



C-XAI: Design Method for Explainable AI Interfaces to Enhance Trust Calibration

Mohammad Naiseh

Bournemouth University

**A thesis submitted in partial fulfilment of the requirements of
Bournemouth University for the degree of Doctor of Philosophy**

[June] 2021

Copyright

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Abstract

Human-AI collaborative decision-making tools are on an accelerated rise in several critical application domains, such as healthcare and military sectors. It is often difficult for users of such systems to understand the AI reasoning and output, particularly when the underlying algorithm and logic are hidden and treated as a black-box for commercial sensitivity and also for challenges of its explainability. Lack of explainability and opacity of the underlying algorithms can perpetuate justice and bias and decrease users' acceptance and satisfaction. Integrating eXplainable AI (XAI) into AI-based decision-making tools has become a crucial requirement for a safe and effective human-AI collaborative environment.

Recently, the impact of explainability on trust calibration has become a main research question. The role refers to how explanations and their communication method to help form a correct mental model of the AI-based tool; thus, the human decision-maker is better informed on whether to trust or distrust the AI recommendations. Although studies showed that explanations could improve trust calibration, such studies often assumed that users would engage cognitively with explanations to calibrate their trust. Recent studies showed that even though explanations are communicated to people, trust calibration is not improved. Such failure of XAI systems in enhancing trust calibration has been linked to factors such as cognitive biases, e.g., people can be selective of what they read and rely on. Also, other studies showed that XAI failed to improve calibrated trust due to the inconsistency in properties of XAI methods which are rarely considered in the XAI interfaces design. Overall, users of XAI systems fail, on average, to calibrate their trust, human decision-makers working collaboratively with an AI can still be notably following incorrect recommendations or rejecting correct ones.

This thesis aims to provide C-XAI, a design method expressly tailored to help trust calibration in the XAI interface. The method identifies properties of XAI methods that may introduce trust calibration risks and help produce designs that mitigate these risks. Trust calibration risk is defined in this thesis as a limitation in the interface design that may hinder users' ability to calibrate their trust. This thesis followed a qualitative research approach with experts, practitioners, and end-users who used AI-based decision-making tools in their work environment. The data collection methods included a literature review, semi-structured interviews, think-aloud sessions, and a co-design approach to develop C-XAI. These data collection methods helped conceptualise various aspects of trust calibration and XAI, including XAI requirements during Human-AI collaborative decision-making tasks, trust calibration risks, and design principles that help trust calibration. The results of these studies were exploited to devise C-XAI. The C-XAI then was evaluated with domain experts and end-users. The evaluation aimed to investigate the effectiveness, completeness, clarity, engagement, and communication between different stakeholders. The evaluation results showed that the method helped stakeholders understand the design problem and develop XAI designs to help trust calibration.

This thesis has four main contributions. First, it conceptualises the trust calibration design problem concerning XAI interface design. Second, it elicits main limitations for XAI interfaces design to support trust calibration. Third, it proposes key design principles that support XAI interface designers to support trust calibration. Finally, the thesis proposes and evaluates the C-XAI design method to guide XAI interface design to enable trust calibration systematically.

Table of Contents

Copyright.....	I
Abstract	II
Acknowledgement.....	X
Glossary.....	XII
1. Chapter 1: Introduction.....	1
1.1 Research problem and motivation.....	2
1.2 Research questions	3
1.3 Research objectives.....	3
1.4 Methodology overview	5
1.5 Thesis structure.....	6
1.6 Publications arising from this thesis.....	9
1.6.1 Declaration of Co-authors contribution.....	9
1.7 Chapter summary.....	10
2. Chapter 2: Literature review.....	11
2.1 Explainable Artificial Intelligence	11
2.1.1 The need for explainability	12
2.1.2 Terminology identification	12
2.1.3 Explainability goals	13
2.1.4 Explainability approaches.....	14
2.1.5 Explainability approaches properties	15
2.1.6 How to measure and evaluate explainability	16
2.1.7 Human-Computer Interaction aspects.....	18
2.2 Deigning for trust calibration	26
2.2.1 Calibrated trust definition and conceptualisation	26
2.2.2 The role of humans' error.....	29
2.2.3 The role of explainability	30
2.2.4 Calibrated trust measurements	31
2.3 Chapter summary.....	32
3. Chapter 3: Research method.....	33
3.1 Research paradigms	33
3.1.1 Positivism	34
3.1.2 Realism.....	34
3.1.3 Interpretivism	35
3.1.4 Pragmatism	35
3.1.5 Rationale for selecting Pragmatism	35
3.2 Research approaches	36
3.3 Research strategies.....	37
3.3.1 Qualitative research approach.....	37
3.4 Time horizons	40
3.5 Adopted data collection methods.....	40
3.5.1 Data collection.....	40
3.5.2 Analysis methods and tools	43
3.6 Design methods	45
3.6.1 Co-design.....	45
3.6.2 Scenario-based approach.....	46
3.7 Software design approaches.....	46
3.7.1 User-centred design	47
3.7.2 Participatory design	48
3.7.3 Value-sensitive design	49

3.7.4	User stories	50
3.8	Research ethics	50
3.9	Chapter summary	51
4.	Chapter 4: A Taxonomy for Explainable AI classes	54
4.1	Introduction	54
4.1.1	Identify different Terminologies	54
4.2	Taxonomy for machine learning explanations	55
4.2.1	Research goal	56
4.2.2	Research methodology	56
4.2.3	Results	60
4.2.4	Forming users' questions and explainable classes taxonomy- prior evaluation..	66
4.2.5	Validation	68
4.3	Chapter summary	72
5.	Chapter 5: Explainability and Calibrated trust: The role of explanation class	74
5.1	Introduction and theoretical background	74
5.2	Research method.....	76
5.2.1	Human-AI task description	77
5.2.2	Explanation classes	78
5.2.3	Study design	79
5.2.4	Study procedure and data collection	80
5.2.5	Participants.....	82
5.2.6	Data analysis	82
5.3	Findings	83
5.3.1	Explanation classes impact on user trust (RQ1.1).....	83
5.3.2	Explanation classes impact on trust calibration process (RQ1.2).....	85
5.3.3	User requirements for enhanced trust calibration	87
5.4	Discussion	91
5.5	Limitations of the study	94
5.6	Chapter summary	94
6.	Chapter 6: Systematic users' errors with AI-based explanations.....	96
6.1	Research rationale	96
6.2	Research method.....	98
6.2.1	Recruitment and participants	98
6.2.2	Consent procedure	99
6.2.3	Study procedure	99
6.2.4	Data analysis	100
6.2.5	Strengths and limitations	100
6.3	Results	101
6.3.1	Skipping explanations	102
6.3.2	Misapplying explanations	104
6.4	Discussion	106
6.5	Chapter summary	107
7.	Chapter 7: XAI interface design principles to help trust calibration	109
7.1	Theoretical background and related work.....	109
7.2	Research method.....	111
7.2.1	Use case and underpinnings	112
7.2.2	Participants.....	113
7.2.3	Co-design	114
7.2.4	Data analysis	116
7.3	Results	117
7.4	Discussion	126
7.5	Conclusion	131

8.	Chapter 8. C-XAI: A method for designing XAI interfaces to help trust calibration	132
8.1	C-XAI method overview	132
8.2	Chapter research goal.....	134
8.3	Phase 1: Identification process	136
8.3.1	Identifying method stakeholders and sampling process	136
8.3.2	Identifying Human-AI collaborative tasks	138
8.3.3	Eliciting users' explainability needs for each task.....	139
8.4	Assessment phase.....	142
8.4.1	Assessing the explanation method	143
8.4.2	Conducting trust calibration risk assessment.....	146
8.5	The selection and implementation phase.....	153
8.6	The evaluation phase	154
8.7	C-XAI workflow	156
8.8	Good design practice for C-XAI.....	157
8.9	Chapter summary	159
9.	Chapter 9: Evaluating C-XAI method.....	160
9.1	Evaluation goals	161
9.2	Evaluation context.....	161
9.3	Evaluation case Study.....	162
9.4	Evaluation study phases.....	162
9.4.1	Phase 2: Expert evaluation.....	164
9.4.2	Phase 3: Designing XAI interface with C-XAI method.....	165
9.5	Study protocol	167
9.6	Participants recruitment process	168
9.7	Data collection and analysis methods.....	169
9.8	Findings	170
9.8.1	Expert evaluation findings	170
9.8.2	Designing XAI interface with C-XAI method findings	173
9.9	Benefits of C-XAI in the design process	183
9.9.1	Effective communication and increased engagement	183
9.9.2	Increased focus on the design problem	184
9.9.3	Increased empathy towards users	184
9.9.4	Better software product design.....	184
9.9.5	Risks of using C-XAI method in the design process	185
9.10	Evaluation validity	185
9.11	Chapter summary.....	186
10.	Chapter 10: Discussion, Future work and Conclusion.....	187
10.1	Research questions and objectives revisited	188
10.2	Contribution to knowledge.....	190
10.3	Thesis limitations	191
10.4	Recommendations for future research.....	192
11.	References.....	194
12.	Appendix.....	216
12.1	Appendix 1 - Explainable models taxonomy validation.....	216
12.2	Appendix 2 – The role of XAI class on trust calibration	222
12.3	Appendix 3. Explainable Recommendation and Calibrated trust: two systematic users' errors 228	
12.4	Appendix 4. Calibrated trust design principles study material.....	234

List of Figures

Figure 1 Thesis chapters and research road map.....	8
Figure 2 Theory-driven user-centric explainable ai framework introduced by Wang et al. (2019).	19
Figure 3 Description of calibrated trust and trust resolution (adopted from Lee and See 2004)	27
Figure 4 Okamura and Yamada (2018) framework for calibrated trust.....	28
Figure 5 Ekman et al. (2017) for calibrated trust	29
Figure 6 Research onion (Saunders et al.2009).....	33
Figure 7 Research approaches according to saunders et al. 2009.....	36
Figure 8 A description of the grounded theory Introduced by CHUN tie et al. (2019).	39
Figure 9 Thesis objectives and research methods.....	52
Figure 10 Comparison between the terms “Interpretable machine learning” and “explainable machine learning” from January 2004 until April 2021.	55
Figure 11 Global Interpretable example introduced by (Guidotti et al. 2018b)	61
Figure 12 Two examples of rule extraction models presented in (Zhou et al. 2003) and Johansson et al. (2004)	62
Figure 13 Counterfactual example-based explanation example (Wachter et al.2017).....	64
Figure 14 Workflow for prescription screening aided by AI-based decision-making tool.	77
Figure 15 A sample of prescribing system interface supported with AI recommendations.	78
Figure 16 Study workflow	80
Figure 17 Mean cognition-based trust components rating per explanation class. Explanation cognition-based trust ratings range from Strongly Disagree (a rating of 1) to Strongly Agree (a rating of 5).	84
Figure 18 mean of trust calibration behavioural indicators	85
Figure 19 Switch percentage and agreement percentage for incorrect recommendations across different XAI classes.....	86
Figure 20 Study workflow	100
Figure 21 Participants' interaction behaviour with AI-based explanations	102
Figure 22 Screening prescription classification AI-based system classification.....	112
Figure 23 A sample of prescribing system interface supported with AI recommendations	113
Figure 24 Co-design session workflow	115
Figure 25 An example of two levels of abstraction design presented in our co-design study ...	118
Figure 26 An example of suggested control techniques designs for Local explanations in the co- design sessions	120
Figure 27 Suggested visual cues resulting from the design phase.....	123
Figure 28 Friction design example for calibrated trust goal	130
Figure 29 Stages of THE CAT method.....	134
Figure 30 C-XAI method activities.....	135
Figure 31 Curiosity checklist	138
Figure 32 HAI-A analysis sheet	139
Figure 33 Intrinsic elicitation for model-agnostic explanations	141
Figure 34 Explanation method functional requirements assessment sheet.....	144
Figure 35 Explanation method operational requirements assessment sheet	145
Figure 36 Explanation method usability assessment sheet.....	146
Figure 37 Explanation method Safety assessment sheet.....	146
Figure 38 Explanation method validation assessment sheet.....	146
Figure 39 Trust calibration risk assessment sheet – Functional dimension	148
Figure 40 Trust calibration risk assessment sheet – Operational dimension	149
Figure 41 Trust calibration risk assessment sheet – Usability dimension	150
Figure 42 Trust calibration risk assessment sheet – Safety dimension	151
Figure 43 Trust calibration risk assessment sheet – Validation dimension	153
Figure 44 C-XAI design method workflow	156

Figure 45 C-XAI method evaluation phases.....	164
Figure 46 Screenshot from with C-XAI method evaluation session	173
Figure 47 HAI-A analysis sheet provided to participants	174
Figure 48 Sample responses for assessing lime operational requirements	175
Figure 49 Sample responses for assessing LIME usability requirements	176
Figure 50 Samples of participants answers to trust calibration risk assessment	179
Figure 51 Final XAI interface generated from C-XAI method.....	182
Figure 52 Multi-stage qualitative study workflow	235
Figure53 Provided patients' profile in the design sessions	237
Figure 54 Global feature importance	238
Figure 55 Local feature importance explanation	239
Figure 56 Counterfactual explanation	239
Figure 57 Example-based explanations using KNN.....	240
Figure 58 Design space provided to our participants	240
Figure 59 Sample of the collected data.	241

List of Tables

Table 1 Mapping the thesis research questions, objectives and chapters.....	5
Table 2 Explanation Styles examples	21
Table 3 Data collection strategies associated with each philosophical approach.....	36
Table 4 Data collection methods used in this thesis	41
Table 5 Participants familiarity with relevant topics	56
Table 6 Participants assessment survey results	57
Table 7 Selected research databases	57
Table 8 Final search string.....	59
Table 9 descriptive statistics for each database search results	59
Table 10 Five main explanations classes	60
Table 11 The categorisation of the reviewed papers.....	65
Table 12 Mapping between users' questions and explainable models	66
Table 13 The structure of Taxonomy evaluation study.....	68
Table 14 The provided documents during the focus group session	69
Table 15 Final taxonomy	70
Table 16 Users' questions	71
Table 17 Population details	82
Table 18 Table 2 Issues and needs applicability to explanation classes	87
Table 19 Study population	98
Table 20 Human-explanation cognition-based trust components definitions adapted from (Madson and Gregor, 2000)	110
Table 21 Participants demographics for exploration and design stage	113
Table 22 The four main themes that emerged from the co-design phase.....	117
Table 23 Description for methods' activities	135
Table 24 C-XAI method stakeholders.....	136
Table 25 Criteria and rationale for stakeholders recruitment	137
Table 26 Users' needs elicitation methods	141
Table 27 Trust calibration risks design guidelines	154
Table 28 Calibrated trust behavioural measures.....	155
Table 29 Evaluation methods for Users' engagement by (Doherty and Doherty, 2018)	155
Table 30 Expert participants demographics.....	165
Table 31 Expert evaluation guided questions	165
Table 32 Stakeholders evaluation.....	168
Table 33 Summary of expert evaluation phase	172
Table 34 Samples of implemented design techniques used by participants	181
Table 35 Scenarios characteristics. Scenarios numbers do not represent the order of presentation.	224
Table 36 Four examples of four patients' profiles presented in the scenarios.	225

Acknowledgement

I would like to say a big thank you to my first supervisor Dr Nan Jiang for his patience, understanding, guidance, and motivation throughout my PhD journey. He has encouraged me to exercise professionalism and taught me about the significance of paying attention to the small details, which has made me a much better researcher. Without his constant support, expertise, and experience, my research goal would not have been achieved. I was fortunate to have a supervisor like you, and it was a privilege to work with you. I would also like to thank my second supervisor Prof Raian Ali. I have learned from him lots of skills and knowledge that helped me to complete my PhD thesis. Raian has encouraged me to be involved in other activities to increase my visibility and opportunities in the academic field. His guidance aided my entire PhD journey, starting from conceptualising the research problem to writing up the thesis. I admire your attitude and the effort you invest in your work. Finally, I would also like to thank all the supervisory team, Dr Shamal Faily, Dr Shahin Rostami, Dr Jianbing Ma and Prof Tamas Hickish, for their valuable feedback and comments during multiple stages of my PhD journey.

I would also like to express my appreciation and thanks to Bournemouth University and IQHealthTech for the PhD Studentship, research facilities, and research support. I could not have pursued my dream of getting a PhD and becoming a researcher without their financial support.

Big thanks also to my friends and colleagues in the Computing and Informatics Department at Bournemouth University for their support, encouragement, discussions and advice, especially during some difficult times in the PhD. To all those who participated in my studies, thank you for your input. Without your valuable input, I could not have achieved the thesis aims.

Finally, I want to acknowledge the love, support, and encouragement of my family and friends, particularly my father Nidal, mother, and beautiful sisters. Without your support, patience and understanding, I could not have completed my PhD journey.

Glossary

XAI interface	A user interface that presents an explanation generated from the eXplainable AI method.
XAI method	It refers to an algorithm or a model that generates an explanation from an AI.
Trust calibration	It refers to a process where users of an AI can appropriately judge the capability and limitations of an AI.
Human-AI collaborative environment	It refers to a family of applications where the human decision-maker interacts with an AI to perform a task.
Explanation instability	Explanation model ability to generate similar explanations for similar input.
XAI method assumptions	It refers to any functional assumptions that have been made during the development of the XAI method. Specifically, constraints concerning the input and output of the XAI method, e.g., XAI method for black and white images only.
Explanation misuse	This property refers to generating the right information to the right users, i.e., avoiding providing information irrelevant and redundant information to users. Failing to achieve this goal could lead to explanation misuse
Coherence	Explanation ability to be consistent with the knowledge and beliefs of end-users.
Complexity	Explanation level of details.
Contextual information	Information provided with the explanation help users to interpret the explanation and prevent forming incorrect conclusions.
Completeness	This property tells the users of the explanation model how well the explanation can be generalised beyond specific recommendations.
Soundness	This property measures how well the explanation is accurate with the underlying machine learning model.

Novelty	Explanation ability to provide surprising and unexpected information to end-users.
Actionability	Explanation ability to provide guidelines to the desired outcome.
Causality	Explanation ability to provide a causal relationship between the cause and the effect.
Explanation Scope	This property tells the users of the explanation method to what extent an explanation can be generalised.

1. CHAPTER 1: INTRODUCTION

Artificial Intelligence (AI), especially Machine Learning algorithms (ML), have become widely used in various everyday Human-AI applications, ranging from commercial recommender systems to social robots. While these algorithms can provide impressive performance and accuracy in many applications (Abdul et al., 2018), full automation to ML algorithms is not implemented yet. This is because of their probabilistic and dynamic nature, which means there is no guarantee of correctness for a particular ML recommendation (Zhang et al., 2020). Furthermore, the ML model's accuracy depends on the historical data used to train them, and this data may suffer from input errors, unknown flaws and biases (Bansal et al., 2019). ML models are mostly used as assisted decision-making tools to support human decision-makers toward a collaborative decision outcome. Researchers showed such approaches increased the effectiveness and accuracy of humans' decision-making (Tulabandhula and Rudin, 2014, Char et al., 2018).

A fundamental success for AI-based assisted decision-making tools is to support users in forming a correct mental model of when to follow (trust) the recommendation of the system or when to reject it (distrust) (Bansal et al., 2019). When users fail to calibrate their trust, collaborative decision-making would be affected, and dramatic failures could happen in high-stakes application domains (Dietvorst et al. 2015). The research community discussed the challenges for humans to understand and develop an accurate mental model of an AI since opaque ML black-box models are increasingly used (Springer, 2019, Zhang et al., 2020, Bansal et al., 2019). For instance, users may fail to follow up with the AI-based recommendations because of their dynamic and uncertain nature (Zhang et al., 2020a). In such cases, users might follow an incorrect recommendation (over-trust) or reject a correct recommendation (under-trust). A design goal that aims to attain and manage trust refers to calibrated trust. This thesis emphasises that this goal is distinct from enhancing trust in AI. For example, enhancing users' trust could be done by providing indicators and metrics for AI abilities and performance (Yin et al., 2019). Enhancing trust also does not require AI users to understand and develop accurate mental models of the AI. On the other hand, calibrating users' trust may require extra effort from the users' and shall be done on the recommendation level where decision-makers can identify situations when to rely on the AI and use their judgment (Yu et al., 2019).

Humans' decision-makers require a user interface to reflect the current state or logic of the recommendation to attain trust calibration. Studies such as (Muir, 1994, Zhang et al., 2020b, Yang et al., 1994) have emphasised the role of eXplainable AI (XAI) in calibrating users' trust. Such communication is a major facilitator of trust calibration within a Human-AI collaborative decision-making task. Explanations help calibrated trust by showing the rationale and reasoning behind single recommendations and their overall logic (Cai et al. 2019). However, such studies

often assumed that users would engage cognitively with explanations and form correct interpretations from them. Recent studies showed that even though explanations are communicated to people, trust calibration is not improved (Zhang et al., 2020a, Bussone et al., 2015). Such failure of XAI systems in enhancing trust calibration has been linked to factors such as humans' cognitive biases, e.g., people are selective of what they read and rely on (Naiseh et al., 2020b). Also, others showed that XAI failed to improve calibrated trust because of undesired human behaviour with AI-based explanations, e.g., human laziness to engage in what they perceived as effortful behaviour (Wagner and Robinette, 2021). Overall, users of XAI systems fail, on average, to calibrate their trust, i.e., human decision-makers working collaboratively with an AI can still be notably following incorrect recommendations or rejecting correct ones. The main focus of the literature was to suggest explanation types (Zhang et al., 2020a) and presentation format (Tomsett et al., 2020) for improved trust calibration. Still, the research did not provide structured studies to explore how users interact with AI-based explanations, what mistakes users do with AI-based explanations and why. Such knowledge might be required to design XAI interfaces that help trust calibration.

