompanied by additional symptoms like those you're experiencing.
12. Balanced model and explanation fidelity: The SymptomSolver's algorithm has identified your symptoms of a severe headache, nausea, and sensitivity to light as being highly correlated with a migraine based on patterns observed in large datasets of patient experiences. In the real-world context, migraines are a common neurological disorder characterized by intense, throbbing pain on one side of the head, and can be accompanied by symptoms like those you're experiencing. Various factors, such as stress, hormonal changes, or certain foods, can trigger migraines.

In this set of explanations, the first one focuses on model fidelity, explaining the algorithm's reasoning behind the diagnosis based on patterns in data. The second explanation emphasizes explanation fidelity, describing the real-world context of migraines and their common symptoms. The third explanation balances model and explanation fidelity, providing insights into both the algorithm's diagnostic process and the real-world context of migraines.

Set 6: Trustworthiness and relevance

Trustworthiness vs. Relevance: Trustworthy explanations come from credible sources, but they might not always be relevant to the user's specific goals or context. Finding a balance between trustworthiness and relevance is important to ensure that explanations are both reliable and useful.

Context: You have been experiencing a persistent dry cough, fatigue, and shortness of breath for several days. You decide to use a self-diagnosis app, SymptomSolver, to determine the potential cause of your symptoms.

These are the explanations that are presented to the user in the survey, without naming the factors. Merely the explanation.

1. High trustworthiness: According to a recent publication in a leading medical journal, your symptoms of a persistent dry cough, fatigue, and shortness of breath are commonly associated with a respiratory infection, such as bronchitis. The research article is authored by a team of renowned experts in the field of respiratory medicine.
2. High relevance: Your symptoms of a persistent dry cough, fatigue, and shortness of breath suggest that you might be experiencing a respiratory infection. It is important to seek medical advice as soon as possible, as untreated respiratory infections can lead to complications, especially if you have pre-existing health conditions or a weakened immune system.
3. Balanced trustworthiness and relevance: A recent publication in a leading medical journal, authored by a team of renowned experts in the field of respiratory medicine, indicates that your symptoms of a persistent dry cough, fatigue, and shortness of breath are commonly associated with a respiratory infection, such as bronchitis. It is crucial to seek medical advice as soon as possible to prevent potential complications, particularly if you have pre-existing health conditions or a weakened immune system.

In this set of explanations, the first one emphasizes trustworthiness by citing a credible source (a leading medical journal) and renowned experts in the field. The second explanation focuses on relevance, providing practical advice for the user's specific situation. The third explanation balances trustworthiness and relevance, combining the credibility of the source with pertinent advice tailored to the user's needs.

Section 3. Who are you?

The third section provides us with information of the participant. The most important question is the fourth question of this section. Since the framework is intended to be used on laypeople, and not on experts, it must be ensured that answers given by medical experts (however interesting they may be) are not taken into account during analysis of final results. Additionally, the third question is what is known as an attention check. As explained in the methodology chapter, this question allows us to, to some extent, guarantee the quality of the responses.

“Finally, I would like to know something more about your background. To that extent, please answer the following four questions:

1. *Have you ever used an Artificially Intelligent (AI) system before?*
    a. *Yes*
    b. *Maybe*
    c. *No*

2. *Have you ever received an explanation from an AI system before?*
    a. *Yes*
    b. *Maybe*
    c. *No*

3. *How do you view the following statement: I have never used a computer-like device before?*
    a. *Yes, this is true for me.*
    b. *No, this is not true for me.*

4. *Would you consider yourself a medical expert? Are you for example a doctor or medicine student?*
    a. *Yes*
    b. *No*

5. *If yes: why would you consider yourself a medical expert?"*

Section 4. Final statement

The final section of the survey provides a very small final message to the participant. It also includes contact information in case the participant has any questions about the survey.

Thank you for your participation in this survey. Your input is invaluable for my research! Suppose you have any questions about this research or what will happen with the results you have given, please feel free to contact David on D.A.Lensen@student.tudelft.nl.

Appendix G. Survey Data Cleaning

Before cleaning the data, a total of 204 responses were captured by the survey.

Section 1. Laypeople – expert distinction

To ensure that only the results of laypeople in the medical field are being used for processing, medical experts are being filtered out. In total, 8 respondents have presented themselves as medical experts (see Figure 20).

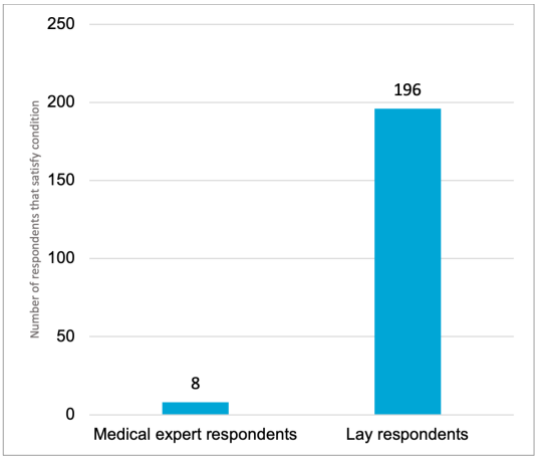

Figure 20. Number of experts respondents in the survey

In the final question, these 8 respondents had to argue why they are medical experts. The following answers were presented:

- *'Im a 4th year medical student'*
- *'I am a family doctor by profession.'*
- *'I'm a clinical biochemist'*
- *'I am a Holistic therapist and studied anatomy, physiology and pathology to qualify'*
- *'I am studying cardiology at university'*
- *'I am a healthcare professional'*
- *'I am a doctor'*
- *'Medical intern'*

All responses are considered valid. Therefore, these 8 respondents are filtered out, keeping 196 responses in the dataset.