Furthermore, recently a surge of advances in explainability led to increasing interest in model-agnostic explanation methods which interpret any machine learning model by focusing primarily on the input and the output of the machine learning model (Hohman et al. 2019). Model-agnostic explainable models generate different classes of explanations to answer different user questions (Carvalho et al. 2019). This approach is motivated by preserving the models' confidentiality, increasing the cost-efficiency of generating the explanation, and increasing its usability (Feng and Boyd-Graber, 2019, Zhang et al., 2020, Sokol and Flach, 2019). Different model-agnostic methods may generate explanations with distinct explanation output, but they may vary in their performance, fidelity and completeness of the underlying AI model (Arrieta et al., 2020). Model-agnostic explanation methods have a wide range of properties and features that may not be appropriate for a particular Human-AI task or require an intervention on the design level to operationalise these explanations (Sokol and Flach, 2020). For example, an XAI method that has a high probability to generate novel explanations may need the design team to make the explanation noticeable and engage the users with the explanations in each interaction. Neglecting the properties of the XAI method and its relation to the design was another factor to explainability limitations to enhance trust calibration (Sokol and Flach, 2020a).

1.1 RESEARCH PROBLEM AND MOTIVATION

Due to the aforementioned reasons in the introduction, current methods for designing XAI interfaces, such as user-centred approaches e.g. (Eiband et al., 2018a), would not be applicable to designing an XAI interface with a trust calibration goal in mind. XAI interface design for trust calibration may require more attention than the explanation ease of use, explanation

content and format (Miller, 2019). The design may need to debias users' behaviour and persuade them to read explanations. For instance, distraction in a typical system is to avoid, but it might not be the case for XAI as a distraction may be needed to nudge people towards looking at explanation. Furthermore, the variety of XAI classes and their properties would require systematic approach to evaluate XAI methods and their generated explanations on trust calibration.

This thesis aims to conceptualise the trust calibration design problem concerning XAI interfaces and propose C-XAI design method that aids XAI interfaces' designers to help trust calibration. It will help them understand the design problem, identify explanation properties that may introduce trust calibration risks and propose design principles that mitigate them. This thesis will take the AI-based screening prescription tool as an exemplary case study. The choice of this application domain was to reflect an everyday Human-AI collaborative decision-making task where trust calibration errors are possible. To the researcher's best knowledge, this research will be the first of its kind, and it will lay the foundation for further research in this area.

1.2 RESEARCH QUESTIONS

Based on the thesis aim, the questions below are developed to set the boundary of this research.

- **RQ1** – As an explanation can be related to an XAI class and have different properties, what is the role of XAI class and explanation properties in calibrating users' trust and what makes explanations effective in calibrating users' trust?
- **RQ2** – What risks related to XAI interfaces hinder users' ability to calibrate their trust?
- **RQ3** – How the design of XAI interfaces can play a role in helping trust calibration?
- **RQ4** – How can the findings of RQ1, RQ2 and RQ3 be used to develop a design method for enhancing trust calibration given an explanation method.

1.3 RESEARCH OBJECTIVES

In order to answer the thesis research questions, the following objectives were formulated.

- **Objective 1: To conduct a literature review on calibrated trust design, XAI and related areas.**

The research will conduct a literature review around Human-AI trust and trust calibration. Furthermore, the review will look into other elements, including XAI from the HCI and AI perspectives, and software design approaches. The review strives to gain background knowledge and further understand relevant concepts related to this research. Achieving this objective will support the studies to be conducted in the thesis and the thesis solution design.

- **Objective 2: To provide a classification of XAI methods.**

To provide a classification for different XAI methods in the literature of XAI, a systematic literature review will be conducted. Specifically, the review will look at a family of XAI methods that can explain any black-box ML model, i.e., model-agnostic XAI methods. The research will then map several users' questions that each XAI class could answer. Finally, the research will conduct an expert evaluation to evaluate the completeness, coherence and understandability of the generated classification. The goal of this objective is to provide a reference point for the stakeholders of the C-XAI method about what can be explained to users given a black-box model. Furthermore, the results from this stage will be used in further user studies.

- **Objective 3: To explore the lived experience of XAI users during a Human-AI collaborative decision-making task.**

The exploration is built based on the results from Objective 1 and Objective 2. Objective 3 informs the research regarding the role of XAI class to calibrate users' trust and potential trust calibration risks. This Objective will be achieved through a multi-stage qualitative approach which includes semi-structured interviews and a think-aloud protocol. Categories of trust calibration risks and XAI interface requirements to help trust calibration will be provided. These results will be used as a basis for exploring design principles for enhancing trust calibration in the XAI interface. Furthermore, it could enable the design method users to become more informed about why users' may fail to calibrate their trust with the XAI interface, and how to support trust calibration in the design.

- **Objective 4: To propose XAI interface design principles that help trust calibration.**

This objective explores different design principles that could guide the design of the XAI interface to enhance trust calibration. These principles can provide C-XAI design method users with various strategies for supporting trust calibration and mitigating potential risks. To achieve that, a co-design method will be conducted with end-users, in which the researcher and users will propose the design principles that could help explainability design to support trust calibration. Co-design will consider the classification of trust calibration risks suggesting how these risks could be mitigated.

- **Objective 5: To create, evaluate and refine a method for helping trust calibration in the XAI interface.**

The method helps system analysts and requirement engineers identify the required XAI classes for a given Human-AI task. C-XAI method also uses findings from Objectives 1,2,3 and 4 to develop templates for assessing XAI methods and identifying their potential trust calibration risks, as well as selecting appropriate design principles to mitigate those risks. The design

method is meant for system analysts and the design team who would like to enhance the role of explainability in calibrating users' trust. The proposed method will enable end-users, AI experts, and other stakeholders to be involved in the design process.

The evaluation will follow a case study approach. The evaluation will be performed from both the experts and the user's perspective. The evaluation is meant to assess C-XAIs' ability to aid the XAI interface design and also to refine the proposed templates and supporting documents. The evaluation focuses on the effectiveness of the method to help C-XAI users in understanding the design problem. The evaluation also considers the completeness and the clarity of the C-XAI method and its supporting material. It also examines the C-XAIs' engagement and the communication between different stakeholders.

TABLE 1 MAPPING THE THESIS RESEARCH QUESTIONS, OBJECTIVES AND CHAPTERS

Research Question	Objectives Outcome	Chapters
RQ1	Objective 2	Chapter 4
	Objective 3	Chapter 5
RQ2	Objective 3	Chapter 5
		Chapter 6
RQ3	Objective 4	Chapter 7
RQ4	Objective 5	Chapter 8
		Chapter 9

1.4 METHODOLOGY OVERVIEW

The research methodology followed in this thesis is discussed in **Chapter 3**. This chapter explores various research methods, potential options and the selected methodological approaches for this thesis. This section briefly summarises the adopted research methods that will be employed to achieve each objective as follow:

Objective 1: The research reviews the literature concerning XAI, Human-AI trust and trust calibration. Also, software design approaches are reviewed.

Objective 2: A systematic literature review and expert validation methods will be followed. The systematic literature review will focus on classifying different XAI methods used in XAI literature to generate explanations given a black-box model. The literature review will provide the research with a comprehensive taxonomy for different classes of XAI methods. In this stage, a mapping between XAI classes and different users' questions that each XAI class could answer is also provided. This is meant to provide a reference point for C-XAI design method users in eliciting users' explainability needs for a given Human-AI task. Expert evaluation study will be

conducted to gather their opinions about the completeness, validity and clarity of the emerged concepts. The evaluation also focuses on assessing the mapping between users' questions and XAI classes.

Objective 3: To achieve Objective 3, a multi-stage qualitative approach will be followed. This will include three studies. This approach aims to explore the lived experience of participants during a Human-AI collaborative decision-making task, and their usage and interaction style with XAI interfaces. This is meant to identify potential users' errors and difficulties while interacting with AI-based explanations. Furthermore, this approach aims to explore the role of XAI class in calibrating users' trust and what makes the explanation effective in calibrating users' trust. This study will use a think-aloud protocol during a decision-making task study, semi-structured interviews, and follow-up interviews in terms of data collection. This would allow the research to explore different trust calibration risks and allow participants to discuss their opinions and concerns regarding AI-based explanations in their everyday Human-AI collaborative decision-making task.

Objective 4: The research will explore design principles that can enhance the role of explainability in calibrating users' trust. Based on the findings from Objective 3, the research has identified various trust calibration risks and requirements for XAI interface to calibrate users' trust. Findings from Objective 4 will illustrate how to implement AI-based explanations in an XAI interface to enhance the explanation role to calibrate users' trust. A co-design approach will be followed to gather how the solution would look from the users' perspective. In this stage, the researcher discussed and negotiated with representative users' ways of utilising AI-based explanations to serve their needs, enhance trust calibration, and mitigate potential trust calibration risks. This has been achieved by giving the participants initial prototypes of the design problem to help them visualise the idea and then provoke brainstorming related to the design problem. The deriving framework to discover protentional design solutions was digital nudging (Caraban et al.) and also the principles of de-biasing (Soll et al., 2014).

Objective 5: The researcher will develop a method to assist designers and system analysts in developing XAI interfaces to help trust calibration. The method adopts a participatory design approach to ensure that all relevant stakeholders are involved in the early and later stages of XAI interface design. The method includes seven activities supported with templates and supporting documents from previous objectives. The method will be evaluated using a case study approach to assess the effectiveness, completeness and clarity of the methods. The evaluation also focuses on the engagement and communication between stakeholders.

1.5 THESIS STRUCTURE

This thesis structure is illustrated in **Figure 1**. In **Chapter 2**, a literature review is presented; this chapter addresses various relevant topics related to the research problem. The research

methodology followed in this thesis is discussed in **Chapter 3**. It discusses various research approaches, potential choices, and the selected methods in this thesis. In **Chapter 4**, results from a systematic literature review on XAI methods is presented. **Chapter 5** and **Chapter 6** discussed the findings from multi-stage qualitative studies to conceptualise trust calibration related to XAI interfaces. **Chapter 7** presents Co-design findings focusing on design principles and techniques that can help trust calibration in XAI interface. In **Chapter 8**, the C-XAI method is presented. **Chapter 9** presents the results from a two-stage evaluation approach for C-XAI method. Finally, **Chapter 10** summarises the research thesis results and discusses future work.

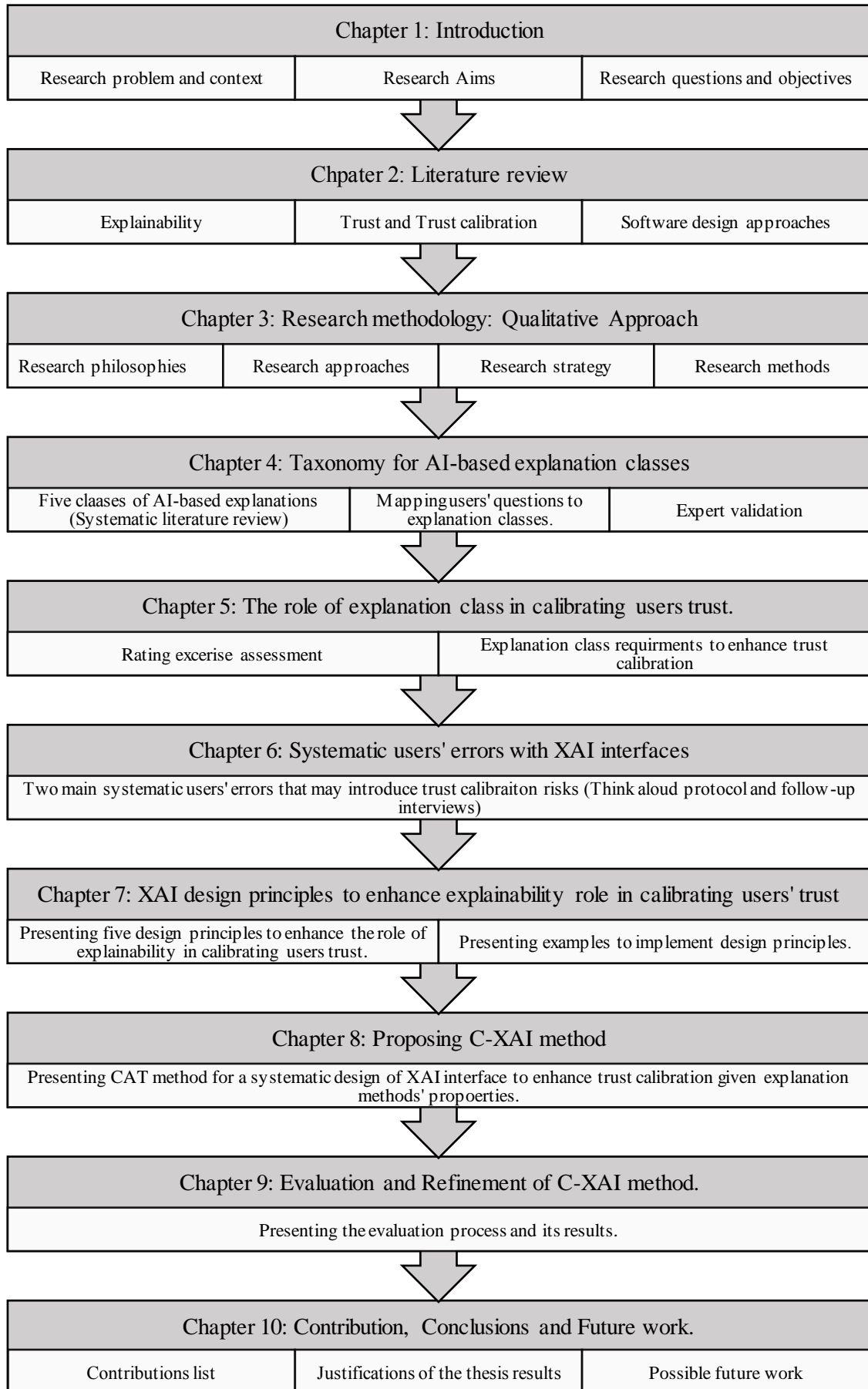


FIGURE 1 THESIS CHAPTERS AND RESEARCH ROAD MAP

Journals:

- Naiseh, M., Al-Thani, D., Jiang, N., and Ali, R., Explainable Recommendations and Calibrated Trust: Two systematic users' errors. *Special Issue on Explainable AI and Machine Learning. In Computer Journal*. **[Published]**
- Naiseh, M., Al-Thani, D., Jiang, N., and Ali, R., Explainable Recommendation: When design meets trust calibration. *Special Issue on Explainability in the Web. In World Wide Web Journal*. **[Published]**
- Naiseh, M., Jiang, N., and Ali, R., How different explanations impact trust calibration: The case of clinical decision support systems. **[Conditionally accepted]**

Conferences:

- Naiseh, M., Jiang, N., Ma, J. and Ali, R., 2020, September. Explainable recommendations in intelligent systems: delivery methods, modalities, and risks. In *International Conference on Research Challenges in Information Science* (pp. 212-228). Springer, Cham. **[Published]**
- Naiseh, M., Jiang, N., Ma, J. and Ali, R., 2020, April. Personalising Explainable Recommendations: Literature and Conceptualisation. In *World Conference on Information Systems and Technologies* (pp. 518-533). Springer, Cham. **[Published]**
- Naiseh, M., 2020, September. Explainability Design Patterns in Clinical Decision Support Systems. In *International Conference on Research Challenges in Information Science* (pp. 613-620). Springer, Cham. **[Published]**
- Naiseh, M., Almansori, R., Althani, D., Jiang, N., and Ali, R., 2021, July. Nudging through Friction: an Approach for Calibrating Trust in Explainable AI. In *2021 8th International Conference on Behavioural and Social Computing (BESC)*. IEEE. **[Published]**

The author contributed as a co-author in related research:

- Cemiloglu, D., Naiseh, M., and Ali, R., 2021, The fine line between persuading and addicting. In *The 16th International Conference on Persuasive Technologies*. **[Published]**
- Aldhayan, M., Naiseh, M., McAlaney, J. and Ali, R., 2020, September. Online Peer Support Groups for Behavior Change: Moderation Requirements. In *International Conference on Research Challenges in Information Science* (pp. 157-173). Springer, Cham. **[Published]**
- Almansori, R., Naiseh, M., Althani, D., and Ali, R., 2021, July. Digital Wellbeing for all: expanding inclusivity to embrace diversity in socio-emotional status. In *British Human-Computer Interaction Conference*. **[Published]**

1.6.1 DECLARATION OF CO-AUTHORS CONTRIBUTION

The author of this thesis was the first author of the publications that originated from the thesis.

The contribution of the first author was as follows:

- Establishing and elaborating the concepts and the research aim of each paper.
- Determining appropriate research methodology to be used in each paper (e.g., multi-stage qualitative approach).
- Designing and conducting the empirical studies presented in each paper (e.g., designing the study, recruiting study participants and collecting and transcribing the data).

- Analysing and interpreting the collected data.
- Writing up the study findings and completely drafting each paper.

The co-authors contributed, verified, and validated the conducted studies in this thesis and reviewed each published and submitted paper. They also gave guidance and feedback on the structure and overall articulation of the papers' argument and goals. They also added insights into the research methodologies and checked the quality of the papers in terms of the writing style and content. Finally, the co-authors improved the papers with the appropriate tools and technologies.

1.7 CHAPTER SUMMARY

This chapter gave an introduction to the structure of the thesis. Also, presented the research goal and research questions, research objectives, methodology summary and publications that emerged from the thesis. The next chapter will provide a literature review on relevant research on XAI, Human-AI trust and trust calibration.

2. CHAPTER 2: LITERATURE REVIEW

The research has argued that Human-AI collaborative decision-making tasks are similar to Human-Human decision-making tasks (Kessler et al., 2017). Madhavan and Wiegmann (2007) developed a framework that describes the similarities between trust development in human-human and human-automation tasks. People tend to apply norms of human-human interpersonal interaction to human-AI interaction (Madhavan and Wiegmann, 2007). However, in human-AI, the human decision-maker role changes from being the primary decision-maker to an active teammate and shares the control of the task with an AI. As a result, the human-AI task's success is always influenced by the humans' trust towards the AI team member. Therefore, calibrating users' trust is a crucial design requirement due to the dynamic and uncertain nature of AI (Cheshire, 2011).

Various factors can influence calibrated trust during the human-AI collaborative task: trust disposition, risk, task complexity, attitude towards AI, and self-confidence (Parasuraman and Riley, 1997). Two critical factors define and attain calibrated trust: transparency and explainability (Lee and See, 2004). This information guides the expectations regarding when to trust or distrust an AI-based recommendation. Enhancing calibrated trust lies behind operationalising explainability and transparency during the Human-AI decision-making task, i.e., helping and supporting users benefiting from explanations to develop accurate mental model about the AI.

This chapter aims to provide an outline of the research conducted in the area of explainability and calibrated trust design. The chapter will provide a grounding for this research and identify clear boundaries for this thesis and other researchers who share the need to design explanations for a calibrated trust goal.

2.1 EXPLAINABLE ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) has been adopted in many everyday applications (Russell and Norvig, 2002). While the origins of AI trace back to decades ago, there is a clear consensus on the need for such machine intelligence to improve many activities of humans' lives, e.g., supporting medical practitioners in their decision-making. It has been shown that the capabilities of AI-based applications are achieving a high level of performance in complex tasks, making them a core element for future industrial development (West, 2018). However, as AI-based applications are deployed into humans' lives and affecting them (e.g., medicine, law or defence), there is an emerging need both from an ethical and legal perspective to make them explainable and transparent (Goodman and Flaxman, 2017). In the following section, a discussion about different aspects of explainable AI and the scope of this thesis are discussed.

2.1.1 THE NEED FOR EXPLAINABILITY

While the early applications and methods of AI-based solutions were easily interpretable and understandable by humans, the last years have witnessed a rise of complex and opaque machine learning algorithms such as Deep Learning (DL) (West, 2018). The high performance and the ability to process millions of parameters of such algorithms were the main reasons for their increasing adoption (West, 2018). Such features of complex machine learning algorithms are also considered as *black-box*. Therefore, explanations about decisions or the recommendations provided by such complex AI-based applications are crucial for human-AI collaborative decision-making tasks. For instance, a medical practitioner working with an AI-based application requires information from the AI-based application to make the final decision affecting their patients' lives (Payrovnaziri et al., 2020). Other examples can also be seen in autonomous vehicles in transportation, finance and security.

In general, explainability considerations in an AI-based application is essential in the development for three main reasons (Arrieta et al., 2020):

- Explainability helps ensure the correctness of the AI decisions and recommendations, i.e., detecting and avoiding any bias inherited in the training dataset.
- Explainability can facilitate the AI-based application's trust by pointing out potential errors that can damage the collaborative decision-making between humans and AI.
- Explainability can ensure that the AI-based application is compatible with current ethical principles.

2.1.2 TERMINOLOGY IDENTIFICATION

A common issue that hinders the establishment of common ground in the explainable artificial intelligence literature is the frequent misuse of interpretability and explainability terms (Abdul et al., 2018, Arrieta et al., 2020). There is a slight difference between the terms. Interpretability describes the silent features of an ML model that can make it understandable by the human operator. This term is also described as transparency. On the other hand, explainability refers to an active feature of an ML model that can generate an explanation for its underlying logic.

The next section describes the most common terms in the literature of explainable artificial intelligence to identify their similarities and differences.

- Understandability, intelligibility or transparency. These terms refer to the features of an ML model to make it understandable to the human operator without explaining its reasoning. (Montavon et al., 2018).
- Comprehensibility. It refers to the ML model ability to represent its underlying logic in a human-understandable way (Fernandez et al., 2019).