Section 2. Attention check

By distributing the survey on the Prolific platform, participants are being paid to complete it, which introduces external motivation (financial compensation) as opposed to internal motivation (a genuine desire to contribute to sound research). This distinction in motivation could potentially lead to biased or less reliable results, as some participants may rush through the survey or provide low-quality responses just to receive the payment. Therefore, as a measure to mitigate the potential risk of processing low quality results, an attention check has been added to the survey. This has been done by adding the following question to the survey:

How do you view the following statement: I have never used a computer-like device before?

a. *Yes, this is true for me.*
b. *No, this is not true for me.*

Since all surveys are being filled in on either a smartphone, tablet, or computer (all computer-like devices), everyone should be answering B to this question. All responses that have not

selected answer B, will automatically be discarded from the final dataset. In total, only 2 out of the remaining 196 respondents have failed to answer the attention check satisfactory. Therefore keeping 194 responses in the dataset after this step. See Figure 21.

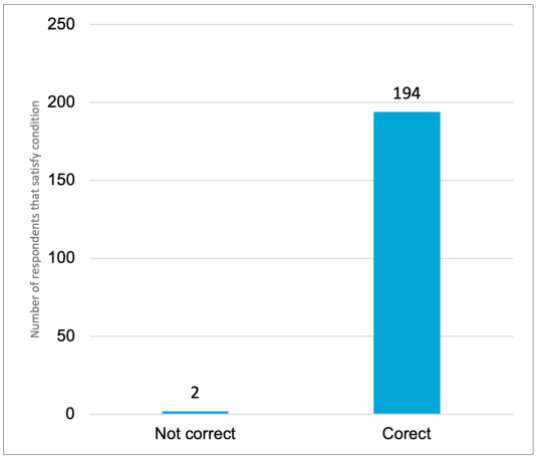

Figure 21. Number of respondents that have failed to complete attention check

Section 3. Amount of time taken

Thirdly, the amount of time that people took was considered when cleaning the data.

To further mitigate the potential risk of receiving low quality data, an estimation has been made that the survey should take the average person 8 minutes to complete. This duration is a benchmark to ensure that participants are dedicating an appropriate amount of time and attention to the survey questions, which should result in more thoughtful and accurate responses. To further reduce the risk of biased or unreliable data, the decision has been made to exclude all respondents who took less than 4 minutes to complete the survey. This threshold has been created by assuming that it would be possible for people that are capable of reading rather quickly and processing information quickly, to take half the amount of time as the average person would. This threshold is set to filter out participants who may have rushed through the survey without carefully considering their answers. By removing these respondents from the dataset, the quality of the remaining responses should be higher, leading to more reliable insights and conclusions.

This process of excluding participants based on their response time aims to minimize the impact of external motivation on the survey results and to encourage participants to be more engaged and attentive when providing their answers. The ultimate goal is to collect high-quality data that genuinely reflects the opinions and experiences of the participants, leading to more accurate and robust findings from the research.

Figure 22 presents a histogram that shows the respondents and a distribution of their time taken. Based on this histogram, a further 37 responses were excluded from the survey, therefore having a final number of 157 responses in the dataset. Two other relevant metrics to base the 4-minute threshold on, could only be presented after the survey has been completed by all respondents. These include the mean and median time of filling in the survey. The average time is 7 minutes and 35 seconds and the median time is 6 minutes and 23 seconds. This further justifies that excluding all responses below 4 minutes is feasible.

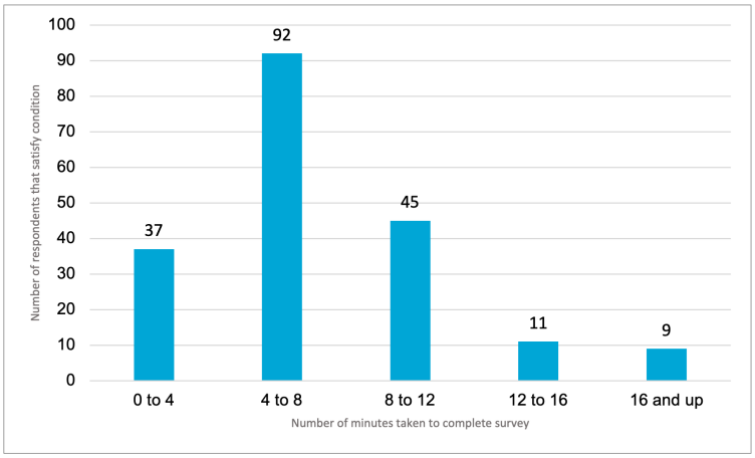

Figure 22. Histogram of time taken to complete survey

Appendix H. Demographic data

As explained in appendix G, three measures were undertaken to ensure high validity. After cleaning the data, 157 responses were used in the final analysis. This section will provide insights into how the survey was distributed. Answering the question: 'Who are we basing our results on?'. It should be noted participants either had the chance to revoke consent on demographic questions, or that the data was no longer considered up to date. This can be found in the graphs.