- Interpretability. It measures the ability of the model to explain or provide a justification in a human-understandable way.
- Explainability. It is linked to the term explanation, as the interface between the AI-based system and the human operator. This thesis uses this term for the rest of the thesis.

2.1.3 EXPLAINABILITY GOALS

XAI research has identified several goals for explaining the AI-based system recommendations to end-users. This subsection highlights the most frequent goals of explainability provided in the recent XAI surveys (Arrieta et al., 2020, Carvalho et al., 2019) and the thesis systematic literature review in Chapter 4.

- Trustworthiness. It refers to users' confidence an AI-based system will act as intended when facing a problem. Although every trustworthiness AI-based system shall be explainable, explainability does not guarantee trustworthiness. Trust might be a more complex goal to achieve through explainability (Arrieta et al., 2020).
- Persuasiveness. Several researchers identified explainability as a persuasive feature to persuade users to follow AI-based recommendations (Gkika and Lekakos, 2014, Wang et al., 2014). This goal is widely used in commercial recommender systems applications to convince users to buy items or watch videos.
- Privacy awareness. An AI-based system might have learned complex patterns from the data. The inability to inspect these patterns in a human-understandable way might be considered as a privacy breach for an AI-based system (Castelvecchi, 2016, Eslami et al., 2018). Scholars used explainability in AI-based systems to enhance the awareness of the privacy of the model. This also includes explaining the most important privacy patterns only to authorised entities in the system.
- Fairness. From a social perspective, explainability can be considered as a way to convey the fairness of the AI to end-users (Chouldechova, 2017, Benmetot et al., 2019). The impact of the algorithmic decision on humans' lives has increased recently, thus, explainability shall be utilised to avoid and detect unfair or unethical situations. A recent study examined the relationship between explanation perceived fairness and the type of explanation provided to users (Dodge et al., 2019). They concluded that certain explanation types can be perceived as less fair by humans.
- Curiosity. From a psychology perspective, people seek an explanation while they are interacting with their environment. As AI-based systems are implemented in everyday human activities, they shall be explainable to satisfy people's desire to know (Miller, 2019).

- Interactivity. Some scholars included the ability of an AI-based system to be interactive with the human operator as a goal of communicating explainability (Arya et al., 2019, Naiseh et al., 2020a). This goal is related to the applications domain where end-users are working together with an AI-based system to solve a problem. The role of the explanation in such environments is to ensure the success between the human operator and the AI-based system.
- Calibrated trust. Calibrating user trust might be a different goal from the one focusing on increasing the trust of users in AI-based applications (Zhang et al., 2020b, Bussone et al., 2015). Inspiring trust can be done without necessarily improving people mental models and communicating explanations. However, calibrated trust requires explainability as a crucial component in the system to enhance users' mental model of true AI capability and limitations. This thesis studies explainability as a way of calibrating users' trust.

2.1.4 EXPLAINABILITY APPROACHES

The explainability approach refers to the mechanism in which the explanation is generated from the AI-based system. Multiple prior surveys in the literature have tried to generate different taxonomy of explainability approaches (Hall, 2019, Ras et al., 2018, Guidotti et al., 2018b). The following section provides classification to explainable methods based on different criteria. Also, section 2.1.4.2 discusses a trade-off between choosing different explainable approaches and discusses this thesis's focus.

2.1.4.1 INTRINSIC AND POST-HOC

This classification criterion is followed to differentiate two types of machine learning models that are explainable (intrinsic) or black-box models that can be analysed after training (post-hoc). Considering the intrinsic models, explainability can be achieved through the imposition of constraints on the model, such as sparsity, causality, monotonicity, or physical constraints that are related to the nature of the machine learning model and the domain knowledge (Carvalho et al., 2019). Post-hoc explainability refers to explainable models that are applied after the model training. These models usually use reverse engineering methods on machine learning models to explore how a specific output can be generated given a single input.

2.1.4.2 MODEL-SPECIFIC AND MODEL-AGNOSTIC

Another classification for explainability approaches in the literature is *models-specific* and *model-agnostic*. Model-specific explainable models are limited to a particular machine learning model as it interprets the output of the machine learning model based on its internal features (Ras et al., 2018). For example, interpreting the weights in linear models are model-specific

interpretation as the interpretation of intrinsically transparent machine learning models are always model-specific.

On the other hand, model-agnostic explainable models are methods that can be applied to any machine learning model. These models are applied after the training phase of the machine learning model. In general, these models cannot have access to the internal mechanisms of the underlying machine learning model, such as the structure of the deep neural network structure (Arrieta et al., 2020). These models are an extra layer added to the machine learning model to interpret its predictions. These models can also interpret the machine learning model prediction without affecting its accuracy, as they are implemented after the training phase. This thesis focuses on this type of model as there is no clear evidence in the literature on how to utilise these models in real-world scenarios to calibrate users' trust (Zhang et al., 2020b). Furthermore, these models are usually utilised by machine learning experts and data scientists to debug the model. Their implementation to real-world application scenarios to calibrate users' trust is still missing (Miller et al., 2017). Finally, model-agnostic explanation has gained attention recently to be utilised due to its ability to preserve the confidentiality of the ML model (Arrieta et al., 2020, Adadi and Berrada, 2018). **Chapter 4** provides a systematic literature review to classify these models.

2.1.5 EXPLAINABILITY APPROACHES PROPERTIES

With the current advances in explainable Artificial Intelligence (XAI) research, it has grown a challenge to maintain a record of analysing and comparing many explanation methods (Weld and Bansal, 2019). The XAI research community has sought to define properties that explainable methods should be evaluated against clearly. The need for such a framework is inspired by the fact the some of the explanation methods properties might be lost during the implementation (Sokol and Flach, 2020a). For instance, an explanation method is particularly designed for medical data and cannot be used in other domains. As a result, many implementations' properties may not be exploited or cause a faulty on the XAI interface level. For example, an XAI interface method based on counterfactuals might not take advantage of its interactive property.

This demand for a systematic framework that assesses an explanation method has encouraged the researchers to provide a desired list of properties. Miller (2019) presented a list of desired features for explanation models using social sciences insights. Weld and Bansal (2019) discussed a subset of explainable methods requirements supported with examples to help the users of an explanation method familiarise themselves with the explanation output. On the other hand, organisations such as IEEE provided standards for the transparency of autonomous systems (Association, 2018). Other studies chose a subset of properties and used them to evaluate the explanation method in a particular task or application domain. For example,

Kulesza et al. (2013) evaluate music recommender systems' explanations in terms of explanation fidelity. In another study, Kulesza et al. (2015) evaluated interactive email classification applications concerning fidelity, interactivity, parsimony and actionability.

Recent research by Sokol and Flach (2020) developed a coherent framework to mitigate the lack of consensus regarding a common set of properties that each explanation method should be evaluated against. The framework has five main dimensions:

1. **Functional.** This dimension can help to evaluate whether a particular explanation model is suitable for the desired task.
2. **Operational.** This dimension can support designers in understanding how users would interact with an explanation.
3. **Usability.** It helps to evaluate the explanation model from the users' point of view. Usability dimension includes properties of an explanation method related to theories of explainability from social sciences.
4. **Safety.** Explanation models communicate partial information about the data set used to train the AI-based model. The safety dimension evaluates the effect of the explainable method on the security and privacy of the AI-based model.
5. **Validation.** It reveals methods of evaluation have been done in the literature to evaluate the given explanation method.

Although the research sought to propose frameworks for explanation methods properties, these properties are rarely discussed in terms of designing XAI interfaces, and how different values for each property could affect designers' decisions. Such a gap in the literature could be potential explainability limitations to support trust calibration (Zhang et al., 2020b, Bussone et al., 2015).

2.1.6 HOW TO MEASURE AND EVALUATE EXPLAINABILITY

The various benefits of explainability discussed in the previous section have introduced a raised development of various explainable models developed in the context of a particular application. Most of the explainable models depend on the underlying machine learning model, the type of data used in the underlying model, the type of information provided in the explanation, or the explanation's scope (Local or Global). This increase of explainable models has led to a unified definition and measurements of explainability. This also introduced a lack of a common framework to evaluate the explainable models and compare their performance.

A recent survey (Carvalho et al., 2019) in interpretable machine learning algorithms have identified the following evaluation metrics which explainable model should optimise:

1. **High Fidelity.** It represents the degree to which the explainable model matches the black-box model i.e., how truthfully the generated explanation represents the underlying black-box model. This metric has appeared in (Lakkaraju et al., 2017, Ribeiro et al., 2016a). Fidelity is an essential metric for optimising the model-agnostic model because low fidelity explanations are argued to be misleading and confusing for users e.g., an explanation model might use completely different features to the actual features used in the underlying logic of the black-box model.
2. **High accuracy.** It measures the accuracy of the generated explanation for unseen data, i.e., an explainable model might perform well on the training data and has low accuracy for unseen data. Low accuracy explanation might leave out the end-users with information that makes no sense. Thus, explanations might be wrong and lead to wrong representation and low trust in the black-box model even when the black-box model has high prediction accuracy. This term appeared in (Ribeiro et al., 2016a, Lakkaraju et al., 2017).
3. **Generalisability.** This metric represents to what extent the explanation can be generalised across the data. The literature has identified three main scopes for generalisability: *local* - for single prediction, *group* – for a cluster of data points and *global* – for the entire underlying model.
4. **High Explanatory power.** It is associated with the number of users' questions can the generated explanation answer. For example, global decision tree approximation methods might answer users' questions such as why, why not, and what if for local predictions (Lim et al., 2009).
5. **High Stability.** It refers to the level of similarity between two explanations generated for similar AI-based recommendations. An explainable model's high stability means that a little change of the input data to the explainable model shall not introduce a high change in the generated explanation unless this change has a huge impact on the AI-based decision (Carvalho et al., 2019).
6. **High Consistency.** While stability focuses on two explanations samples from the same explainable model, consistency related measures the difference of two explanations from two different explainable models regarding the same AI-based recommendation. If the explanations are similar, the explanation is highly consistent.

7. **Comprehensibility.** This evaluation metric measures the understandability of the explanation from the humans' perspective. This metric could be a subjective metric, where it depends on the audience of the explanation and the context.

2.1.7 HUMAN-COMPUTER INTERACTION ASPECTS

Unlink the different AI fields that rarely involve the end-users in the development and evaluation phases. The XAI field has given more attention to end-users. This section presents an overview of the relevant literature and factors considered vital to Human-Computer interaction literature.

2.1.7.1 HUMAN REASONING AND EXPLANATIONS

The deviation from expected behaviour triggers the demand for explanations, such as curiosity, inconsistent, or abnormal events (Miller, 2019). With the various triggers for explanations, people would then ask to find a cause or generalise their knowledge based on the perceived explanations. Pierce (1997) identified three main types of inferences: deduction, induction, and abduction. Deductive reasoning is a top-down logic where the process of human reasoning from premises to a conclusion. Inductive reasoning is bottom-up and reverses the process of reasoning from a single observation to a probable explanation. Abductive reasoning is also a reverse of deductive reasoning and reasons from observation to the most likely explanation. Abductive reasoning is also recognised as “inference to the best explanation”. It is more selective than inductive reasoning since it prioritised hypotheses (Peirce, 1997).

Researchers sought to develop frameworks that justify the strengths of XAI literature in supporting human reasoning. For instance, Theory-driven user-centric explainable AI is a framework that highlights the main features of XAI and their ability to support human reasoning processes (Wang et al., 2019). Such mapping between explainability features and human reasoning processes helps mitigate human biases while interacting with AI-based explanations. For instance, What-if explanations can support counterfactual reasoning; explanations about the prior probability can mitigate confirmation bias. The framework is presented in Figure 2.

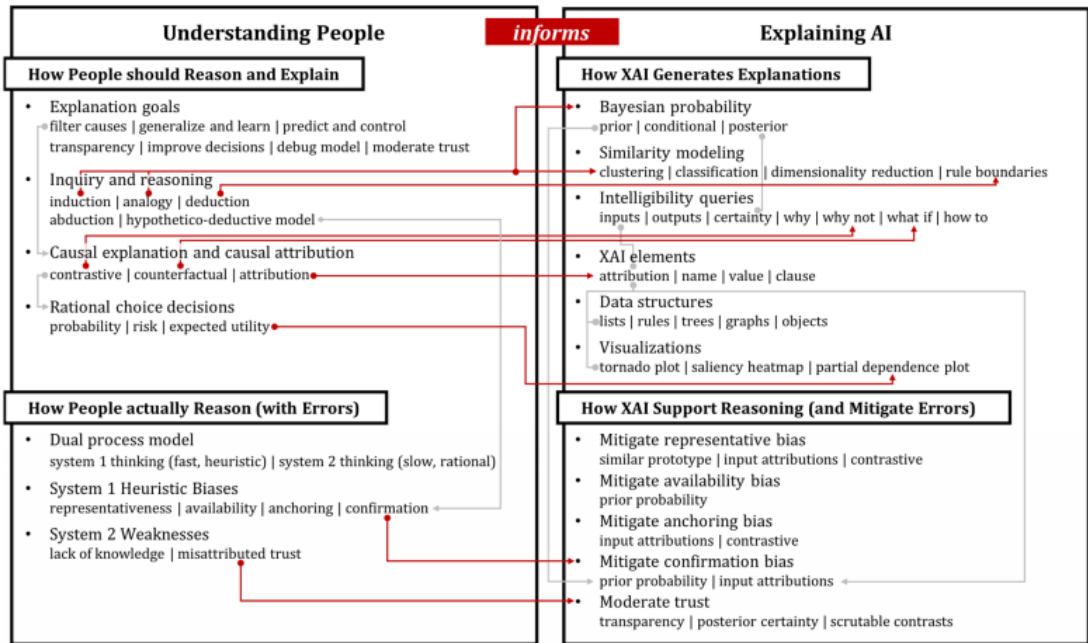


FIGURE 2 THEORY-DRIVEN USER-CENTRIC EXPLAINABLE AI FRAMEWORK INTRODUCED BY WANG ET AL. (2019).

Scholars have used this framework to theoretically support their decisions of implementing an explanation over another explanation (Lim et al., 2019); this includes connecting the design choice to the users' goals in the system. For example, a user trying to recognise a specific reason for an AI recommendation may follow contrastive reasoning, which can be supported by using "why not explanations". Such explanations provide information about why another AI decision has not been recommended. However, the framework lacks consideration of the environment in which the recommendation is being issued. Moreover, the framework links humans' reasoning processes to current explainable models and techniques, which limits the guidance of emerging explainable models in the literature.

2.1.7.2 EXPLANATION PRESENTATION

It refers to the adopted method to convey and present explanations. The most common approach to convey explanation to end-users is using textual presentation (Nunes and Jannach, 2017). Also, different styles of visualisations, e.g., in form of graphs, are common approaches to convey the explanations to users (Eiband et al., 2018a). Previous work compared textual and visualisation approaches to explain AI-based recommendations and suggest that standard rules for interface design are the best guide for explanation presentation choice (Gregor and Benbasat, 1999). Zanker and Schoberegger (2014) claim that text-based explanations should make use of conversational language. Naveed et al. (2014) recommend using argumentation-based language to explain recommendations. Herlocker et al. (2000) explored various explanation presentation styles and formats; their results showed that simple graphs were preferable. Furthermore, recent research showed that preferences towards presentation methods could differ between users based on their goals, level of knowledge and familiarity with the method. Feng et al. (2019)

indicated the importance of personalising explanations to end-user by studying the effect of the presentation method on expert and novice users. Results from their study produced a more accurate and realistic evaluation for machine learning explanations methods. Kouki et al. (2019) used the explanation presentation method as a control variable to find which visualisation method is more persuasive for end-users.

Researchers agree that presenting explanations to end-users shall be done to reduce users' cognitive overload (Kulesza et al., 2013, Sokol and Flach, 2020b, Naiseh et al., 2020c). In conclusion, the literature had provided limited guidelines and no best practices concerning the explanation presentation format. The guidelines focus on an abstract level and rely on the designers to interpret and choose the best explanation presentation in a particular application. To date, there is no consensus as to when to use text-based format or visualisation-based format, and XAI designers have to explore the best way to visualise explanation (Eiband et al., 2018a).

2.1.7.3 EXPLANATION STYLE

This section describes various styles in explaining the recommendations, highlighting different possible motivations and effects. Explanation style indicates the orientation, level, granularity and framing adopted to explain the recommendation (see Table 2). Those may accompany a particular algorithm or be used for a specific goal. For example, Zanker et al. (2014) found that fact-based explanations make users change their preferences about the recommended item and can be used for persuasiveness goals. Explanation styles can contain information ranging from high-level descriptions of how the algorithm works to low-level explanations of the specific factors influencing the algorithm decisions. Furthermore, the design can combine various styles in one explanation (Sato et al., 2018, Kaptein et al., 2017). Sato et al. (2018) combined context-based explanation with other styles e.g., demographic-based explanation, and confirmed that this combination of styles improved the persuasiveness and usefulness of explanation.

Explanation style is one factor that influences the users' decisions to follow the recommendations (Kouki et al., 2019). Also, the design of the explanation is typically based on system goals and motivations. A typical goal for recommender systems is to introduce users to new solutions that match their interests, needs and intentions (Tintarev and Masthoff, 2011), but other goals differ based on the application domain and this affect the suitability of the explanation style. For instance, Chen (2009) used the Argumentation style to explain the recommendations that have a multi-objective decision process, e.g., Products with different quality attributes, so that the explanation can show the trade-off between the objectives (e.g., price vs. quality and comfort vs. durability). Sharma et al. (2013) relate the effect of social-based style to persuasiveness goal which increases the users' listening rate to recommended songs. Our analysis revealed other explanation styles typically found in group recommender systems to reach an acceptable consensus within the group of users. For example, Najafian et al.

(2018) present two explanation styles to demonstrate the agreement and disagreement on a specific item of the group members.

The personalisation of explanation style was rarely studied in the literature despite being a demanded requirement due to the different effects, positive and negative, of each style on end-users (Sato et al., 2018, Naveed et al., 2018). Situational and personal variable, such as gender (Kleinerman et al., 2018) and age (Kaptein et al., 2017) have been studied to choose an explanation style that best fits the context, end-user characteristics and end-user goals. Yet, most of the provided methods and approaches focused on the improvement of the system goal by choosing an explanation style that fits it and the explanation content. Little has been done to personalise the explanation style and the level of the explanation complexity based on users' characteristics such as their cognitive style of the consumer, personality characteristics and the way users interact with the explanations (ask for alternative explanation style, configure their style, hybrid methods of interaction). Personalising the explanation styles could provide a basis for design choices at the interface level, as well as improve the perceived quality and trust of recommender systems (Kulesza et al., 2013). However, considering this personalisation in the design of the explanation interface could be challenging where more data should be collected from the end-user, this will require developing approaches to detect the preferred explanation style for the end-users. A summary of various explanation styles from the literature is presented in Table 2.

TABLE 2 EXPLANATION STYLES EXAMPLES

Explanation Style	Description	Reference
Fact-based	It is based on a formal argumentation style. E.g., with A and B as premises and C as a consequent, the explanation would be, from A and B, therefore, C	(Zanker and Schoberegger, 2014, Gasparic et al., 2017)
Goal-based	It answers questions like "to what end?" or "for what purpose ", e.g., Recommended diet plan is trying to balance between diet and activity.	(Braunhofer et al., 2014, Kaptein et al., 2017)
Context-based	It is based on presenting a context and usage scenario to demonstrate the recommendation rationale, e.g., a waterproof jacket would fit UK weather	(Sato et al., 2018, Wiebe et al., 2016)
Argumentation-based	It generates a set of pros and cons about a specific recommendation to help make informed decisions, e.g., Speed vs durability.	(Muhammad et al., 2015, Naveed et al., 2018, Chen, 2009, Jugovac et al., 2018)

Influenced-based	It relates to explaining to the recommendation impact deemed by the systems to be important, e.g., this is be-because it has a V-neck collar.	(Berkovsky et al., 2017, Chen et al., 2019)
Belief-based	It presents the context and reasons that influence the algorithm to choose this recommendation over another, e.g., the highlighted route has less traffic jam right now	(Kouki et al., 2019, Kulesza et al., 2013)
Demographic-based	Explanation points out the relationship between the demographic data (age, gender, location, etc.) and the recommendation, e.g. young males prefer racer cars.	(Tintarev and Masthoff, 2011)

2.1.7.4 DELIVERY METHODS

Delivery methods refer to the way that explanations are communicated to end-users (Naiseh et al., 2020b). Delivery methods of explanations focus on how the delivery methods inter-relate with other design considerations. Delivery methods are not mutually exclusive, and multiple delivery options can be used in the same interface based on the context, the recipient of the explanation and the nature of the application. The results from the literature review revealed four delivery methods.

Persistent-specific: Explanations are delivered to the users along with the recommendation in a straightforward and accessible way and without waiting for the user to request the explanation. The lifetime of the explanation in this method is specific to the user interaction time with the recommendation. In other words, the user is unable to consume the explanation after finishing the task. The main goal is to inform the user decides whether to accept the recommendation. This method is used in the literature to foster trust (Gedikli et al., 2014), transparency (Gedikli et al., 2014), persuasiveness (Sato et al., 2013), user acceptance (Karga and Satratzemi, 2019) and prevent errors and bias (Schäfer et al., 2019). The cost-benefit analysis is a challenging design consideration (Kulesza et al., 2013) as users may perceive the cost of reading explanations to exceed their benefits (Bunt et al., 2012).

Ad-hoc: The explanation in this category is designed to be delivered to the end-users when it is necessary and needed. This method is used in the literature in two ways:

On-demand: This method enables the users to request the explanation where the explanation is embedded in a separate view, and the users can ask for it. This is meant to reduce information overload in the interface (Millecamp et al., 2019, Barria-Pineda et al., 2019) when explanations are not always beneficial or crucial for the performing task (Bunt et al., 2012, Wiebe et al., 2016). Also, this delivery method could blend well with the persistent-specific method, e.g., when users ask for further details to reveal the full set of explanation features (Muhammad et al., 2015). On-demand method is useful where explanations contain a high level

of information so users may get distracted and need more time to consume it (Gedikli et al., 2014, Springer and Whittaker, 2019). Also, it is argued to be more effective to reduce users' cognitive effort and avoiding overwhelming end-users with unnecessary information (Poursabzi-Sangdeh et al., 2018). On the other hand, embedding explanations in a separate view argued in the literature might not fulfil the goal of presenting the explanation and become an additional burden on user experience. Eslami et al. (2018) and Leon et al. (2012) found that users might not benefit from this method as end-users may hardly notice the on-demand button due to factors like their main focus and flow state.

Exploration: The users in this method can explore the nature of the explanation and the agent process and increase the understanding of the reasoning behind the recommendation (Bostandjiev et al., 2012, Verbert et al., 2013). This exploration could be (a) feature-based exploration, where user can investigate how individual feature contributes to the recommendation and explanation output (Lamche et al., 2014), (b) subset-based where in-put features specified by users are leveraged (Kulesza et al., 2012) and (c) global exploration where the nature of the data and its distribution are exploited (Schaffer et al., 2015). Exploration techniques help users to build useful mental models and provide the user with the ability to discover more knowledge and about the agent interactively and engagingly (Kulesza et al., 2015). Examples of such tools help the users in some problems like detecting bias in data (Krause et al., 2018), combat the filter-bubble effect in social media (Kang et al., 2016).

Persistent-generic: Explanations are stored as a report for later investigation, and the explanation is persistent without a time limit. The report may include more information compared to persistent-specific and ad-hoc methods. For instance, information about the underlying processes of the algorithm decision-making on each step of the process and the reasons for selecting each decision point (Langley et al., 2017). This is essential in some application domains, such as clinical decision support systems where the explanation is a crucial factor for accountability, traceability and ethics (Binns et al., 2018, Dodge et al., 2019). Most of our reviewed studies did not focus on developing approaches with the ability to access the explanation after finishing the task. The main approaches provided in the literature to apply this method include i) embedding the explanation in the "help page" (Eiband et al., 2018b) and ii) providing a dialogue interface to navigate the archive interactively (Zhao et al., 2019).

Autonomous: This method appeared twice in our reviewed studies. The system in this method is responsible for deriving users' needs for an explanation based on the context. In other words, it is about the autonomy of the system to choose the time and the context to deliver the explanation. In contrary to the ad-hoc approach, which is a user-based delivery method, autonomous approaches are a systems-based method. Lim et al. (2009) argued that this method could be used to provide privacy-sensitive information when the recommendation could provoke privacy violation so that it acts as a precautionary measure. Understanding the nature of

the application and the different users' personas is essential to launch this approach in human-agent systems. The papers that studied this method appeared in the domains of ubiquitous computing (Lim and Dey, 2009) and robots (Huang et al., 2018). For instance, Huang et al. (2018) develop an approach to explain the intelligent agent behaviour only in critical situations, e.g., there is no need to explain why the autonomous vehicles slow down when the road is empty. This method was helpful to calibrate user trust and avoid over-trust and under-trust states.

2.1.7.5 MODALITIES

Explanations in common applications are presented either as text or graphical representations in a static way (Nunes and Jannach, 2017). However, explanations can be designed as interactive systems where the initial explanation represents a starting point for further user interactions, e.g., asking the user for correct parts of the explanation. Designers use such modalities to streamline user functionalities to explore more details about the underlying algorithm and put the user into control of the output. Providing such interactive explainable interfaces can fulfil both persuasion and over-trust reduction requirements by demonstrating the algorithmic reasoning in thoroughness and experimental way to the end-users (Schäfer et al., 2019). Research in this area is still limited, and it is unclear how to design interactive explanation interfaces in a way that is tailored and fit to users in standardised or personalised ways. Kulesza et al. (2013) mentioned that supporting users with interactive explanations could lead to more complexity, as it needs a level of knowledge in software engineering and machine learning and also burden on the user experience. This section focuses on discovering common input modalities in the literature that typical explanation not only conveys information but also might trigger an interactive approach. We highlight these types and their potential usage scenarios.

Control. Users are enabled to play, change, regenerate or elicit some preferences about the agent to enhance their understanding of the underlying system (Le Bras et al., 2018). The main principles behind this interaction style include boosting transparency and interpretability of the system processes and giving users control over their output (Lamche et al., 2014). Studies found that the control functionalities can enhance the user experience as well as enrich mental models. The research in this area focused on providing dynamic explanations more than static approaches when users need to observe the inter-relation between different factors that influence the output - e.g. Recent research presented an approach based on a user-controlled function that includes different explanation components, which allows tuning recommendation parameters for exploring social contacts at academic conferences (Tsai and Brusilovsky, 2017).

Configure. This modality gives end-users the ability to choose what information, presentation, colours, order and size are suitable to reflect the importance, relevance and focus of certain parts

of the explanation. This method is rarely studied in the literature as it appeared twice in our selected studies and without elaborating on the design considerations (Díaz-Agudo et al., 2018).

Dialogue. It indicates the explanations provided to the end-users in an inter-active bi-directional style. The user can ask for specific information about the recommendations (Zhao et al., 2019). This approach is argued to be beneficial for the design of explanation interfaces and balance between the amount of information presented to the end-users and their cognitive efforts to process that information (Ramachandran et al., 2015). Our findings show that users have specific information requirements before they are willing to use recommendations such as system capability, algorithmic reasoning and detailed information about the recommended item. For instance, Eiband et al. (2019) revealed that users called for more accurate information about the recommended item rather than relying on the item-based and user-based explanations. These requirements could be fulfilled by using a dialogue interaction, as the user will be assured asking specific questions about the recommendation.

Debug. In this approach, the system presents its explanation to the end-users, where, on the other hand, users are enabled to provide corrections and feedback about the explanation to the systems to improve output in the future recommendations (Kulesza et al., 2013, Kulesza et al., 2015). User debugging can occur by different types of inputs, such as providing ratings to the explanation (Elahi et al., 2015) and correcting parts of the explanations explicitly (Kulesza et al., 2015). Providing the debug modality is argued to increase the algorithm accuracy by putting the Human-In-The-Loop.

2.1.7.6 PERSONALISATION

The complexity of intelligent systems and their wide adoption in real-world and daily life applications like healthcare made them more visible and familiar but at the same time more questionable. Driven by theories of the ethical responsibility of technology development, informed consent and informed decision making, regulations started to emphasise and even demand the right of citizens to be explained how systems work and how their data is used (Goodman and Flaxman, 2016, Tomsett et al., 2018). Researchers argued that the current approaches to helping trust calibration design require personalising XAI interface design to fit as one explanation does not fit all users (Rosenfeld and Richardson, 2019). For instance, Tomsett et al. (2018) propose a model that defines six different user roles in the machine learning system and argued that designers of the system should consider providing different explanations that match the needs of each role in the system. Also, Rosenfeld and Richardson (2019) presented three categories of users who are different in their demands of the nature and level of explanations: Regular user, Expert user, and External entity. These results are supported by (Millecamp et al., 2019) who investigated how the end-user characteristics, including their profile and role in the system, affect the design of the explanations and trust calibration. This

this thesis does not study personalisation as a target solution to help XAI interface design, but it looks at the XAI methods systematically to propose a solution that map XAI research to HCI research.

2.2 DEIGNING FOR TRUST CALIBRATION

As presented in the previous section, calibrated trust is a key requirement for the success of AI-based decision-making tools in real-world scenarios for achieving safe and accountable design. This section provides an introduction to the concept of calibrated trust and analyses the literature on how AI-based applications shall be designed to achieve appropriate trust.

2.2.1 CALIBRATED TRUST DEFINITION AND CONCEPTUALISATION

Trust has been defined in many fields such as social sciences and psychology literature. This section outlines the definition of trust from Human-Automation Interaction and it is defined as “the attitude that an agent will help achieve an individual’s goal in a situation characterized by uncertainty and vulnerability (Lee and See, 2004)”. The definition uses the term agent to show that trust can exist only between humans and between any two agents collaborating in an active environment. Recent research showed that similar factors exist in trust between humans and trust between humans and machines (Lee and See, 2004, De Visser et al., 2017b). For instance, uncertainty and vulnerability are important aspects that help humans reduce the cognitive effort in a complex situation and guide humans’ trust (Lee and See, 2004).

The increasing complexity and sophistication of AI-based systems make understanding of the system, and its underlying reasoning is unachievable. This makes calibrating trust design an important mediator in the interaction of humans and AI (Lee and See, 2004, Parasuraman and Manzey, 2010). Previous research (Hergeth et al., 2016) showed an inverse relationship between humans’ trust and the ability to succeed in Human-AI collaborative tasks. They concluded that calibrated trust is the key to avoiding misusing the AI-based application (over trust the AI) or disusing it (under trust the AI). This relationship is shown in Figure 3. These findings are in line with the findings by Parasuraman and Manzey (2010) who showed that over-trust is a key cause of automation bias. Over-trust of users was highlighted by a recent article showing a sleeping driver for Tesla car with an automation feature for driving. Overall, calibrated trust does not guarantee an appropriate reliance on the AI-based system, but it helps to guide it (Chavaillaz et al., 2016). Factors such as workload, self-confidence, and repeated success of an AI affect the decision to rely on the system in addition to trust (Lee and See, 2004).

2.2.1.1 *CONCEPT DEFINITION*

Calibrated trust is a design goal to mitigate undesired behaviour with the AI-based system as well as promote desired behaviour (Lee and See, 2004). Trust is associated with the attribute Calibrated when the human agent can understand and match the AI-based system's actual

capabilities. In a recent extensive literature review considering trust in automation, Lee and See (2004) identified three main aspects to achieve calibrated trust in the Human-Automation relationship:

1. Calibration. It is an appropriate matching between users' trust and the actual capability of a system.
2. Resolution. It refers to the degree to which users' trust judgment differentiates several capability levels of the automation, e.g., high-resolution capability changes require high changes in trust.
3. Specificity. It refers to the degree of differentiation between the different aspects of the trustee: (a) Functional specificity: it refers to the differentiation between functions, and (b) temporal specificity: it is the sensitivity to changes (See Figure 3).

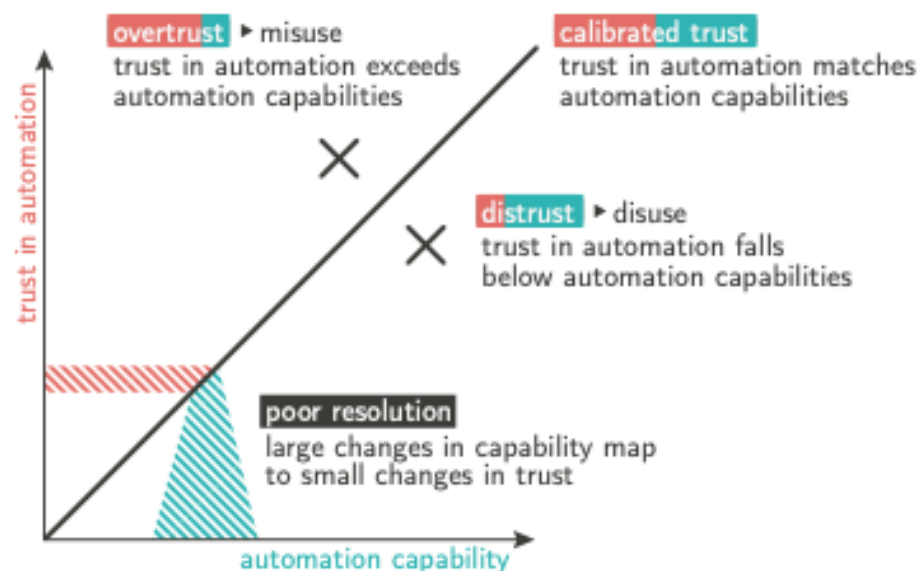


FIGURE 3 DESCRIPTION OF CALIBRATED TRUST AND TRUST RESOLUTION (ADOPTED FROM LEE AND SEE 2004)

Over-trust and under-trust can be alleviated in situations where resolution and specificity are high. During the collaborative decision-making task between Automation and Human, the automation interface shall support the human operator with information to facilitate calibrated trust. Lee and See (2004) identified two dimensions for the human operator's information: (a) Transparency. It refers to information indicating the overall performance of the automation, processes, and automation goal; (b) Explanation. It describes the logic of automation and how the automation decisions are being made. Abstraction and explanation form the basis of calibrated trust in automation.

2.2.1.2 TRUST CALIBRATION DESIGN SPACE

Designing for trust calibration is the active management of users' mental models while using the AI-based decision-making tool (Lee and See, 2004). The user mental model is formed in three steps: understandability, predictability and reliability (Hoffman, 2017). Understandability refers to understanding the AI-based system knowledge and reasoning process. Predictability is the ability to predict future recommendations and decisions for the AI-based system. Reliability is a consistent behaviour of the AI-based system under the same circumstances. Hoffman (2017) defined trust calibration space as the active management of understandability, predictability and reliability.

Several trust calibration frameworks and methods have been proposed in the literature to aid the systematic design of AI-based systems. For instance, Trust Taxonomy Cues methodology (de Visser et al., 2017a) is based on a mapping between trust evidence level (origin, expressiveness, process, performance, and intent) and the information process level (perception, comprehension, projection, decision, and execution). Furthermore, recent research defined the overall interaction of the user from entering the autonomous car until exiting it (Ekman et al., 2017). They identified 13 different interactions in the overall journey and mapped them against 11 factors that can facilitate a calibrated trust (See Figure 5). Another research presented a different view of achieving calibrated trust (Okamura and Yamada, 2018). They presented the use of intuition and logic on one axis and anthropomorphism and machine on the other axis (See Figure 4).

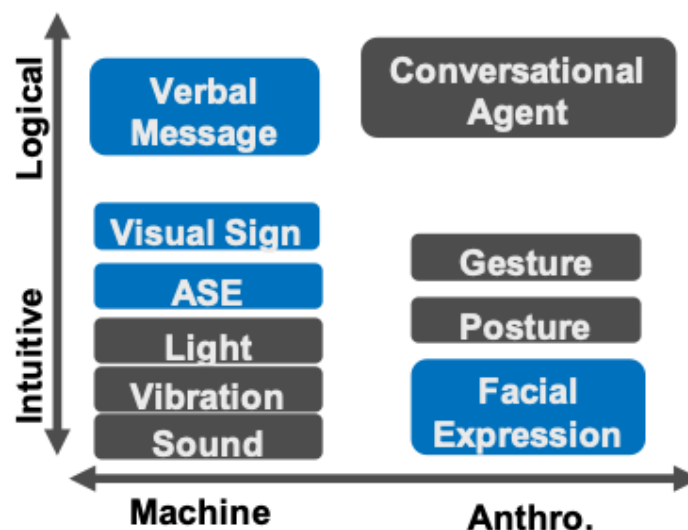


FIGURE 4 OKAMURA AND YAMADA (2018) FRAMEWORK FOR CALIBRATED TRUST

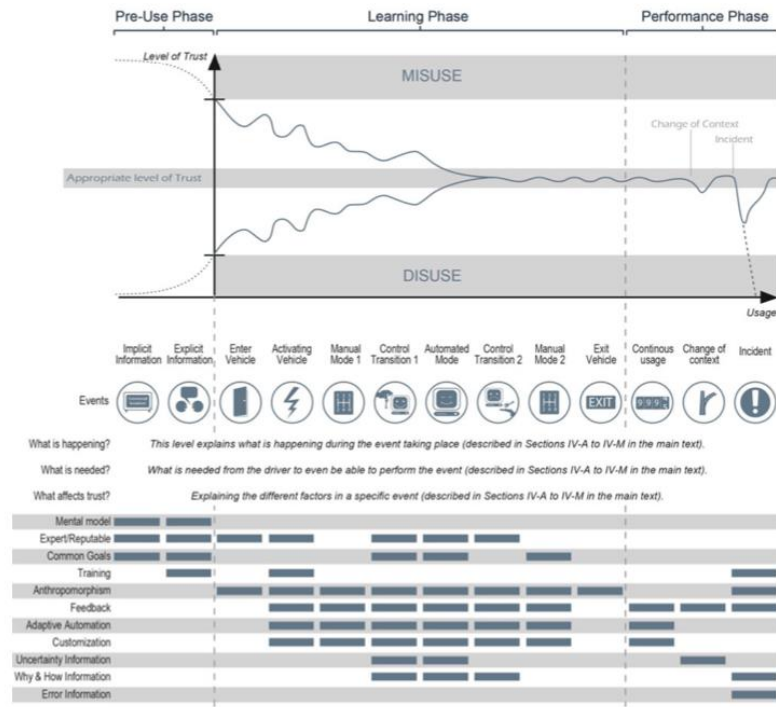


FIGURE 5 EKMANN ET AL. (2017) FOR CALIBRATED TRUST

The mentioned research work presents a primarily work toward designing for trust calibration through communicating information about the AI-based system to the end-user. However, nearly all the approaches hypothesised that users of AI-based systems would benefit from such information and use them to calibrate their trust. Recent research showed that even though the AI-based system provided an explanation about its logic, users failed to calibrate their trust (Bussone et al., 2015, Zhang et al., 2020b).

2.2.2 THE ROLE OF HUMANS' ERROR

Human errors occur when the designed constraints for the system fail (Leveson, 2016). Understanding the nature of the users' errors requires understating their behaviour with the system. Such knowledge would then help to realise the design constraints. It has been identified that humans' behaviour has three different levels: skill, rule, and knowledge-based performance (Rasmussen, 1983). Based on Rasmussen (1983) definition, Skill Based (SB) behaviour is "*sensor–motor performance during acts or activities, take place without conscious control as smooth, automated, and highly integrated patterns of behaviour*". SB behaviour occurs intuitively without the consciousness of the human operator. Rule-Based (RB) behaviour occur consciously, and the human can report the rules used in performing the task. Knowledge-Based (KB) behaviour requires an understanding of the activity to deal with an unfamiliar situation. This feature distinguishes KB from SB and RB, as KB occur when the human faces an unfamiliar situation. In comparison, SB and RB are related to familiar situations to the human. While the high majority of human errors are SB or RB errors as they occur in daily life activities, the probability of KB mistakes is lower compared to SB and RB.

Another classification of users' behaviour with automation was based on trust behaviour (Lee and See, 2004). This classification has identified three main dimensions:

1. Misuse: It refers to using the system in a situation for which it has not been designed. Such behaviour could potentially be a result of over-trust of the system.
2. Disuse. It refers to rejecting engaging with the system even when it is suitable for the task. This behaviour is linked to under-trust the system.
3. Correct use. It refers to utilising and engaging with the system in situations and tasks that fit the system ability. Such behaviour requires the human operator to understand the knowledge and the reasoning of the system.

As discussed in Section 2.2.1, calibrated trust design is a desired human behaviour with the AI-based tool. That means the design of the XAI interface to calibrate users' trust is untimely limited to the nature of the humans' errors and mistakes (Rasmussen, 1983). Thus, understanding the XAI interface's design constraints and the different design features that enable a desired calibrated trust behaviour would increase its efficacy to calibrate users' trust. Also, a deeper understanding of users' behaviour with the AI-based explanations and what triggers misuse and disuse of AI-based explanations is still missing. Such knowledge could be useful for researchers and practitioners to develop calibrated trust XAI interfaces.

2.2.3 THE ROLE OF EXPLAINABILITY

An essential requirement for the success of AI-based decision-making tools in collaborative Human-AI environments is to help users' form a correct mental model of the AI-based tool (Bansal et al., 2019). That means the human operator needs to know when to trust or distrust the AI-based system. Many researchers have shown several challenges for users to form a correct mental model about the AI-based system due to the opaque nature of the underlying black-box models (Chen et al., 2017). Furthermore, researchers have shown that the focus of increasing the performance of the AI-based system does not imply mitigating the risk of trust calibration errors. Human operators still need to understand and form correct mental models about the AI-based system. For instance, the AI-based systems' dynamic nature may confuse the human operator, who may accept or reject AI-based recommendations in the wrong situations. This could happen even when the algorithm's overall performance is high (Bansal et al., 2019).

It has been suggested that the explainability of system reasoning is a crucial factor in supporting correct mental model formation (O'Sullivan et al., 2014). This could help the human operator decide when to accept an AI-based recommendation or reject it. Some of the recent work has focused on showing the reasoning of the AI-based system through several explanation types, such as answering why questions (Gönül et al., 2006) or showing the confidence score (Zhang et al., 2020b). On the other hand, several studies have examined the effectiveness of the

explanation on users' trust calibration process (Bussone et al., 2015, Zhang et al., 2020a, Bansal et al., 2020). A general conclusion is that the current utilisation of explanations in AI-Human collaborative decision-making environments could facilitate trust calibration errors. Besides, the research around developing explainable interfaces for trust calibration argued that explanations should indeed enhance the trust in the system but mainly in terms of its credibility and transparency. However, they should also support users' situational awareness of the environment and other contextual factors (Sanneman and Shah, 2020). Overall, the adoption of AI-based tools increased significantly in recent years, and this raises the question of how to design explanations with trust calibration as a primary goal.