Section 1. Gender

Figure 23 displays the gender distribution of survey participants. Initially, the survey aimed to achieve a perfectly equal distribution of male and female respondents to avoid gender bias in the data. However, after the data cleaning process, the balance between genders might have been slightly altered. The results in Figure 23 show the final gender distribution of participants after the data cleaning process. This gender distribution is an important consideration when interpreting the findings, as it ensures that the study captures a variety of perspectives and experiences from both male and female respondents.

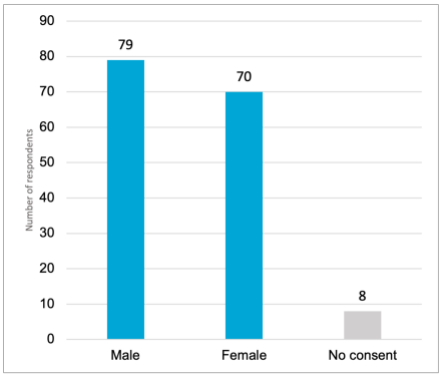

Figure 23. Gender distribution

Section 2. Age

The age distribution of the survey respondents is presented in Figure 20. The survey received responses from a wide range of age groups, covering almost all adult ages from 18 to 70, with the exception of ages 18, 64, and 69. The age range of 20 to 40 is most prominently represented in the sample. This can be explained by reasoning that older adults may be less likely to have access to or be comfortable using the digital platforms through which the survey was distributed. This could have made it difficult for them to participate in the survey. However, this is not considered a problem, since XAI users will most likely have the same age distribution.

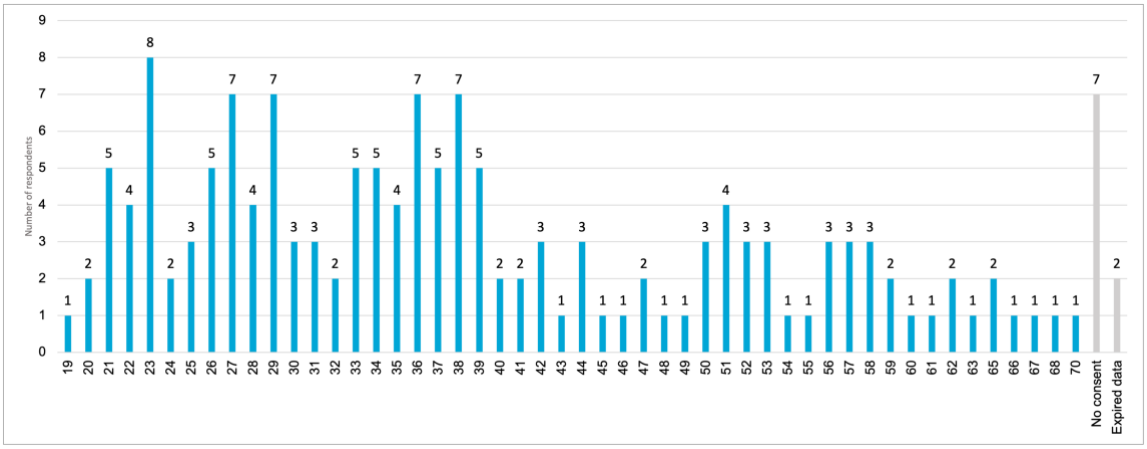

Figure 24. Age distribution

Section 3. Geographical location

The geographical location of the survey participants is illustrated from Figure 25 to Figure 29. A majority of the respondents were from the United Kingdom (124), followed by the Netherlands (18), and the United States (13). In total, there were 157 participants across these three countries. The study therefore predominantly includes perspectives from respondents in the UK, with some representation from the Netherlands and the US. This geographical distribution should be considered when interpreting the findings and generalizing the results, as the experiences and opinions of individuals in these locations may differ from those in other countries or regions.