2.2.4 CALIBRATED TRUST MEASUREMENTS

Prior research suggested that calibrated trust is the active management of trust components (Madsen and Gregor, 2000). Calibrated trust can be measured by self-reporting trust components and comparing them to their actual values (Wang et al., 2016). We followed Madsen and Gregor (2000) conceptualisation of human-computer trust into two main components: cognition-based components and affect-based components. Cognition-based components are based on the users' intellectual perceptions of the systems' characteristics, whereas affect-based components are based on the users' emotional responses to the system. While the research showed that both affect-based trust and cognition-based trust are more likely to affect the calibrated trust (Zucker, 1986). The main difference in their effect is that cognition-based trust is crucial for establishing appropriate trust, whereas affect-based trust is developed as the relation continues (Nah and Davis, 2002). Furthermore, previous research showed that in critical decision-making scenarios it is highly likely that cognition-based trust components have a significant effect on trust calibration (McAllister 1995, Lewis et al. 1985).

Furthermore, recent studies showed that self-reporting measures are not reliable indicators for calibrated trust (Schaffer et al., 2019, Kunkel et al., 2019). Therefore, the following recent studies proposed several behavioural measures for calibrated trust (Zhang et al., 2020a, Wang et al., 2016, Poursabzi-Sangdeh et al., 2018). The following presents the main behavioural calibrated trust measure resulting from the literature review:

- Agreement percentage: the percentage of trials in which the participants decided to agree with the AI-based recommendations.
- Compliance percentage: the percentage of trials in which the participants choose to follow the AI-based recommendation. The main difference between Agreement and Compliance measures is that the participants agreed with the AI-based recommendation and automatically made the final decision in the agreement case. In contrast, compliance only considers the case where the participants disagree with the AI-based recommendation, but they intend to comply with the AI-based recommendation.

- Incorrect decisions. This is a team-performance measure, and it is extracted from incorrect decisions made between the human and the AI.
- Correct decisions: It measures the percentage where the collaborative decision-making between the human and the AI has led to a correct decision.

2.3 CHAPTER SUMMARY

In this chapter, a review of different explainability and calibrated trust aspects is provided. Furthermore, the chapter discussed different methods that might guide the development of the design method. The next chapter will explain the research methodology, assumptions, and different options for achieving this thesis's objectives.

3. CHAPTER 3: RESEARCH METHOD

This chapter discusses the methodological approach in this thesis. It also justifies and discusses research methods choices adopted in this thesis. This chapter also looks at the variability of methodological approaches, as it is vital to find all the options for data gathering in any research (Twycross, 2004). This chapter starts by discussing research approaches, research methods for data collection, and then methods for interpreting the collected data.

The research methodology and research methods used in this thesis are structured based on the research onion framework (Thornhill et al., 2009). The research onion framework adopted in this thesis is presented in Figure 6. The figure clarifies the phases of the research process, e.g., the philosophies, method to theory development, methodological choices, methods and techniques that can be sought to accomplish primary research goals. The underlined words in Figure 7 indicate the followed approaches in this thesis — the following sections describe in

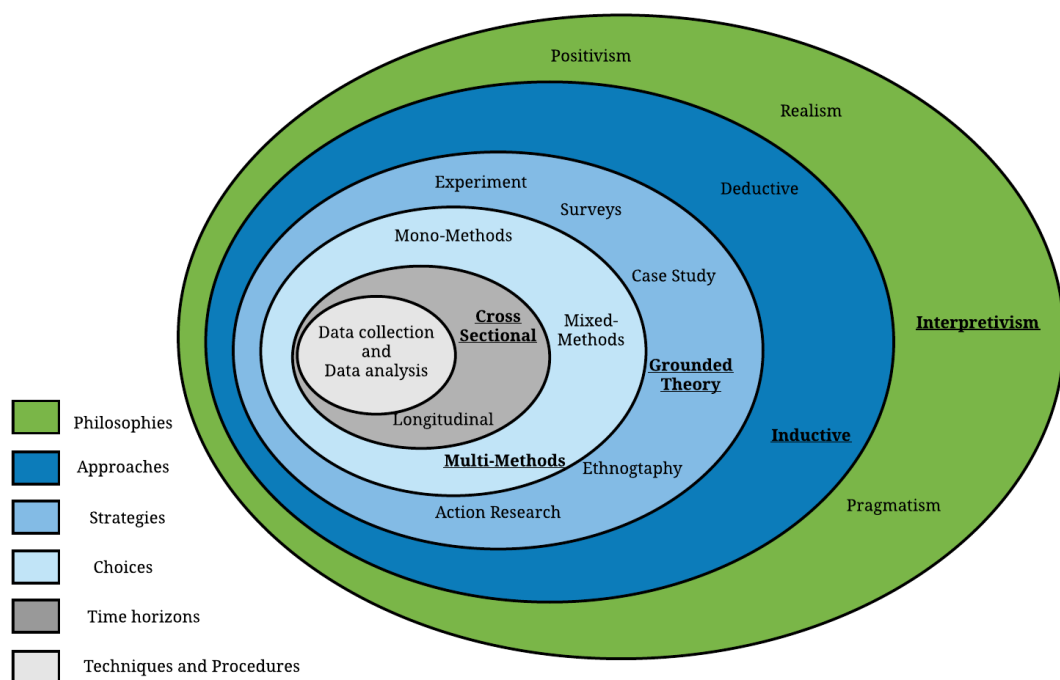


FIGURE 6 RESEARCH ONION (SAUNDERS ET AL.2009)

subsequent these choices.

3.1 RESEARCH PARADIGMS

This section summarises four research paradigms broadly considered in the literature and comprehensively presented in the research onion. Those four paradigms are positivism, realism, interpretivism and pragmatism. Understanding these paradigms will help develop the research beliefs and related assumptions that would guide the research. Such understanding has shown

effective results on the research outcome. It has a substantial impact on the research and the research objective (Gill and Johnson, 2002). The followed research paradigm holds vital assumptions on the reasoning of following research techniques and approaches. Researchers require an understanding of the philosophical aspect because this has a vital impact on what they do, as well as knowing what they are investigating. Research paradigms can be differentiated by three primary research assumptions in the literature (Saunders et al., 2015).

- **Ontology** refers to humans' assumptions regarding the nature of reality, i.e., what humans believe about the nature of reality. These assumptions define the research options for the studied problem.
- **Epistemology** refers to humans' assumptions about ways of acquiring their previous knowledge and ways of making such knowledge acceptable, reliable and legitimate enough to use. These ways are usually used to improve humans decision-making strategy and to convey this knowledge to other people (Burrell and Morgan, 2017). Epistemology raises questions about the reliability of knowledge sources and their accuracy.
- **Axiology** refers to the beliefs and ethics of the research problem. In this category, the scholar task is to establish ways of dealing with humans' values and research problem values. The research philosophy and the data gathering methods reflect the scholars' values for the research.

3.1.1 POSITIVISM

Positivism takes the scientific strategy as the best way to establish the truth and impartial reality. This philosophy includes working with observable social reality, e.g., social entities in an organisation, and utilising measurement that can be confirmed. Positivism provides certainties for clear-cut and accurate knowledge (Thornhill et al., 2009). Based on Thornhill et al. (2009) results, positivism is value-free research. The researchers are independent, isolated and impartial in what they are researching. In practice, a positivist is not involved in the environment they are investigating. In contrast, researchers in other paradigms need to heighten their understanding of their studying environment characteristics.

3.1.2 REALISM

Realism philosophical outlines that reality is the independence of the researchers' mind, that is, outside reality exists (Bhaskar, 2013). Realism uses a scientific method to obtain knowledge (Thornhill et al., 2009). This assumption supports the gathering of data and the understanding of the collected data. The meaning and importance of realism in research can be understood more when comparing two classes of realism (i) direct realism and (ii) critical realism. The first class, direct realism, declares what the researchers observe is what they get; the world is characterised

precisely by what they perceive through their senses. The second class, critical realism, looks at clarifying the things researchers see and experience and claims the world around is about impressions and pictures of the real world (Thornhill et al., 2009), see Figure 21. Critical realism states that there are two phases to know the world (i) feelings and actions humans encounter and (ii) mental process that takes place at some point after the experience.

3.1.3 INTERPRETIVISM

Interpretivism, also known as interpretivism, refers to the researcher's methodology to develop a better understanding of the research problem and its complexity in a particular context, rather than attempting to generalise the understanding for the entire population. In this paradigm, the goal of the research relies on the participant's experience and behaviour in the studied problem (Ortiz and Greene, 2007). Often, the interpretivism paradigm is associated with quality research (Denzin and Lincoln, 2008).

Interpretivism argues that individuals' behaviour and their social environment is better understood physically (Thornhill et al., 2009). Interpretivism research aims at developing novel, deeper insights and understandings of the human's behaviour and contexts, and this means investigating the research problem from different groups of people.

3.1.4 PRAGMATISM

The pragmatist philosophy is linked to how social and physical realities can be interpreted are diverse. Hence, they require multiple modes of inquiry to permit intelligibility (Thornhill et al., 2009). The adoption of such an approach allows researchers to address their research questions to determine methods of inquiry that they utilise to develop knowledge in their field. A variety of approaches and methods, including quantitative and qualitative, are typically included and guided by the pragmatism paradigm (Wilson, 2014). Furthermore, research projects that adopt pragmatism approach might integrate both positivist and interpretivist perspectives (Thornhill et al., 2009)

3.1.5 RATIONALE FOR SELECTING PRAGMATISM

Before following a research philosophy in this thesis, the data collection strategies associated with each philosophical approach is presented in Table 4. After understanding the research paradigm's implications in this section, the research followed a pragmatism approach. The adoption decision was because of the following reasons: (i) its ability to support an in-depth understanding of human decision-making through multiple modes which is the main trigger for trust calibration (ii) pragmatism can combine both, positivist and interpretivism positions within the scope of single research based on the nature of research questions, (iii) it allows researchers to use a multimethod quantitative and qualitative approach to better understand the inter-relation

between explainability and trust calibration. Therefore, this thesis has adopted multiple qualitative data collection techniques, e.g., semi-structured interviews and think-aloud protocol.

TABLE 3 DATA COLLECTION STRATEGIES ASSOCIATED WITH EACH PHILOSOPHICAL APPROACH

Data gathering approaches	
<i>Positivism</i>	<ul style="list-style-type: none"> • Large samples • Predominantly quantitative
<i>Realism</i>	<ul style="list-style-type: none"> • Methods are chosen based on the research problem quantitative or qualitative.
<i>Pragmatism</i>	<ul style="list-style-type: none"> • Multimethod qualitative or quantitative • Mixed methods
<i>Interpretivism</i>	<ul style="list-style-type: none"> • Small samples with in-depth engagement with participants • Predominantly qualitative

3.2 RESEARCH APPROACHES

The research approach identifies the broader conceptual framework that the researcher will follow to manage and order the selected research activities. The literature has identified two primary research approaches: deductive and inductive. Figure 7 illustrated the beforehand mentioned approaches.

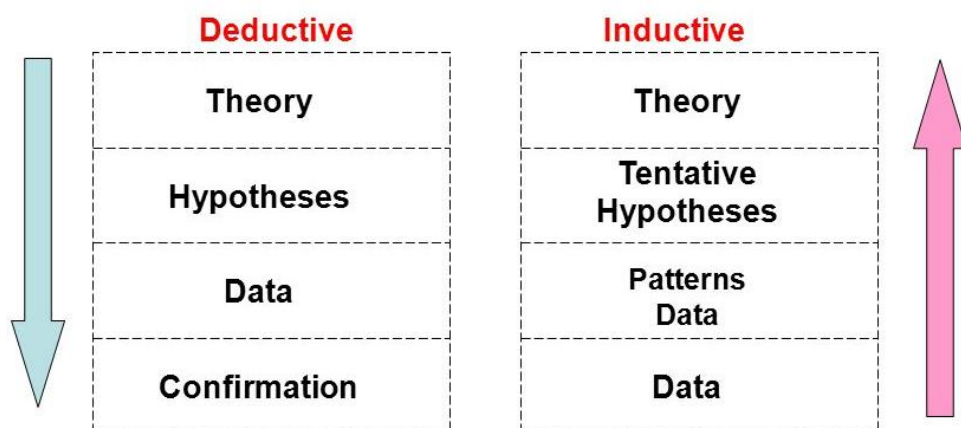


FIGURE 7 RESEARCH APPROACHES ACCORDING TO SAUNDERS ET AL. 2009

As shown in Figure 8, the deductive approach is top-down, while inductive is bottom-up. The deductive approach starts with testing the theory before the actual research, and then, it is followed by data collection methods to confirm or reject the proposed hypothesis. On the other hand, the inductive research approach reverses the previous process. It begins with an

observation, followed by an analysis of that observation, and ended with creating a new hypothesis and theory. Then, the generated theory and hypothesis can be validated using a deductive approach. The deductive approach is suitable for research that starts with an exact and specific research question. In contrast, the inductive approach moves from general to precise research questions. Furthermore, it has been shown that the inductive research approach is more appropriate for a research problem that seeks to investigate complex human behavioural and social issues. It helps to illustrate casual relationships and identify patterns in a specific social context.

Reasons for using the Inductive approach. An inductive approach was selected in this thesis because the research problem seeks to understand trust calibration as a users' behaviour while interacting with XAI interfaces. Such a choice helped the flexibility of understanding the research problem and understand the complexity of human behaviour during the Human-AI collaborative decision-making task. All these dynamics cannot be captured following the deductive approach.

The inductive research approach was especially relevant for this thesis research problem because the research concepts were still researching gradually in the literature and being examined from various perspectives. Thus, there was no substantial research to build upon and obtain hypotheses. More profound knowledge and richer interpretations were needed to create insights that might lead and narrow down the research and provide data regarding trust calibration and XAI interface design.

3.3 RESEARCH STRATEGIES

Research strategies are processes that follow multiple steps, from a general plan to detailed data collection and analysis methods. The research strategy includes several choices that need to be considered to reach the research objectives and answer the main research questions. Factors such as the type of the research problem, skills of the research team involved, and the targeted participants influence the research approach's selection. This section discusses the main research strategy used in this thesis, its main characteristics, and potential disadvantages.

3.3.1 QUALITATIVE RESEARCH APPROACH

Qualitative research is mainly exploratory research (Berg, 2004). It is used to gain knowledge for the main reasons, opinions, and motivation regarding a research area or problem. It produces an extensive understanding of the research problem and supports promoting ideas, conclusions or hypotheses for potential quantitative research. Qualitative research aims to identify and study the targeted participants' behaviour and beliefs in the studied research problem. It employs closed and in-depth studies of small groups of individuals to lead and help construct hypotheses. The results of qualitative research are descriptive rather than predictive.

Qualitative research is described by the aim to study real-world problems as they emerge from humans' behaviour (Creswell and Creswell, 2017). It is also for scholars who are free to new ideas arise, i.e., there are no pre-set constraints on findings. Qualitative research is characterised by the choice of cases, e.g., people, cultures, organisations and communities. They enrich the research with useful and deep insights from their experience. This is different from the quantitative approach, which is an empirical generalisation obtained from an adequate sample from a population.

Recent research by Berg (2004) has identified five potential disadvantages of using qualitative methods into a research problem in the social sciences:

- The possibility of drifting away from the research's main objectives due to changes in the study context during the qualitative studies.
- Qualitative studies can arrive at multiple conclusions and results based on the targeted participants' knowledge, experience, and circumstances as well as the researcher knowledge.
- Qualitative research is usually a time-consuming process.
- The analysis and the interpretations of the collected data could be subjective to the researcher bias.
- A high level of experience from the research is required to gain the targeted audience's required knowledge.

Reasons for using qualitative strategy. A qualitative research approach was chosen for this thesis. As utilising XAI interfaces for trust calibration goals in human-AI decision-making tools is an emerging topic in human-AI interaction and there was not enough knowledge about the problem in the literature, an in-depth exploration of the problem was needed. Such an approach helped the research understand users' behaviour, knowledge, personal experience, issues and concerns regarding the current advances in explainable AI literature.

Qualitative research design has been categorised into five research designs (Creswell and Poth, 2016, Lazar et al., 2017).

1. Grounded theory. It refers to a theory that is grounded or generated inductively from the participants' explanation of the problem (Chun Tie et al., 2019). It is sufficient for investigating peoples' behaviour. This approach is preferred when the research problem is little identified in the literature to generate or build an explanatory theory. In this approach, the research aims to develop an overall understanding of individual daily experiences in a particular setting as well as collect their opinions and insights about the research problem. Two main features have been identified in this research design: (1) constant comparison of the emerging categories, (2)

theoretical sampling of a variety of groups to increase the validity of the data (Chun Tie et al., 2019). The process of conducting grounded theory research is presented in Figure 8.

In this research design, different qualitative data gathering approaches can be followed, e.g., semi-structured interview, think-aloud protocol and co-design sessions (Corbin and Strauss, 2014). It is an effective method in the absence of the theoretical framework that guides the data collection strategy (Chun Tie et al., 2019).

Grounded theory is the research design that has been followed in this thesis. First, the choice is related to the inductive nature implemented in this design approach. Also, it is a valuable method for addressing many of this thesis focal aims (i.e., understanding the concept of calibrated trust in the presence of AI explanations, users' behaviour and interaction style with AI explanations). This has inspired the analysis of the user studies conducted in **Chapter 5** and **Chapter 6**. Finally, the flexibility of this thesis and its question-driven characteristics are compatible with grounded theory principles.

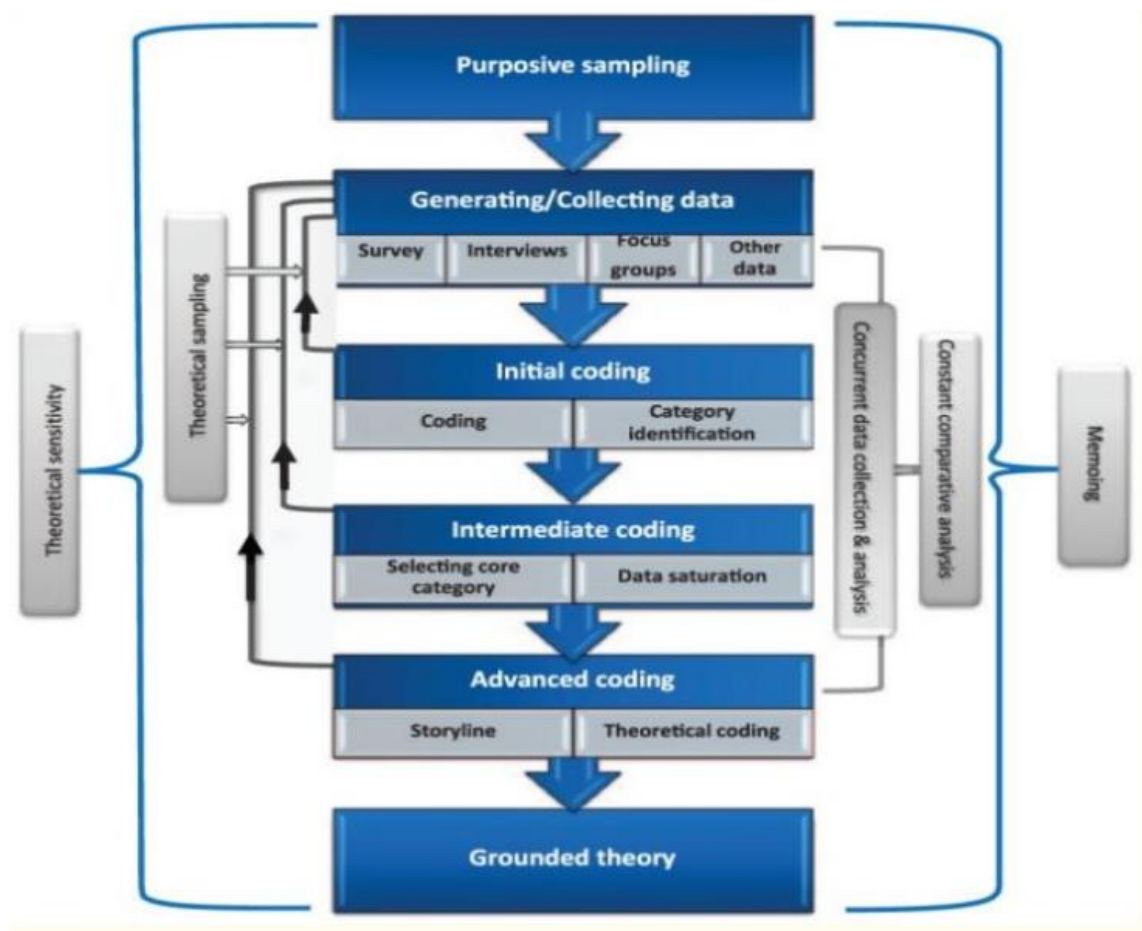


FIGURE 8 A DESCRIPTION OF THE GROUNDED THEORY INTRODUCED BY CHUN TIE ET AL. (2019).

2. Case study. This strategy includes an in-depth examination of an event, activity, software or one or more persons. Case studies can be utilised for a descriptive exploratory as well as an exploratory research goal. Researchers in this strategy can benefit from combining both qualitative and quantitative data collection methods. It is a common approach for evaluating research findings in a real-world environment. Therefore, to achieve **Objective 5**, this thesis conducted a case study approach in **Chapter 9** to evaluate C-XAI design method.

3.4 TIME HORIZONS

It refers to the time frame of conducting the research (Saunders et al., 2015). The time zone can be categorised into two types, (i) cross-sectional time zone and (ii) longitudinal time zone. Choices regarding the time zones depend on the data collection strategies and the research questions under investigation. The following sections describe the two types of time zones:

1. Cross-sectional. It collects data from the targeted population over a specific time frame, e.g., investigating people opinions about a specific topic during an interview study. It is appropriate for both quantitative studies such as surveys and qualitative studies such as interviews.

2. Longitudinal. It requires a longer time to answer the research question under investigation, e.g., studying the changes in human behaviour over a long period of time to understand the development and changes in that behaviour. Adopted methods for such time zone could be a diary study.

This thesis followed a cross-sectional time zone. The reason for that is related to the nature of the research problem in this thesis which does not require a study of humans' participants over a long period of time. In particular, this thesis looks at gathering data regards users' behaviour with the XAI interface and their main issues and difficulties during a specific time frame. Such knowledge would help the research to gain knowledge on why XAI interfaces are not improving trust calibration. The research problem does not focus on the changes in users' behaviour with the XAI interface over time. Therefore, the thesis adopted a multi-stage cross-sectional approach to work closely with the end-users and understood the research problem's dynamics.

3.5 ADOPTED DATA COLLECTION METHODS

The core of the research onion consists of data collection and data analysis methods (Saunders et al., 2015). It focuses on a variety of methods and techniques to be followed to collect data and analyse it. The following sections discuss data collection and analysis methods used in this work.

3.5.1 DATA COLLECTION

This thesis follows an inductive approach, i.e., bottom-up, which starts from the data collection to analyse the data and ends up with the theoretical level. Therefore, a qualitative data collection method is adopted to help achieve this thesis aims and objectives identified in **Chapter 1**. A generic summary of the data collection methods will be discussed in this section. In contrast, a detailed description of how the methods were implemented to help achieve the research outcome will be given in the respective chapters. An overview of the data collection methods used in this thesis is presented in Table 5 and their location in this thesis.

TABLE 4 DATA COLLECTION METHODS USED IN THIS THESIS

Data Collection Method	Type	Location
Semi-structured interviews	Qualitative	Chapter 5
Follow-up interview	Qualitative	Chapter 5, Chapter 6
Think aloud	Qualitative	Chapter 6
Focus group	Qualitative	Chapter 4, Chapter 7, Chapter 9

3.5.1.1 INTERVIEWS

Interviews help researchers collect data relevant to the research objectives, aims, and questions. It is a valid and reliable qualitative data collection method (Saunders et al., 2015). This method involves a one-to-one conversation with the research target audience to gather their opinions and experience regarding the research problem. An interview can take three different styles:

1. Structured interviews. Researchers follow this style to collect precise information about the studied problem, where the interview questions are defined and prearranged. However, a structured interview requires a deep understanding of the research problem and identifying the questions precisely and the time associated with every question—open-ended questions such as what and why and closed-ended questions such as when and where. A structured interview restricts the researcher within the scope of pre-defined questions.

2. Unstructured interviews. Researchers in this method do not prepare questions for the interview, and the collected data is informal. Such a process makes the unstructured interviews an unreliable method from a research point of view. This is because the collected data might be biased and constrained based on the interview conversation and the variations of the interviews between different participants.

3. Semi-structured interviews. In this style, the researcher prepares a set of structured questions before the interview, and the researcher has the freedom to ask unstructured questions during the interview. Additional questions used in this method are usually to elaborate and expand the structured questions.

This thesis adopted the semi-structured interview style for the studies carried on in **Chapter 5** and **Chapter 6** to establish an understanding of the lived users' experience during their interaction with the XAI interface. Also, the semi-structured interview was helpful to identify users' concerns, difficulties and needs for each explanation class. This style helped this research in collecting related data needed for **Objective 3**. The adoption of the semi-structured interview took into account the following guidelines for conducting this style:

- 1. Reproducible.** The procedure of the semi-structured interview can be repeated by another research to produce similar data.
- 2. Systematic.** The process avoids a bias of participants that help the researcher previous idea and insights.
- 3. Credible.** The questions were designed to produce a valid explanation for the research questions.

3.5.1.2 THINK ALOUD METHOD

The Think Aloud (TA) method provides in-depth verbal data about the participants' reasoning process during a problem solving or a decision-making task (Fonteyn et al., 1993). Adopting this method in a research project helps the research identify the information connected to the problem-solving task and how it can solve the problem. Scholars used the TA when investigating the performance of the participants in a particular context or usability testing. According to (Fonteyn et al., 1993), the advantage of using TA compared to other observation studies is linked to the cognitive processes of the participants during the study, thus revealing the available information in the working memory. However, other researchers also criticised it because the thinking aloud process can limit participants' performance during the problem-solving task (Nielsen, 1994). Researchers conducting this method need to have experience in observing people and reliably interpreting their cognitive processes. TA was used for the first time to gain insights into medical practitioners' problems while working with the paper-based guideline (Nielsen, 1994).

In this thesis, the adoption of the think-aloud method is related to **Objective 3**. The TA's main aim was to observe participants behaviour and beliefs about the AI-based explanation during their Human-AI collaborative decision-making tasks. This was also helpful to identify categories of trust calibration risks. To ensure the validity of the TA results, the research followed the guidelines identified in (Nielsen, 1994).

1. The researcher informed the participant that the research team was not involved in creating the explanation interface design. This was an assurance for the participants to give an honest answer.

2. The researcher gave several examples to the participant of how they should think aloud, e.g., what they like and what they do not like in the interface.
3. It is common for participants to forget thinking aloud during the study. Research had given the participants polite and neutral prompts to remind them.

3.5.1.3 *FOCUS GROUP*

Focus groups are a cost-effective way to gather rich and reliable qualitative data from a group of participants focusing on a single research problem (Lazar et al., 2017). Focus groups are structured discussions with the targeted participants to elicit specific data related to a research question (Saunders et al., 2015). The rise of participatory research has led to this method's success (O. Nyumba et al., 2018).

The researcher who runs the focus group is called the facilitator of the session. The facilitators' role includes managing the session, ensuring the participants' discussion is still within the scope of the research problem, and encouraging group discussion. In general, the number of participants in each focus group can range from six to twelve participants (Krueger, 2014). Non-probability sampling is usually followed in focus groups data collection method, e.g., people with the right experience and understanding the particular research problem (Saunders et al., 2015). The facilitator of the focus group shall be aware of the following challenges to data bias, (i) dominant participants influencing other participants opinions, (ii) encouraging silent participants to participate in the discussion, (iii) side discussion between the participants, and (iv) long discussion on a specific question.

In this thesis, focus groups were followed in **Chapter 7** as a validation method to the participants' designs in Co-design sessions. This was meant to critically analyse and evaluate the participants' ideas to formulate robust solutions. This activity allowed our participants to explore various ways of using AI explanations in their work environment, considering trust calibration as the primary goal. Also, a focus group was used in **Chapter 4** to refine the taxonomy results and the mapping between users' questions and each explanation class. Finally, **Chapter 9** used a focus group in the case study evaluation to evaluate C-XAI design method.

3.5.2 ANALYSIS METHODS AND TOOLS

Data analysis methods are usually adopted in research to analyse and evaluate the collected data from the previous data collection methods. This thesis follows a qualitative data collection approach to answering the research questions; the following section discusses two qualitative data analysis methods: thematic analysis and content analysis.

3.5.2.1 *THEMATIC ANALYSIS AND CONTENT ANALYSIS*

Thematic analysis and content analysis are methods used to analyse the collected qualitative data by marking, coding and sorting data based on the emerged themes in the data or predefined themes (Ritchie et al., 2013). According to (Corbin and Strauss, 2014), thematic analysis and content analysis include three main stages. It starts with working on the collected data, describes the group of participants, the studied system, and the interaction between participants and the studied problem. Second, the researcher tries to understand the relation between different elements and the nature of each component. Finally, the collected information from investigating each component is utilised to reach a valid conclusion about the studied problem.

Qualitative content analysis methods can be applied to transcribed communication, such as interview transcriptions or note-taking during an observational study. Content analysis also can be applied to videotapes and written articles. Content analysis method goals include transforming a large amount of qualitative data into valid knowledge that can be understood and represented into valid facts (Krippendorff, 2018). In general, qualitative researchers use content analysis to develop a model, conceptualise a specific problem, and classify several concepts (Elo and Kyngäs, 2008). The main advantages of content analysis are the flexibility, ability to handle complex data, integration with different data material, and integration with the study context. It is also a controlled process where it consists of a set of rule-based steps.

On the other hand, the thematic data analysis method is used to identify, analyse, and report patterns within data (Braun and Clarke, 2006). The thematic analysis process consists of six key stages (Braun and Clarke, 2006):

1. Familiarisation with the data.
2. Generation of initial codes.
3. Searching for themes.
4. Reviewing themes.
5. Specifying and naming themes.
6. Writing up the report.

Content analysis and thematic analysis connect in various ways, including segmenting the collected data, searching for themes. However, they vary in the probability of data quantification. This means that it is possible to calculate the occurrences of the themes in the data in content analysis. On the other hand, the thematic analysis gives more detailed interpretations and a rich understanding of the data.

This thesis used both content analysis and thematic analysis. In **Chapter 4**, content analysis was used to analyse the literature Explainable AI and build the explanation classes taxonomy. In **Chapter 5, Chapter 6 and Chapter 7**, the collected data were analysed using the thematic

analysis method. The research followed the six stages mentioned before to reach **Objective 3** and **Objective 4**.

3.5.2.2 *QUALITATIVE RESEARCH ANALYSIS TOOLS*

NVivo software is a tool that supports qualitative and mixed methods analysis. It is developed to help the researcher to organise, analyse and categorise their unstructured qualitative data, e.g., interviews and open-ended surveys (International Copyright © 1999-2014). The aim of adopting this software in this thesis is its ability to save time. It is also usable and comfortable to use. Also, NVivo software helped to visualise the qualitative data, identify links and find new insights.

3.5.2.3 *HUMAN-COMPUTER TRUST MODEL*

The Human-Computer trust model was introduced as a classification for trust components in human-computer trust environments (Madsen and Gregor, 2000). Two main classifications comprise the model: Cognition-based trust and affect-based trust. Cognition-based components are based on the users' intellectual perceptions of the systems' characteristics, whereas affect-based components are based on the users' emotional responses to the system. Cognition-based trust has three main dimensions: perceived understandability perceived reliability, and perceived technical competence. Affect-based trust components are personal attachment and faith. The Human-Computer trust model was used as a starting point for understanding the features of XAI interfaces that enhance the trust calibration process. This model was used in **Chapter 5**.

3.6 DESIGN METHODS

User-centred design (UCD) and Co-design approaches are discussed in this section. Design approaches, in general, are used in the literature to involve users in the design process. Design approaches can lead to a better understanding of the end-user needs, which enhances the possibility of the solutions' acceptance (Song and Adams, 1993). In this research, we discussed and negotiate how to embed AI-based explanations to serve users' needs, workflow and trust calibration to fulfil **Objective 5**. Together with the participants, the research conceptualised and sketched design features to support users in utilising AI explanation and reduce trust calibration errors. This could be achieved by giving the participants initial prototypes or mock-ups (Clement et al., 2012) of the problem to help them visualise the idea and then provoke brainstorming related to the research problem. All these dynamics were hard to capture during the first phase. Therefore, design approaches helped the research to come up with innovative designs of how the solution should look from the user perspective. In the following sections, a discussion about the adopted design approached is presented.

3.6.1 CO-DESIGN

Co-design involves end-users directly in the design team of the product. It ensures that users' preferences and needs are considered in the earliest stages of the software development process (Sanders, 2002b). The co-design approach and UCD approach are similar to their goal of centring the system's design on the end-users. The main difference is that Co-design places a more prominent focus on user engagement at the design phase.

As explainability is a well-studied phenomenon in social sciences and psychology, therefore implementing AI-Human explanations shall meet the same criteria of human to human explanations (Miller, 2019). As such, explanations that can help users calibrate their trust with a satisfactory design depend on the users classically, e.g., discovering and validating users' needs and their involvement in the design phase. Design solutions to mitigate trust calibration errors can be identified and agreed upon at the design's earliest stages.

Chapter 7 discusses using the Co-design approach in this thesis to explore the design principles and techniques to support users' in calibrating their trust while interacting with AI-based explanations.

3.6.2 SCENARIO-BASED APPROACH

Scenarios are stories about users' activity in a system. For instance, a student wants to complete an online assignment. Thus, a scenario-based approach is introduced to tell different stakeholders involved in the design process a story about a specific problem (Sutcliffe and Carroll, 1998). It aims to provide a narrative description of a real-world scenario to capture its requirement. (Sutcliffe, 2003) classifies scenarios in two main classes, (i) scenarios that describe the system and its social environment, and (ii) scenarios that contain a sequence of events related to a specific problem.

In this thesis, the scenario-based approach was followed to stimulate focus groups and co-design sessions. A detailed description of the method is presented in **Chapter 5** and **Chapter 6**. This approach helped the researcher engage participants in the research problem and understand the main ideas and concepts related to the research problem. As an outcome, the participants could become interested in the research. Scenarios could also be used as a warm-up activity, assisting the participants in developing their scenarios and explaining how trust calibration errors occur or how specific designs could be leveraged to lessen such errors. As such, diverse and rich data could be collected with intense ideas to address every research problem.

3.7 SOFTWARE DESIGN APPROACHES

This section discusses multiple software design approaches. The core design approaches describe different ways users can be integrated into the system design. Collecting data on users' interaction style and identifying their XAI interfaces' requirements could explore different trust

calibration contexts and how XAI interfaces for calibrated trust can be implemented. As the proposed method in this thesis is meant for the elicitation of design principles for XAI interfaces which will ultimately lead to support users in their trust calibration process. Understanding users' explainability requirements for the new design layer would help analyse how the interface would be designed and how it could help users' trust calibration process. To attain this, representative users' active participation in the design process would increase the design acceptance, adoption and mitigate undesired behaviour with the XAI interface. In the following sections, a discussion around the possible approach for the proposed method and the selected approach is provided.

3.7.1 USER-CENTRED DESIGN

User-centred design (UCD) is a design approach that involves the end-user in the design process and focuses its attention on the needs, expectations and disadvantages of the product. It also measures product effectiveness in real-world scenarios (Abrás et al., 2004). UCD aims to avoid asking the users what they want from the product; instead, it makes sure the final product fits users need (Wever et al., 2008). UCD tries to support the designer in achieving the goal of system design by understanding its end-users. Users of a product are involved and considered in the product development process. A UCD approach includes incorporating the user perspective in the product cycle to increase users' acceptance and foster trust (Johnson et al., 2005). Overall, the main aim of the UCD approach is to develop software that fits the users' task and satisfy their needs (Teixeira et al., 2012).

UCD is an association between HCI and design practices when users' engagement is the main principle of ensuring the users' needs (Marcus and Wang, 2017). In the UCD approach, users help the design process by effectively using their knowledge and information to utilise the software. Their involvement can discover various errors and mistakes. The role of the design in UCD is not only to discuss the technical aspects of the software during the analysis process but also to look at users' needs, behaviour and issues (See Figure 6) (Wever et al., 2008).

Previous research (Abrás et al., 2004, Teixeira et al., 2012) described several guidelines on how and when to involve end-users in the design process and what data shall be collected from them:

- Designers of the systems are encouraged to conduct interviews and surveys to understand users' expectations and needs.
- Designers are also expected to conduct interviews and survey to justify the flow of the research at the earlier stages of the design.
- Several focus groups and observational studies on-site are recommended at the early stages of the design to gather environmental settings and knowledge where the proposed system could operate.

- Low-fidelity prototyping, High-fidelity prototypes, role-playing, walkthrough, and simulation are recommended during the middle stages of the design process to assess the design process and gather modified requirements.
- Usability testing, interview and experiments at the final stages of the design to evaluate users' acceptance and systems' goals through a collection of qualitative and quantitative data.

3.7.2 PARTICIPATORY DESIGN

Participatory design (PD) facilitates users' engagement in the design process to affect the system design directly. In PD, end-users are the central part of the design process (Sanders, 2002a). Under PD, users' needs and behaviour and their understanding of the system guide the process of the system design. Several researchers have proposed tools and techniques for implementing the PD approach. The user engagement model in online behaviour change interventions design was presented in (Short et al., 2015). They used the PD approach to design an online behavioural change tool so that the design is effective and accepted by end-users. PD encourages users to participate in the design process from the early stages. This includes solving the design problem based on users' understanding of the problem. This would enhance the system designer's ability to collect features, elements, requirements, and functionalities crucial for end-users.

The goal of PD is to incorporate all stakeholders in each stage of the system design. Stakeholders of a system could include designers, developers, end-users, clients, examiners and domain experts (Kang et al., 2015). PD facilitates users' perception of ownership of the system, acceptance and thus better results. PD can be implemented using various methods such as focus groups, ethnography, prototyping and card sorting (Kang et al., 2015). Such methods can generate design solutions that can be used for the full benefit of each stakeholder.

Common stages for implementing PD in real-world scenarios are presented in recent research (Spinuzzi, 2005).

- Stage 1 Exploration of the work. This phase includes meeting the end-users to understand their work environment and their natural setting. Ethnographic methods, such as observations and interviews, are standard methods used in this phase.
- Stage 2 Processes discovery. It refers to a cooperative interaction between the end-users and the design team to prioritise and identify goals, features and expected results. Furthermore, this stage includes defining main design concepts and techniques in the system, e.g., adaptive and interactive.

- Stage 3. Iterative prototyping. It includes several numbers of methods to design the final system iteratively. These activities can be performed on-site or in laboratory settings (Spinuzzi, 2005).

The design team implementing the PD approach shall consider the users' interventions effect on the design. For instance, users' interests and needs might clash with the system's goals and values (Kujala and Väänänen-Vainio-Mattila, 2009). Therefore, some of the users' needs and goals are not suitable for implementation. For good practice, researchers and designers following PD approach shall create approaches and guidelines to manage users' involvement in the design process, particularly, a problematic user who might deny the existence of the problem.

The success of major design firms such as IDEO (Kelley, 2001) has shown the benefits of user-centred design (UCD). This approach focuses on users to be the main source of innovation, and design companies can recognise novel designs through eliciting users' needs and preferences (Jokela et al. 2003). Companies conduct UCD approach by conducting user-based research, e.g., asking about user needs, or more effectively, observing their behaviour and usage style. On the other hand, participatory design is an approach that endeavours to actively involve all stakeholders (e.g., employees, partners, customers, end-users) in the design process to help ensure that the design meets users' needs. In participatory design, participants are invited to collaborate with designers, researchers and developers during an innovation process. Potentially, they engage during various stages of this process: they engage during the initial exploration and problem formation both to help define the problem and to focus ideas for a solution. Participants also participate during the evaluation phase where they help evaluate proposed solutions (Sanders, 2006). Participatory design, which is used to engage actual users in design activities, represents an example of a research method developed to support design work during concept generation and development phases.

This thesis argues that the nature of the trust calibration problem requires involvement from multiple stakeholders in the design process rather than only focusing on users' behaviour and needs. Research has shown that calibrating users' trust requires understating the complexity of human trust, decision-making theories and cognitive biases (Lee and See 2004). Furthermore, generating AI explanations requires a technical understanding of the underlying XAI methods. All these dynamics require input into the design process from AI experts, psychologists, human factor experts is crucial to develop XAI interfaces where calibrated trust is the main design goal.

3.7.3 VALUE-SENSITIVE DESIGN

The value-sensitive design approach is adopted in the design process when the goal of the system is to consider human values. Value in values-sensitive design is defined as “what a

person or group of people deem important?”(Friedman et al., 2008). Value-sensitive researchers argued that considering and recognising humans’ values in the software can considerably enrich the software design. Neglecting humans values in the design process has shown several side effects on user experience (Nathan et al., 2007). The main challenge for value-sensitive designers is deciding which values have a higher importance in software design.

System values can be gathered from multiple sources. These sources include individuals, colleagues, social groups, organisations or groups of people from the same demographic (Shilton et al., 2014). In general, setting the research boundaries and several aspects related to the study depend on choosing the values’ sources. Shilton et al. (2014) presented seven techniques to collect human values in social computing systems, for example, surveys, interviews, values advocacy, content analysis, design practices and ethnographies.

3.7.4 USER STORIES

User stories are considered a useful tool by software engineers to gather users’ requirements and depict software requirements. It is mainly used with agile approaches. User stories consist of a designed template and a set of guidelines to complete the template. A strong association between user stories and scrum has been shown(Lucassen et al., 2016); (a) user stories assist software developers to outline the software requirement, (b) user stories justify the requirement.

According to Lucassen et al. (2016), increasing the number of user stories in the system developed increased the collected requirements' efficacy. They compared the case of having too many short user stories with having fewer long user stories. They argued that when the story is long, it is called an epic. Epics then can be divided into shorter user stories. The recommended approach of breaking down an epic into multiple user stories is to discuss such information with the end-users. Thus, the story would be relevant and worthy to the end-user.