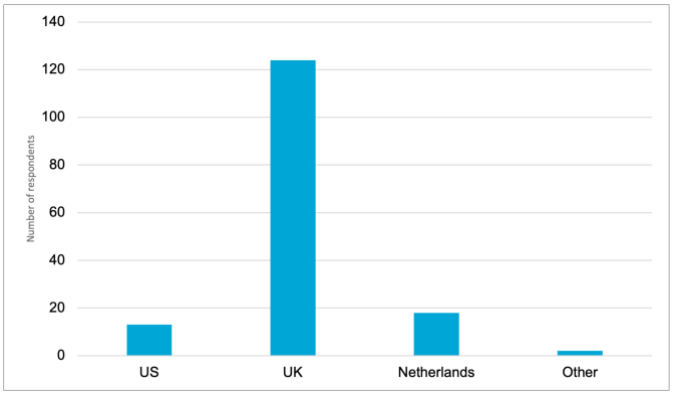

Figure 25. Geographical distribution

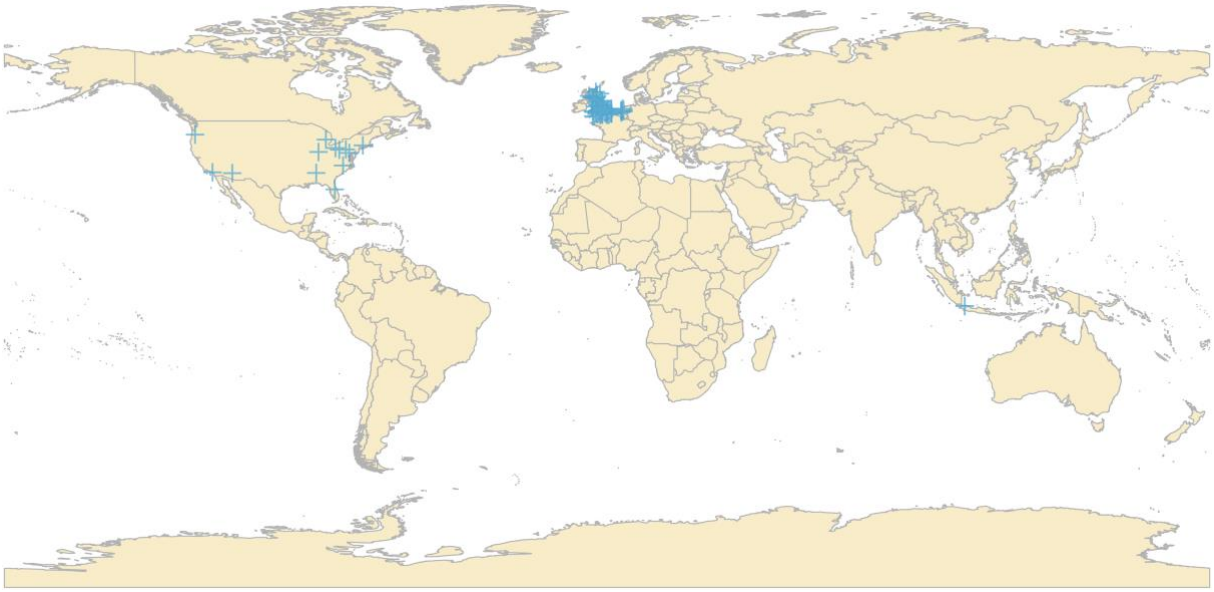

Figure 26. Participant distribution across the world (n = 157)

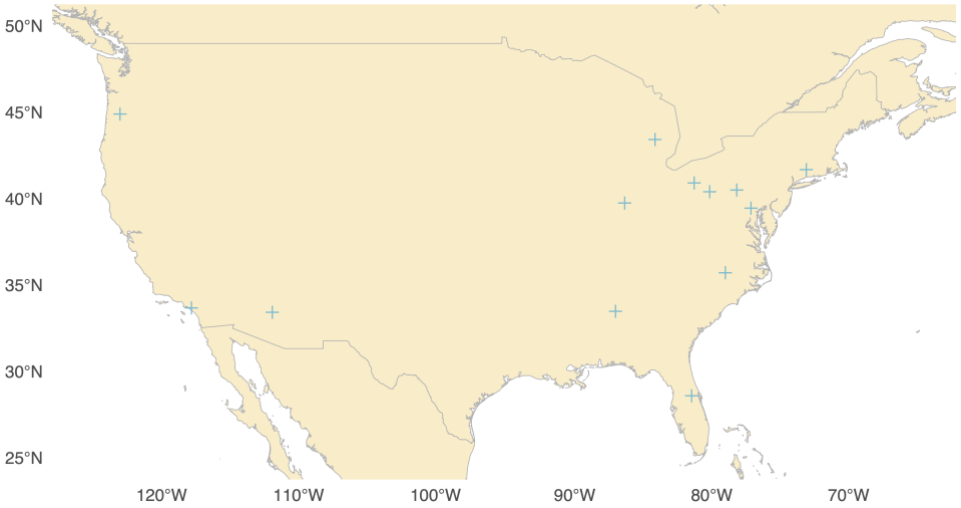

Figure 27. Participant distribution across the United States (n = 13)

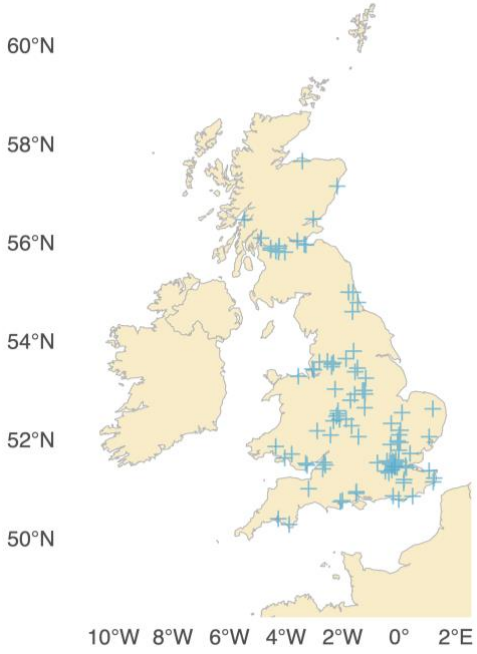

Figure 28. Participant distribution across the United Kingdom (n = 124)