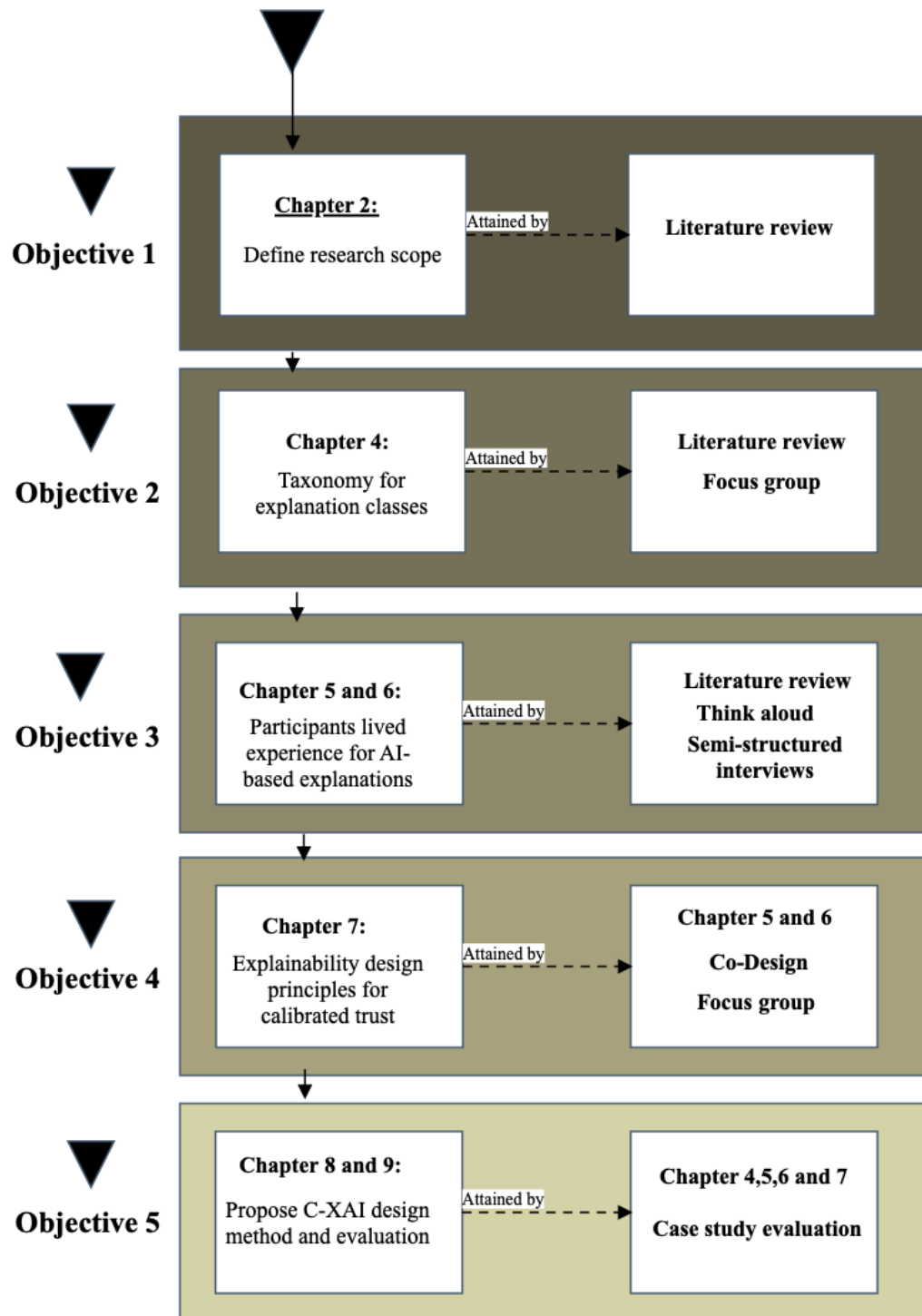
3.8 RESEARCH ETHICS

Research ethics are made to protect the rights of the research participants (Orb et al., 2001). Research ethics is based on the principle of voluntary participation which demands that individuals should not be obliged to participate in a research study. Furthermore, the principle of voluntary participants is the main requirement for informed consent (Berg, 2004). Informed consent means that participants need to be informed precisely about the procedure of the study and the risks involved in the study. As an ethical standard, the researcher should avoid putting the participant in a situation where they may face harm due to their participation in the study. Risk can be either physical or psychological. Researchers have to protect the privacy of the participants and guarantee their confidentiality. Researchers should also inform the participants about the accessibility permissions for their data, i.e., who will access their data. Finally, the researcher should also ensure that participants will remain anonymous during the study.

To address the different ethical considerations in this thesis, the researcher obtained ethical approval from Bournemouth University Research Ethics Committee (UREC). The ethics checklist was reviewed and approved. The ethics ID is 32253. The studies conducted in this thesis were below the minimum risks based on the ethical code document. This means that there were no specific risks for participating in these studies. The research also submitted ethical documentation, including a study information sheet and study agreement form. These documents were proposed to all participants in this thesis before conducting the study. The agreement form was signed by participants and indicated the consent to participate in the studies. The information sheet helped participants to familiarise themselves with the study and understand their rights during and after the study. All data collected during this thesis was anonymised data. The recorded files were transcribed and deleted after that. Samples of the documents are available in the Appendix Section.

3.9 CHAPTER SUMMARY

This chapter summarises research paradigms, research approaches, and strategies as choices that this research might follow. The steps and processes involved in each approach and design adopted are not provided in this chapter, and they can be found in their corresponding chapters (See Figure 9). This is meant to provide the information associated with the activities carried towards driving every step and process employed from the research methodology. The next chapter reviews Explainable AI literature and maps users' questions to each explanation class. The chapter also presents a detailed description of each emerged class from the literature.



4. CHAPTER 4: A TAXONOMY FOR EXPLAINABLE AI CLASSES

4.1 INTRODUCTION

The inherited complexity of machine learning models is an unquestionable feature. These algorithms becoming increasingly adopted in different application domains and becoming increasingly able to be applied in different tasks. This expansion of such models has accelerated the need for and importance of explaining the machine learning predictions and recommendations. While the huge efforts of improving machine learning models in terms of accuracy and performance, interpretable machine learning research has gained a lot of attention in recent years. This chapter is divided into two sections. Section one presents the analysis performed to derive to provide a conceptualisation of interpretable machine learning models. Section two presents a taxonomy for explainable models in the literature of interpretable machine learning.

4.1.1 IDENTIFY DIFFERENT TERMINOLOGIES

The literature lacks a standard terminology of explainability in machine learning, where several different terminologies have been used in the literature of explaining machine learning models. “Interpretability” and “explainability” are the most used concepts in the literature, often used equivalently. Adadi and Berrade (2018) stated that explainability and interpretability are closely related concepts: when the system is interpretable, it is explainable if its components can be understandable by the human. However, in the ML research community, the concept of interpretability is more used than explainability. Results from Google Trends comparison between both concepts confirm that finding (See Figure 10). The Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

Although many researchers used the same concept, other authors develop a definition to describe the difference between the concepts. Furthermore, a recent review from the UK Government House of Lords of AI provided fundamental conceptualisation for different terminologies(Lords, 2018). They noted, “the terminology to describe interpretability used by researcher varied widely”. Many used the concept transparency, while the other used explainability or interpretability, even sometimes intelligibility”. Other authors also used legibility. This thesis will adopt explainability as a general concept in the sense of its understandability for humans.



FIGURE 10 COMPARISON BETWEEN THE TERMS “INTERPRETABLE MACHINE LEARNING” AND “EXPLAINABLE MACHINE LEARNING” FROM JANUARY 2004 UNTIL APRIL 2021.

4.2 TAXONOMY FOR MACHINE LEARNING EXPLANATIONS

This section introduces a holistic taxonomy of the machine learning explainable model provided in explainable machine learning literature. This is meant to help recognise the different explanation classes and techniques that can generate meaningful explanations from a given black-box model. As such, creating a taxonomy was needed to facilitate a systematic overview and encapsulate the pertinent concepts that belong to explainable machine learning. This taxonomy will be used as a reference point for the stakeholders of the C-XAI method about what can be explained to users’ given a black-box model. It will also help the research in designing further user studies in this thesis.

There have been many efforts in the literature to create taxonomies of interpretable machine learning models (Guidotti et al., 2018b, Arrieta et al., 2020). Researchers used common dimensions to classify explainable models: 1) The scope of the explanation, i.e., either to explain the entire model results (global explanation) or explain specific prediction (local explanation); 2) The degree of complexity of the model. 3) Main prosperities of the generated explanation such as accuracy, fidelity and consistency (Carvalho et al., 2019, Arrieta et al., 2020). However, explanations are answers to users’ questions, and humans are the main recipient of such explanations. The emerging taxonomies in the literature focus on providing these taxonomies for machine learning experts and data scientists’ where the underlying explainable model is the main focus. Hence, unlike previous work, this study focuses on analysing the XAI literature to identify classes of explanations and construct a taxonomy that maps each class to users’ questions. Similar work has identified such a framework, but their work stayed on the high level of abstraction to explainable models and users’ questions (Liao et al., 2020), i.e. they identified the high level of categorisation for explainable models. Until the day of conducting this study, no previous work has aimed to provide a comprehensive overview of explainable models that answer the users’ questions as a list of capabilities presented to them.

4.2.1 RESEARCH GOAL

The goal of this research is to identify different explanation classes used in the literature of interpretable machine learning and map these models to identifiable questions and answers.

4.2.2 RESEARCH METHODOLOGY

Given the diversity of the terminologies and the multidisciplinary nature of the explainability, the bottom-up approach with the aid of the content analysis approach (Elo and Kyngäs, 2008) was used to create an initial taxonomy. Thematic analysis (Hsieh 2005) has been adopted to infer some conclusions. The expert checking method was used to provide an in-depth evaluation of the emerged themes and concepts as well as trustworthiness. To increase the credibility of the methodology, the coder did not start analysing the data unless the text identification step was completed. This meant to eliminate any biased coding, such as refusing non-supportive inference (Hsieh 2005). This study conducts a systematic review in the field of explainable machine learning models to analyse and classify the literature and identify the different explanation capabilities.

To evaluate the taxonomy, the research conducted a focus group study with machine learning and Human-computer interaction experts, see Table 6. Since the research addresses a multidisciplinary problem, having a diverse viewpoint is an essential requirement. The research used judgemental sampling (Westfall, 2009) to select experts who know relevant research including HCI, machine learning and explainable AI. To ensure that the participants have a sufficient understanding of the problem, an assessment survey in Table 7 was used.

While there are various evaluation approaches, e.g. golden standard, application-based, data-driven, and assessment by humans, the purpose of the evaluation should guide the selection among them (Brank and Grobelnik, 2005). As the research evaluation was centred around evaluating the emerging concepts from XAI literature and mapping them to users' questions. Manual evaluation with the aid of domain experts was needed (Brank and Grobelnik, 2005). The manual evaluation method is done by humans who assess how robust the taxonomy satisfies a set of predefined criteria, standards and requirements.

TABLE 5 PARTICIPANTS FAMILIARITY WITH RELEVANT TOPICS

Participants	Age group	Gander	Domain of expert	Experience
P1	30-40	Male	HCI	12
P2	40-50	Male	HCI	23
P3	30-40	Male	Machine learning	9

P4	30-40	Female	Machine learning	11
P5	30-40	Male	Machine ethics	12

TABLE 6 PARTICIPANTS ASSESSMENT SURVEY RESULTS

Participant ID	P1					P2					P3					P4					P4				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Machine Learning			x						x					x					x						x
User experience								x		x				x					x						x
Explainable AI			x						x					x						x					x
Human-AI Interaction				x				x		x				x					x						x
Transparency				x			x			x		x						x						x	
Decision making				x				x				x							x						x
Human-Centred computing							x			x			x					x						x	
The questionnaire was based on the Likert scale which can be interpreted as follows:																									
(1) Very poor (2) Poor (3) Average (4) Good (5) Very good. The above cells represent the 5-points Likert scale, and the x sign shows the experts' answers.																									

4.2.2.1 SEARCH STRATEGIES

To find relevant studies that are related to this systematic review, the research selected databases proved in Table 8, where the research ran different search queries. Other research databases that gather research papers automatically from different sources, such as GoogleScholar, or provide nor-reviewed papers such as arXiv, were excluded from the study search scope. The research adopted databases that include peer-reviewed papers published in computer science. This strategy is meant to provide some evidence regarding the quality of the relevant papers.

TABLE 7 SELECTED RESEARCH DATABASES

Source	URL
ACM Digital Library	http://portal.acm.org
IEEE Xplore Digital Library	http://ieeexplore.ieee.org
ScienceDirect	http://www.sciencedirect.com
Springer Link	http://link.springer.com

4.2.2.2 *SELECTION CRITERIA*

Relevant papers retrieved from the selected databases were filtered using a set of criteria. The research considers three inclusion criteria (IC) and four exclusion criteria (EC) to select relevant papers relevant to the scope of this study. The inclusion criteria and the exclusion criteria are described and summarised in the following points:

- 1- Novelty (IC1): The study proposes a novel technique for explaining the machine learning model.
- 2- Foundation (IC2): The study presents an investigation towards the foundations of the explanations in machine learning systems.
- 3- Language (EC1): This paper is not written in English.
- 4- Duplicated (EC2): The content of the paper was published in another completed form.
- 5- Full content (EC3): The research excluded the papers with no access to the full content.
- 6- Domain-related (EC4): The paper must be centred around explainable models and techniques in machine learning. For example, the search results introduced papers addressing the explanations from psychology, social science and education without direct relation to machine learning; these papers were excluded.

Concerning EC3, the research proceeded as follows. To obtain the papers in which they were published, the search strategy started to access them through the Bournemouth University network. If the full text was not available, the search strategy searched for the paper on the web using the author websites, Google search and other repositories of scientific papers such as Google Scholar and ReasearchGate.

4.2.2.3 *SEARCH STRING CONSTRUCTION*

The search string in this study is based on two main concepts of interest and their synonyms, which both have to appear in the protentional selected papers. The first concept is explanations, which is the main term of this study. However, this thesis identified that the research community had used different terminologies to describe the explanations, and those were added as synonyms of the concept explanation. The second concept is machine learning, which is an umbrella term that covers different domains. As synonyms, the research considers sub-classes of machine learning, including, in particular deep learning, neural networks, reinforcement learning, supervised learning and unsupervised learning. The final search string is shown in Table 9.

TABLE 8 FINAL SEARCH STRING

(explanation OR justification OR interpretation OR intelligibility OR explainable OR interpretable OR intelligible) AND (machine learning OR reinforcement learning OR supervised learning OR unsupervised learning OR deep learning OR neural networks)

The search string was not capable of being applied in all the target databases due to the syntax restrictions. In these cases, the search string was customised according to the new syntax. The search within the abstract was also available in all the databases except the Springer Link database due to API limitations. The research thus searched for the terms in the keywords of the papers in this case.

4.2.2.4 SELECTION OF RELEVANT STUDIES

After searching four selected databases on January 12, 2019, the search resulted in 1285 papers (excluding duplicates). The detailed descriptive statistics for each database are shown in Table 3.

TABLE 9 DESCRIPTIVE STATISTICS FOR EACH DATABASE SEARCH RESULTS

Database	Number of studies
ACM Digital Library	378
IEEE Xplore Digital Library	356
ScienceDirect	344
Springer Link	207
Duplicates	25
Total (including duplicates)	1285
Total (excluding duplicates)	1310

Then, the research conducted two main filtering procedures. In the first filtering step, the researcher analysed the paper title and abstract for each 1285 paper. If the title and the abstract presented an explicit relation to the inclusion criteria, the paper was selected to be further analysed in detail. Then, the research applied full text read for each of the papers selected in the first step. The research checked each of the exclusion criteria and re-evaluation the relevance of the paper to the research. Each abstract and paper was analysed by the author of this thesis, following a specific predefined protocol. When there were some doubts about the paper's relevance to the research questions, the opinion of a member of the supervisory team was requested to minimise potential bias. Following the previous procedure, the filtering stage ended up with a total of 190 selected papers.

4.2.3 RESULTS

Analysing the literature has led to the identification of two categories of explainable models 1) those that are to explain any machine learning models (Model-agnostic); and 2) those that are designed for reverse-engineered specific machine learning models, thus cannot be used for other ML models. The research excludes the papers that only generate explanations for specific ML models. Model-agnostic explanation models are techniques that are designed to interpret the prediction of any machine learning model to generate some information from its prediction. The main purpose of such models is to provide extracted knowledge from the ML model, simplify the underlying logic and provide generalisability for other predictions. The results identify a list of five main explanation classes that are supported by current model-agnostic explanations. Five main explanation classes that emerged from the literature review are presented in Table 11.

TABLE 10 FIVE MAIN EXPLANATIONS CLASSES

Category of explainable methods	Definition	Examples
Explain the model (Global)	The explanation presents the weights of global features used in the model. This includes different visualisations of the feature's weights, such as approximating the model into interpretable global decision-tree or set of rules i.e., if-else.	(Henelius et al., 2014, Lou et al., 2013, Nguyen et al., 2016, Tolomei et al., 2017, Bastani et al., 2017, Johansson and Niklasson, 2009, Krishnan et al., 1999, Dash et al., 2018)
Explain a local prediction (Local)	The explanation describes the contribution of the local instance features to the model prediction. This also includes different visualisations such as pointing out a single part of the image, local decision tree and local if-else rules.	(Lundberg and Lee, 2017, Ribeiro et al., 2016b, Ribeiro et al., 2018)
Counterfactual explanation	This explanation class shows how the prediction will change concerning a change in the features' values. This also includes the changes in the prediction in the absence or presence of specific features.	(Friedman, 2001, Apley, 2016, Dhurandhar et al., 2018, Wachter et al., 2017)
Example-based	The explanation presents examples that are similar or have small differences from the current prediction.	(Bien and Tibshirani, 2011, Kim et al., 2014, Koh and Liang, 2017)
Confidence explanation	The explanation indicates the certainty values given a prediction.	(Subbaswamy and Saria, 2018, Gal and Ghahramani, 2016)

Explain the model. The global explanation model is developed to generate an explanation from a black-box model through an interpretable and explainable model. The generated explainable model is an approximation of the black-box model that explains the global behaviour of the model. This model should have similar accuracy and performance to the black-box model. Various papers in our literature review described novel techniques to solve the

global explanation problem and generate an explainable model derived from the black-box model. The analysis of such methods introduced three sub-classes of global explanation models.

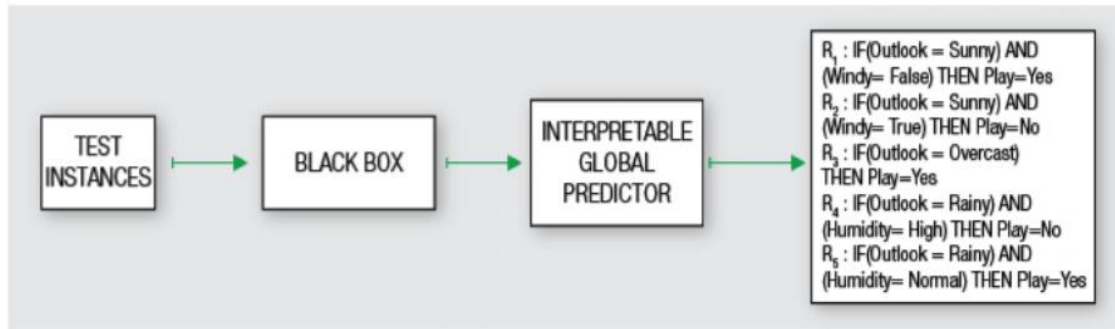


FIGURE 11 GLOBAL INTERPRETABLE EXAMPLE INTRODUCED BY (GUIDOTTI ET AL. 2018B)

- 1- Global feature importance. This category aims to find a group of data features that affects the performance and the prediction of the machine learning model. For instance, recent work presented a naïve Bayes classifier that can derive the dependencies between the data features attributes from any machine learning model (Henelius et al., 2014, Henelius et al., 2017). Their approach can generate an importance score for each feature and reveal the association between different data features. Lou et al. (2013), Nguyen et al. (2013) and Tolomei et al. (2017) developed an explainable technique to rank all possible data features that contributes to the model overall prediction. Furthermore, the work proposed by (Datta et al., 2016) presents an explanation model that measures the degree of influence that a data feature inputs to the overall prediction of the model.
- 2- Decision tree approximation. This category of contributions presents an explainable model that can generate a global decision tree given a black-box model. For example, the following contributions (Bastani et al., 2017, Johansson and Niklasson, 2009, Zhou and Hooker, 2016) proposed approaches based on model extraction techniques that approximate any machine learning model to a simple and explainable decision tree model. Krishnan et al. (1999) overcome the limitation of the complexity of such generated decision tree by controlling the size of the decision tree. Their approach used a genetic algorithm to propose decision trees with varying sizes. In the same direction, Thiagarajan et al. (2016) developed a TreeView approach to generate a human-friendly decision tree by an iterative rejection of an unlikely prediction label until the correct prediction appears.
- 3- Rule approximation. In the same way as the decision tree approximation, the explainable model generates a set of rules that explain the global reasoning of the model. Dash et al. (2018) and Aung et al. (2007) developed an explainable model that learns the boolean rules in either disjunctive normal form (OR-of-ANDs) or conjunctive normal forms of (AND-of-ORs). Similarly, the following contributions (Wei et al., 2019, Tan et al., 2018) proposed a generalised linear rule model that detects the interaction and the relationship between data

features in the forms of decision rules. Also, Zhou et al. (2003) developed REFNE, an interpretable model that can extract rules with strong generalisation ability or with high fidelity and conciseness. The other type of rule extraction is *if-then* form (Johansson et al., 2004, Quinlan, 1987). Figure 12 presents an example of the generated rules in both types of rules.