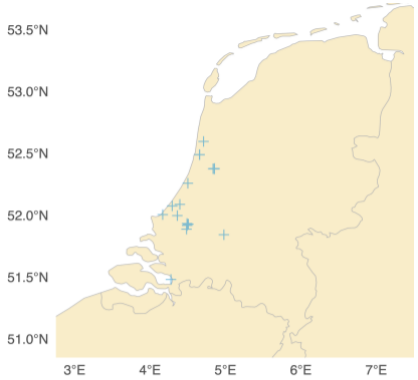

Figure 29. Participant distribution across the world (n = 18)

Section 4. Ethnicity

Related to geographical location, is ethnicity of participants, as shown in Figure 22. Since the majority of respondents are located in the UK, the predominant ethnical category of the survey respondents was Caucasian (117). There was also representation from other ethnic backgrounds, including Asian (8), Black (3), mixed (8), and other (2). While there is some diversity in the ethnic backgrounds of the participants, the overrepresentation of Caucasian respondents may limit the generalizability of the study's findings to other ethnic groups.

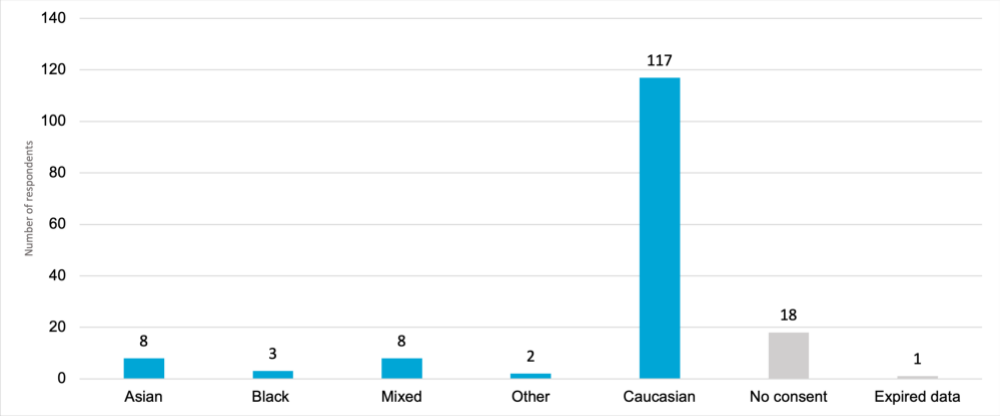

Figure 30. Ethnic distribution

Section 5. Employment status

The survey participants were diverse in terms of their employment status, as seen in Figure 31. The majority of the respondents were employed full-time (57), followed by part-time employees (22). A smaller number of respondents were either unemployed and seeking work (8), not in paid work (10), or due to start a new job within the next month (1). Additionally, one participant indicated "other" for their employment status.

This diversity in employment status helps to ensure that the findings of the study reflect the perspectives of individuals with varying work experiences and backgrounds. However, the relatively high number of participants with expired data or no consent to share their employment status could limit the generalizability of the findings related to employment.

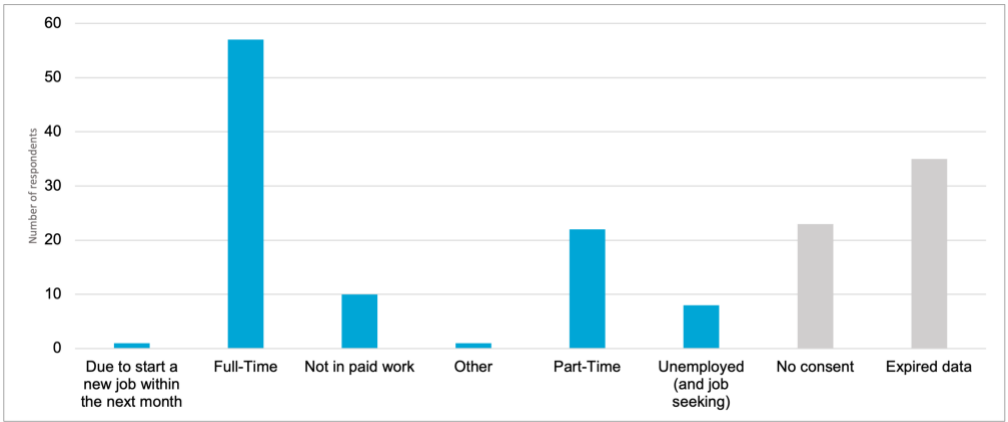

Figure 31. Employment distribution

Section 6. Student status

The student status of the participants is depicted in Figure 32. A total of 23 respondents identified as students, while 90 respondents were not students. The inclusion of both student and non-student respondents in the survey helps to capture a wider range of experiences and perspectives related to the research topic. However, the high number of expired data and no consent responses may affect the generalizability of the findings related to student status.

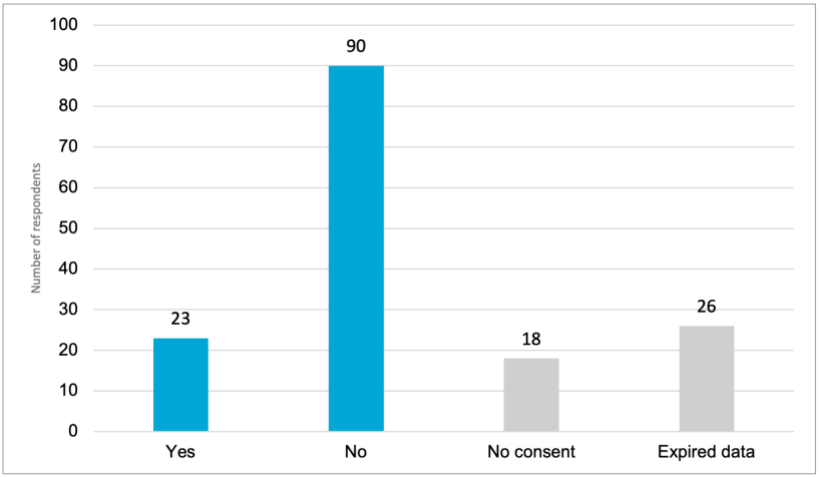

Figure 32. Student distribution

Section 7. AI & XAI experience

Figure 33 displays the distribution of survey participants based on their experience with AI and XAI. Participants were asked to report their experience in these two areas by answering the following two questions:

1. *Have you ever used an Artificially Intelligent (AI) system before?*
2. *Have you ever received an explanation from an AI system before?*

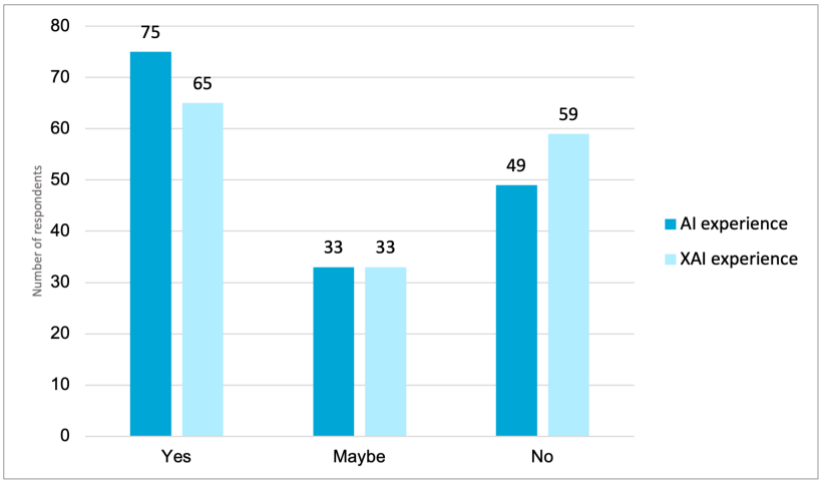

Figure 33. AI & XAI experience distribution

The responses regarding AI experience reveal a diverse among the participants. A significant number of respondents (75) reported having experience with AI, while 33 participants were unsure ("maybe") and 49 indicated no experience with AI. This diversity in AI experience helps to capture a variety of perspectives and understanding of AI, potentially leading to more comprehensive insights into the research topic. Regarding XAI experience, the distribution of responses was also diverse. A total of 65 participants reported having experience with XAI, while 33 were unsure, and 59 indicated having no experience with XAI. This range of XAI experience among respondents allows the study to explore different levels of familiarity with

XAI and its interpretability. Combining these two questions presents us with the fact that there are fewer people that indicated having experience with XAI in comparison two AI. Most likely, there were 10 people that only (thought) they had experience with regular AI, as opposed to XAI.

The distribution of AI and XAI experience among the participants is an important factor to consider when interpreting the findings of the study. The diverse range of experiences provides a broader perspective on the research topic and may contribute to a more nuanced understanding of XAI interpretability. However, potential biases may arise if certain experience levels are over- or underrepresented in the sample.

Appendix I. Human Research Ethics: DMP

TPM - MSc How Interpretable is Explainable?

0. Administrative questions

1. Name of data management support staff consulted during the preparation of this plan.

My faculty data steward, Nicolas Dintzner, has reviewed this DMP on the first of March 2023.

2. Date of consultation with support staff.

2023-03-01

I. Data description and collection or re-use of existing data

3. Provide a general description of the type of data you will be working with, including any re-used data:

Type of data	File format(s)	How will data be collected (for re-used data: source and terms of use)?	Purpose of processing	Storage location	Who will have access to the data
Interview date and time	Word file 1 (any text file would do)	Looking at the date and time	To be able to distinguish easily between different interviews.	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier
First name of the XAI developers and experts.	Word file 1 (any text file would do)	During the 'recruiting' phase.	To be able to distinguish easily between different interviews.	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier
Organization affiliated with the XAI developer/ expert	Word file 1 (any text file would do)	This will be the second question of the interview.	To gain a better understanding in the nature of the expert/ developer	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier
Email address of the XAI developer/ expert	Word file 1 (any text file would do)	During the 'recruiting' phase.	For communication purposes	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier
Comments about the created framework, by XAI developers & experts. No personal information. Solely data about the framework.	Word file 1 (any text file would do)	Through interviews, both in person as well as online interviews	To gain a better understanding into the efficiency and workability of the created framework.	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier
The expert/ developer interviews will be	MP4 file 1	Through recording the interviews, both in person as well as online interviews	To gain a better understanding into the efficiency and workability of the created framework.	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier
Case study date and time	Word file 2 (any text file would do)	Looking at the date and time	To be able to distinguish easily between different case studies.	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier
First name of the lay-user of the XAI system.	Word file 2 (any text file would do)	During the 'recruiting' phase.	To be able to distinguish easily between different case studies.	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier
Email address of the layperson	Word file 2 (any text file would do)	During the 'recruiting' phase.	For communication purposes	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier
Comments about several interpretability related factors regarding an multiple different XAI explanations. Provided by lay-users of the XAI.	Word file 2 (any text file would do)	Through a case study.	To gain a better understanding into the efficiency and workability of the created framework.	TUD's OneDrive	Me, Aaron Ding, Marcus Westberg, and potentially Martijn Warnier

4. How much data storage will you require during the project lifetime?

- < 250 GB

II. Documentation and data quality

5. What documentation will accompany data?

- Methodology of data collection

III. Storage and backup during research process

6. Where will the data (and code, if applicable) be stored and backed-up during the project lifetime?

- OneDrive

IV. Legal and ethical requirements, codes of conduct

7. Does your research involve human subjects or 3rd party datasets collected from human participants?

- Yes

8A. Will you work with personal data? (information about an identified or identifiable natural person)

If you are not sure which option to select, ask your Faculty Data Steward for advice. You can also check with the privacy website or contact the privacy team: privacy-tud@tudelft.nl

- Yes

I will collect the utmost minimal required personal data (only their name, email, and (if applicable) affiliated organization). This will reduce the ethical implications of the research.

8B. Will you work with any other types of confidential or classified data or code as listed below? (tick all that apply)

If you are not sure which option to select, ask your Faculty Data Steward for advice.

- No, I will not work with any confidential or classified data/code

9. How will ownership of the data and intellectual property rights to the data be managed?

For projects involving commercially-sensitive research or research involving third parties, seek advice of your Faculty Contract Manager when answering this question. If this is not the case, you can use the example below.

The project is an internal TU Delft Msc project. The datasets underlying the published papers will NOT be publicly released at any time during or after the research. During the active phase of research, both me, and the first supervisor of the research (Aaron Ding) from TU Delft will oversee the access rights to data (and other outputs), as well as any requests for access from external parties.

10. Which personal data will you process? Tick all that apply

- Signed consent forms
- Data collected in Informed Consent form (names and email addresses)
- Photographs, video materials, performance appraisals or student results

- Names and addresses
- Email addresses and/or other addresses for digital communication

Only names (not addresses).

11. Please list the categories of data subjects

1. XAI developers-experts
2. Laypeople concerning the contents of the XAI (an ordinary Logistics TU Delft researcher for example will most likely be a layperson on an XAI that determines the disease on a self-diagnosis application).

12. Will you be sharing personal data with individuals/organisations outside of the EEA (European Economic Area)?

- No

15. What is the legal ground for personal data processing?

- Informed consent

16. Please describe the informed consent procedure you will follow:

All study participants will be asked for their written consent for taking part in the study and for data processing before the start of the interview.

17. Where will you store the signed consent forms?

- Same storage solutions as explained in question 6

18. Does the processing of the personal data result in a high risk to the data subjects?

If the processing of the personal data results in a high risk to the data subjects, it is required to perform a Data Protection Impact Assessment (DPIA). In order to determine if there is a high risk for the data subjects, please check if any of the options below that are applicable to the processing of the personal data during your research (check all that apply).
If two or more of the options listed below apply, you will have to complete the DPIA. Please get in touch with the privacy team: privacy-tud@tudelft.nl to receive support with DPIA.
If only one of the options listed below applies, your project might need a DPIA. Please get in touch with the privacy team: privacy-tud@tudelft.nl to get advice as to whether DPIA is necessary.
If you have any additional comments, please add them in the box below.

- None of the above applies

22. What will happen with personal research data after the end of the research project?

- Personal research data will be destroyed after the end of the research project
- Anonymised or aggregated data will be shared with others

After the defense and publication into the TU Delft repository are both completed, all data will be completely anonymized. Names, email addresses and organizations will all be removed from the data. Audio files will be destroyed completely.

23. How long will (pseudonymised) personal data be stored for?

- Other - please state the duration and explain the rationale below

We will not be storing (pseudonymised) personal data.

24. What is the purpose of sharing personal data?

- Other - please explain below

We will not be sharing personal data.

25. Will your study participants be asked for their consent for data sharing?

- Yes, in consent form - please explain below what you will do with data from participants who did not consent to data sharing

V. Data sharing and long-term preservation

27. Apart from personal data mentioned in question 22, will any other data be publicly shared?

- All other non-personal data (and code) underlying published articles / reports / theses

The anonymized aggregated data will be used to create the final thesis. After destroying the non-anonymized data, all that is left are interview transcriptions with non-identifiable people and comments about the framework.

29. How will you share research data (and code), including the one mentioned in question 22?

- My data will be shared in a different way - please explain below

It can be shared, however, we will choose not to share it. The anonymized and aggregated data used in the report can be found directly in the report in TUD educational repository. The raw (also anonymized, not aggregated) data will be hold onto by dr. Aaron Ding.

30. How much of your data will be shared in a research data repository?

- < 100 GB

31. When will the data (or code) be shared?

- As soon as corresponding results (papers, theses, reports) are published

32. Under what licence will be the data/code released?

- CC BY
- Other - Please explain

Since its msc thesis: it will be under the same license as the master thesis. If possible: CC BY.

VI. Data management responsibilities and resources

33. Is TU Delft the lead institution for this project?

- Yes, the only institution involved

34. If you leave TU Delft (or are unavailable), who is going to be responsible for the data resulting from this project?

This will be Dr. Aaron Ding. In the end, it is most likely that the non aggregated, anonymized data will also be destroyed. As there is no apparent purpose for it.

35. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Most likely only a small amount of time after the end of the project.

Appendix J. Human Research Ethics: Informed Consent

Informed consent form

You are being invited to participate in a research study titled 'How Interpretable is Explainable?'. This study is being done by master student David Lensen from the TU Delft, under supervision of Dr. Aaron Ding, Dr. Marcus Westberg, and Prof. Dr. Martijn Warnier.

The purpose of this interview/study is to gain insights into the workability and efficiency of the newly created framework that helps when assessing XAI interpretability, and will take you approximately 25 minutes to complete. The data will be used for evaluation purposes. We will be asking you to evaluate the framework by answering multiple questions.

As with any (online) activity the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. We will minimize any risks by asking you as little personal information as possible, and storing all information in a secure manner. Once no longer necessary, the (possible) audio recording will be deleted, and an anonymous summary will be used. This summary will be included in the master thesis and will be made publicly available in the TU Delft educational repository.

Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any questions. You can contact David on d.a.lensen@student.tudelft.nl at any time.

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
3. I agree to the fact that this interview will be recorded, knowing that the audio recording will be destroyed as soon as a summary of the interview is made.	☐	☐
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
4. I understand that taking part in the study also involves collecting specific personally identifiable information (PII) through my name and email address with the potential risk of my identity being revealed.	☐	☐
5. I understand that the following steps will be taken to minimise the threat of a data breach and protect my identity in the event of such a breach: anonymization of data, secure data storage, using summary.	☐	☐
6. I understand that personal information collected about me that can identify me, such as my name, email address, and potential audio recording, will not be shared beyond the study team.	☐	☐
7. I understand that the (identifiable) personal data I provide will be destroyed after the research is completed at last, or earlier once deemed no longer necessary.	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
8. I understand that the summary of our discussion will be used for evaluating an XAI framework.	☐	☐
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
9. I give permission for the summary of the discussion to be archived in the TU Delft repository (in the form of part of one of the appendices to the thesis) so it can be used for future research and learning.	☐	☐

Signature

_____	_____	_____
Name of participant	Signature	Date

Contact details for further information: David Lensen – +316 28250018 – D.A.Lensen@student.tudelft.nl

Appendix K. Human Research Ethics: Approval

Date 10-Mar-2023
Contact person Dr. Cath Cotton, Policy Advisor Academic Integrity
E-mail c.m.cotton@tudelft.nl

Human Research Ethics Committee
TU Delft
(http://hrec.tudelft.nl/)

Visiting address
Jaffalaan 5 (building 31)
2628 BX Delft

Postal address
P.O. Box 5015 2600 GA Delft
The Netherlands

Ethics Approval Application: How interpretable is explainable?
Applicant: Lensen, David

Dear David Lensen,

It is a pleasure to inform you that your application mentioned above has been approved.

In addition to any specific conditions or notes, the HREC provides the following standard advice to all applicants:
• In light of recent tax changes, we advise that you confirm any proposed remuneration of research subjects with your faculty contract manager before going ahead.
• Please make sure when you carry out your research that you confirm contemporary covid protocols with your faculty HSE advisor, and that ongoing covid risks and precautions are flagged in the informed consent - with particular attention to this where there are physically vulnerable (eg: elderly or with underlying conditions) participants involved.
• Our default advice is not to publish transcripts or transcript summaries, but to retain these privately for specific purposes/checking; and if they are to be made public then only if fully anonymised and the transcript/summary itself approved by participants for specific purpose.
• Where there are collaborating (including funding) partners, appropriate formal agreements including clarity on responsibilities, including data ownership, responsibilities and access, should be in place and that relevant aspects of such agreements (such as access to raw or other data) are clear in the Informed Consent.

Good luck with your research!

Sincerely,

Dr. Ir. U. Pesch
Chair HREC
Faculty of Technology, Policy and Management