No.	Rule	
1	physician-fee-freeze \rightarrow democrat	(if (< TV0 17)
2	\rightarrow education-spending \rightarrow republican	(if (< TV2 36)
3	\rightarrow handicapped-infants \wedge adoption-of-the-budget-resolution \rightarrow republican	(if (> OI1 40) 82 72)
4	\rightarrow handicapped-infants \wedge aid-to-nicaraguan-contras \rightarrow republican	(if (< TV2 97) 83 97))
5	\rightarrow water-project-cost-sharing \wedge adoption-of-the-budget-resolution \rightarrow republican	(if (> OP1 216)
6	water-project-cost-sharing \wedge mx-missile \rightarrow republican	(if (< TV2 97) 85 112)
7	\rightarrow handicapped-infants \wedge \rightarrow superfund-right-to-sue \rightarrow democrat	(if (< TV2 40) 85 112)))
8	\rightarrow handicapped-infants \wedge \rightarrow mx-missile \rightarrow republican	
9	\rightarrow water-project-cost-sharing \wedge religious-groups-in-schools \rightarrow republican	

FIGURE 12 TWO EXAMPLES OF RULE EXTRACTION MODELS PRESENTED IN (ZHOU ET AL. 2003) AND JOHANSSON ET AL. (2004)

Explain a local prediction (Local). In the context of a single prediction or a recommendation, the explanation model can indicate reasons for specific predictions. Model agnostic derives the explanation from a local model that approximates the machine learning model well in a neighbour cluster of data points around a specific data point (Ribeiro et al., 2016b). The analysis of the literature identifies two main categories of explaining local prediction:

- 1- Local Feature Importance. The explanation shows how data features of a prediction contribute to the machine learning model prediction such as including parts of an image or text. Among the different contributions in this category, the results identified LIME (Ribeiro et al., 2016b) and all its variation (Mishra et al., 2017, Ribeiro et al., 2016a) as a novel technique that can explain any machine learning classifier by learning an interpretable model locally around the prediction. Also, Lundberg and Lee (2017) developed a novel technique called SHAP (SHapley Additive exPlanations) framework as a unified measure of feature importance that various methods approximate. Similarly, Zhou et al. (2003) developed a general method that perturbs all subsets of features to deal with the shortcomings of other existing feature importance explanation methods. Their approach was focused on considering the interaction between data features. Other contributions such as (Simonyan et al., 2013, Fong and Vedaldi, 2017, Dabkowski and Gal, 2017) proposed an image saliency method, which is applicable for differentiable image classifiers.
- 2- Local rules and trees. These techniques for model-agnostic explanations are designed to be plugged into any machine learning model to extract some information from its local prediction and present it as local rules and trees. For instance, the following contributions (Konig et al., 2008, Johansson et al., 2004) developed a rule extraction method (termed G-

REX) based on genetic programming. In line with rule extraction, (Ribeiro et al., 2018) presents a novel method (termed Anchors) with high-precision rules representing sufficient condition for a single prediction. Local decision trees are another example of an explanation model to explain a single prediction through local rule extraction. This approach was preferable for many researchers in the analysis as it has a human-friendly nature (Guidotti et al., 2018a, Krishnan and Wu, 2017)

Example-based explanation. Example-based explanation models select instances from the dataset to explain the behaviour of the black-box machine learning model. These models come to mimic the explainability behaviour between humans and can be effective for explaining complex connotations (Renkl, 2014, Renkl et al., 2009). Example-based explanations could potentially give the users some intuition about the black-box model that is complex to be understandable through other model-agnostic models. Table 12 presents a categorisation of the reviewed papers. The analysis reveals three categories of explaining black-box models through examples:

- 1- **Prototype:** The examples in this category are representative samples of instances from the dataset with the same record as the prediction (Bien and Tibshirani, 2011, Kim et al., 2016). Prototype methods seek a minimal selection of similar instances as a general performance and accuracy goal (Kim et al., 2014). Furthermore, the following contributions (Kim et al., 2016, Kanehira and Harada, 2019) emphasise that explaining through prototype can lead to over-generalisation or misunderstanding of the presented explanations. They argued that examples might be useful when the distribution of the training data is clean, i.e., prototypical examples represent the current recommendation. However, this case is rare in real-world scenarios. Therefore, to help human decision making, their approach was to select and classify examples in the dataset into good examples and bad examples. Good examples represent the model behaviour in high accuracy, whereas bad examples do not fit the model reasoning.
- 2- **Counterfactual.** This refers to the explainable model that explains the black-box behaviour providing similar instances to the prediction with small differences (Laugel et al., 2017, Mothilal et al., 2020). It also answers the question “What-if” an input changes through examples; example provided in Figure 13 by (Wachter et al., 2017). Some researchers used heuristics for generating counterfactual explanations by amending some input features (Martens and Provost, 2014, Chen et al., 2017).

The prediction for a woman with Pima heritage to be at risk of diabetes is 0.5.
 Other persons that have similar scores:

- A. If your 2-Hour serum insulin level was 154.3 like **Person 1**, you would have a score of 0.51.
- B. If your 2-Hour serum insulin level was 169.5 like **Person 2**, you would have a score of 0.51.

C. If your Plasma glucose concentration was 158.3 and your 2-Hour serum insulin level was 160.5 like **Person 3**, you would have a score of 0.51.

FIGURE 13 COUNTERFACTUAL EXAMPLE-BASED EXPLANATION EXAMPLE (WACHTER ET AL. 2017)

- 3- Influential example. This explains the model prediction based on training instances that are most responsible for influencing the prediction (Koh and Liang, 2017). Influence explainable models capture the idea of inspecting the black-box models through the lens of their training dataset. For instance, Goodfellow et al. (2014) identify that influence examples could be applied for various data science tasks such as understanding the model behaviour, debugging the black-box model and detecting errors. Similarly, the following contributions (Yuan et al., 2019, Dong et al., 2017, Szegedy et al., 2013) developed adversarial examples which are example-based methods with small, intentional feature perturbations that influence the black-box to false prediction. Although the literature provided theoretical foundations for the usefulness and effectiveness of influential examples on humans' decision-making, the research lacks user studies to understand the effect of influential examples on trust and calibrated trust.

Counterfactual explanation. This refers to explainable models that address the question of how the prediction would have been changed with a different set of input (Woodward, 1997). Counterfactual statements are usually taking the form: *Prediction P was made because the feature F has the value f_i . However, if F had the value f'_i , where other features had remained constant, Prediction P_i would have been returned* (Wachter et al., 2017). They are designed in a way to convey a minimal amount of information capable of amending a prediction. Researchers argue that counterfactual explanations are human-understandable explanations and they do not require the user to understand the underlying logic of the model (Arrieta et al., 2020). Designers and developers of such models often assume a clear and relation from recommended changes in feature values to actions in the real world (Barocas et al., 2020). However, in many cases such as medicine, this assumption will fail counterfactual, e.g., an explanation might ask the doctor to change the age of the patient. The analysis identifies two main categories of counterfactual explainable models:

- 1- Feature Influence. It refers to an explainable model that shows how a prediction could change regarding a change of a feature either in a static way (Apley, 2016, Friedman, 2001, Goldstein et al., 2015) or interactive way (Krause et al., 2016). Furthermore, it supports a localised inspection and feature tweak of a prediction to answer how and why a specific prediction is predicted.
- 2- Counterfactual features are techniques that aim to describe the features that will change the prediction when it is amended or deleted (Dhurandhar et al., 2018, Wachter et al., 2017). This method is argued to be efficient to support the user with feedback when the model

prediction is different from the desired prediction e.g. rejected loan application(Zhang et al., 2018).

Confidence explanations. It is an explanation class that shows the rationale of a given recommendation by presenting its certainty score. The confidence score can be generated from the machine learning models from two main sources (Gal and Ghahramani, 2016): model and data. Researchers generated confidence scores at the model level by computing the distributional differences during the model training stage (Gal and Ghahramani, 2016, Schulam and Saria, 2019). Whereas, data confidence scores can come from noisy, missing or predefined assumptions on the data (Josse et al., 2019). The common technique in the literature to assess the confidence is using Bayesian methods e.g. (Graves, 2011, Blundell et al., 2015). There has also been additional work on other techniques such as Dropout (Srivastava et al., 2014), tree-based density (Hooker, 2004) and simple heuristic using SoftMax (Hendrycks and Gimpel, 2016). Researchers argued that such confidence explanations can be used for trust calibration goals when the designer of the system wants to inform the user about the appropriate level of trust (Bussone et al., 2015, Helldin et al., 2013).

TABLE 11 THE CATEGORISATION OF THE REVIEWED PAPERS

Global explanations	Global feature importance	Ranking the data features.	(Lou et al., 2013, Nguyen et al., 2016, Tolomei et al., 2017)
		Dependencies between data features	(Henelius et al., 2014, Henelius et al., 2017)
		Influence Function	(Datta et al., 2016)
	Decision tree approximation	(Bastani et al., 2017, Johansson and Niklasson, 2009, Krishnan et al., 1999, Bastani et al., 2017, Johansson and Niklasson, 2009, Zhou and Hooker, 2016, Thiagarajan et al., 2016)	
	Rule	AND-OR rules	(Dash et al., 2018, Aung et al., 2007, Wei et al., 2019, Tan et al., 2018, Zhou et al., 2003)

	extraction	If-then rules	(Johansson et al., 2004, Quinlan, 1987)
Explain prediction	Local feature importance	(Ribeiro et al., 2016b, Lundberg and Lee, 2017, Simonyan et al., 2013, Fong and Vedaldi, 2017, Dabkowski and Gal, 2017, Zhou et al., 2003, Mishra et al., 2017, Ribeiro et al., 2016a)	
	Local rules and trees	(Guidotti et al., 2018a, Krishnan and Wu, 2017) (Ribeiro et al., 2018) (Konig et al., 2008, Johansson et al., 2004, Soares and Angelov, 2019)	
Example-based	Prototype	(Bien and Tibshirani, 2011, Kim et al., 2016) (Kim et al., 2014) (Kim et al., 2016, Kanehira and Harada, 2019)	
	Counterfactual example	(Wachter et al., 2017) (Martens and Provost, 2014, Chen et al., 2017) (Laugel et al., 2017, Mothilal et al., 2020)	
	Influential example	(Koh and Liang, 2017) (Goodfellow et al., 2014) (Yuan et al., 2019, Dong et al., 2017, Szegedy et al., 2013)	
Counterfactual	Feature Influence	(Woodward, 1997) (Apley, 2016, Friedman, 2001, Goldstein et al., 2015) (Krause et al., 2016)	
	Counterfactual features	(Wachter et al., 2017) (Dhurandhar et al., 2018, Wachter et al., 2017) (Zhang et al., 2018) (Krause et al., 2016) (Barocas et al., 2020)	
Confidence	(Zhang et al., 2020a, Bussone et al., 2015) (Gal and Ghahramani, 2016, Schulam and Saria, 2019) (Josse et al., 2019, Graves, 2011, Blundell et al., 2015) (Srivastava et al., 2014) (Hooker, 2004) (Hendrycks and Gimpel, 2016).		

4.2.4 FORMING USERS' QUESTIONS AND EXPLAINABLE CLASSES TAXONOMY- PRIOR EVALUATION.

In this stage, the research adopts the taxonomy provided by (Lim and Dey, 2010, Lim, 2012) as a reference model for users' questions. Their approach provided a systematic set of questions that end-users may ask to the intelligent system. While their taxonomy does not consider the current state-of-art explainable models and it is built on early explainable models in context-aware systems. This stage conducts a mapping between the current state-of-art model-agnostic explainable models and users' questions, as well as introduces new questions. We describe and situate the users' questions in the context of model-agnostic explainable models.

TABLE 12 MAPPING BETWEEN USERS' QUESTIONS AND EXPLAINABLE MODELS

		Ranking the	Why does the system think
--	--	-------------	---------------------------

Global explanations	Global feature importance	data features.	so? What are the main reasons for that?
		Dependencies between data features	What is the relation between features A and B?
		Influence Function	What is the effect of feature A?
	Decision tree approximation	How the system reaches that conclusion? What if? Why not? Why?	
	Rule extraction	AND-OR rules	How the system reaches that conclusion? What if? Why not? Why?
		If-then rules	How the system reaches that conclusion? What if? Why not? Why?
Explain prediction	Local feature importance	Why the system generates this prediction? What is the effect of a specific feature on local prediction?	
	Local rules and trees	Why the system generates this prediction? How does the system reach this prediction?	
Example-based	Prototype	With what example could this prediction have happened? What cases could result from the same prediction? What else is similar to this prediction?	
	Counterfactual example	What examples could a change A to this prediction result? What else (small changes)?	
	Influential example	What examples make this prediction happen? What examples could affect the prediction?	
	Feature Influence	What if feature A has the value B?	

Counterfactual	Counterfactual features	How to change the prediction? Why not?
Confidence	How certain is the system of this prediction? How accurate the system is?	

4.2.5 VALIDATION

The taxonomy has been validated by domain knowledge experts following a focus group approach. Participants were asked to evaluate and amend the emerged concepts, users' questions and the mapping between them. Table 15 presents the documents used by the expert during the study. Table 14 lists the session procedure and settings. These documents as well as the procedure were sent to the experts via email before the study. Before conducting the study, the participant information sheet, background questionnaire and the consent form were provided. All these documents can be found in (Appendix 1).

TABLE 13 THE STRUCTURE OF TAXONOMY EVALUATION STUDY

Stage No.	Name	Description	Note	Est time.
1	Prepare	The researcher will brief the participants about the study goals and structure.	-----	10 mins
2	Evaluate	The participants will be provided with a copy of the taxonomy to individually evaluate and make notes in the notes form.	Notes might include missing users' questions or concepts, structuring issues and probably refining suggestions	15 mins
3	Map	Each group will be provided with the same set of taxonomy concepts to start mapping the available users' questions into the explainable model with an option to add questions. This is meant to be a group activity [REF].	Participants will be informed to delete/add questions and re-map them as they think appropriate. Disagreements were expected to arise but resolved during the session.	15 mins
4	Discussion	Each group discussed the other groups' findings and highlighted the agreements and recommendations for further corrections.	Each group was given 10 mins	20 mins

TABLE 14 THE PROVIDED DOCUMENTS DURING THE FOCUS GROUP SESSION

Document No.	Documents	Description
1	Taxonomy structure	The first draft of the taxonomy that appeared from our literature review was provided in the email invitation.
2	Notes form	Each participant used this form to amend the taxonomy and make notes about the taxonomy structure.
3	Catalogue	The catalogue defines the relative concepts of the explainable machine learning models and their usage scenarios.
4	Users' questions	Potential users' questions with extra blank ones were provided during the session.

The focus group validation has introduced several changes to the taxonomy. Table 16 presents the final version of users' questions and explainable capabilities taxonomy. The main observations emerged from the focus group, and some important changes applied to the taxonomy are summarised below:

- A consensus on categorising the explainable models into five groups: Global explanations, local explanations, counterfactual explanations, example-based explanations and confidence score explanations.
- “*When*” questions should be added to the example-based explanations category. Example-based explanations relatively explain the model behaviour through a situation-based in which the “*When*” question could be the best. In other words, example-based explanations represent the situations and scenarios would a particular model output happen.
- The main reason for having this taxonomy is to provide a holistic picture of what questions could model-agnostic explanations answer having machine learning and HCI experts' perspectives. Participants mentioned that the taxonomy questions should be written to be generalised for any potential sub-questions and differentiate between any potential conflicts. For instance, partial questions could be derived from Why questions with the same answer, e.g., why the model is generating this output? and why does the model think so?. Participants suggested including the main family of a question without adding additional details to the question. For instance, instead of writing “*Why the AI is recommending specific output based on a given input*”, participants suggested adding the *Why* question.
- To operationalise the previous amendment to the taxonomy, participants suggested building another document that justifies users' question families.

They mentioned that generalising the taxonomy to fit different purposes will require further justification about users' questions and what they mean from the users' perspective. For this purpose, the researcher generated DOC 1, a supporting document that explains each of the users' questions family (See Table 17).

- For usability goal, participants suggested having the following format when writing the questions: (user question family) + (suffix to represent the answer) + (scope of the explanation i.e., Model or recommendation). For example, "Why" questions for local decision trees and global decision trees should have the suffix (trace-recommendation) or (trace-model). Such suffix would tell the taxonomy users the expected answers and the answer level.
- Participants agreed on five main suffixes to represent the answers to users' questions:
 - Influence: This is typically represented as a set of feature contributions to the AI output or weights of evidence.
 - Dependences: This information has typically represented a set of correlated features and their contribution to the AI output.
 - Trace: This information is typically represented as a set of triggered rules for a rule-based system.
 - Exemplar: This information is typically represented as a set of examples similar to the current AI output.
 - Uncertainty: This information informs users how uncertain the AI is of the output.
- Suffix in example-based explanations should refer to the example's types could the model generates. Therefore, three main concepts were added to justify "When" questions: *exemplar*, *exemplar with small changes* and *abstract exemplar*.

TABLE 15 FINAL TAXONOMY

Global explanations	Global feature importance	Ranking the data features	Why – Influence - Model
		Correlation between features	Why – Dependences – Model
	Decision tree approximations	How to – trace – Model What if – trace - Model Why – trace – Model Why not – trace – Model	

	Rule extraction	AND-OR rules	How – trace – Model What if – trace - Model Why – trace – Model Why not – trace – Model
		IF-ELSE rules	How – trace – Model What if – trace - Model Why – trace – Model Why not – trace – Model
Local Explanations	Local feature importance	Why – Influence – recommendation	
	Local Trees	How – trace – recommendation What if – trace - recommendation Why – trace – recommendation Why not – trace – recommendation	
Example-based	Prototype	When – exemplar – recommendation	
	Counterfactual	When – exemplar with small changes – recommendation	
	Influential	When – abstract exemplar – recommendation	
Counterfactual	Feature Influence	What if – Influence - recommendation	
	Counterfactual features	When – influence – recommendation Why – influence – recommendation How to – influence - recommendation	
Confidence	How accurate – uncertainty – recommendation		
	How accurate – uncertainty – model		

TABLE 16 USERS' QUESTIONS

User question	Definition
How, How to	This family of users' questions ask the system logic and how the system can reach a specific output.
Why	It refers to asking the AI about deriving its output value from the current input values.

Why not	It refers to asking the AI why not alternative output is not presented.
What	It refers to asking the AI about the main reasons that derived an output. This could be related to feature weights, overall logic or similar examples.
What if	It refers to asking the AI to anticipate or simulate user-set input values.
When	It refers to asking the AI under what circumstances or scenarios, or with what instance case would a particular output happen.
How accurate	It refers to asking the AI how certain the AI is of the proposed recommendation or prediction.

4.3 CHAPTER SUMMARY

In this chapter, the researcher analysed the literature to answer the main research question “*What can be explained given a black-box model?*”. Then, the research categorised different classes of model-agnostic explainable models and analysed which users’ questions can each explainable model answer. Researchers or XAI interface designers can use this taxonomy as a foundation and a reference point for eliciting users’ mental models and identifying appropriate explanation classes and models to fit the Human-AI collaborative decision-making tasks. The next chapter analyses the five explanation classes from users’ perspectives during a decision-making task and presents their requirements to enhance trust calibration.

5. CHAPTER 5: EXPLAINABILITY AND CALIBRATED TRUST: THE ROLE OF EXPLANATION CLASS

Machine learning has made rapid advances in safety-critical applications, such as traffic control, finance, and clinical decision support systems. With the criticality of decisions, they support and the potential consequences of following their recommendations, it also became critical to provide users with explanations to interpret machine learning models in general, and black-box models in particular. However, despite the agreement on explainability as a necessity, there is little evidence on how recent advances in eXplainable Artificial Intelligence literature (XAI) can be applied in collaborative decision-making tasks to contribute to the process of trust calibration effectively. This chapter presents an empirical study to evaluate four XAI classes for their impact on users' trust and trust calibration. It takes clinical decision support systems as a case study and adopts a within-subject design followed by semi-structured interviews. The researcher gave participants scenarios and XAI interfaces as a basis for decision-making and rating tasks. The study involved 41 medical practitioners who use clinical decision support systems frequently. Findings showed that users perceive the contribution of explanations to trust calibration differently according to the XAI class and to whether XAI design fits their job constraints and scope. This chapter also reveals additional requirements on how explanations shall be instantiated and designed to help a better trust calibration. Finally, the chapter builds on the findings and presents guidelines for designing explainable recommendations.

5.1 INTRODUCTION AND THEORETICAL BACKGROUND

Recent advances in machine learning have increased the adoption of human-AI collaborative decision-making tools in safety-critical applications such as medical systems (Bayati et al., 2014, Caruana et al., 2015) and criminal justice systems (Flores et al., 2016). Combining humans and AI in a collaborative decision-making task is expected to increase the quality of the decision-making (Green and Chen, 2019, Jacobs et al., 2021). However, recent studies showed that humans frequently make trust calibration mistakes by following incorrect recommendations or rejecting correct ones (Jacobs et al., 2021, Bussone et al., 2015, Zhang et al., 2020). Hence, calibrating users' trust has been identified as a key design goal for safe and effective Human-AI collaboration.

One approach to successful trust calibration is eXplainable AI (XAI) which refers to an AI component that explains AI recommendations to humans receiving them (Naiseh et al., 2021). Explanations can help the human operator understand the AI rationale and reasoning as well as decide when to accept an AI-based recommendation or reject it (Yang et al., 2020, Lai and Tan, 2019, Cai et al., 2019). XAI is also a critical factor when establishing liability and accountability for the final decisions. Two main streams emerged in the fast-growing XAI

research. The first suggests new interpretable machine learning models that are mathematically explainable and transparent, which can compete with black-box models' performance (Arrieta et al., 2020). On the other hand, model-agnostic approaches suggest interpreting any machine learning model, whether interpretable or black-box models, by focusing primarily on the input and the output of the machine learning model (Hohman et al., 2019). This approach is motivated by preserving the models' confidentiality, increasing the cost-efficiency of generating the explanation, and increasing its usability (Feng and Boyd-Graber, 2019, Zhang et al., 2020, Sokol and Flach, 2019). Different model-agnostic methods may generate explanations with distinct explanation output, but they may vary in their performance, fidelity and completeness of the underlying AI model (Arrieta et al., 2020). This paper refers to the family of model agnostic methods that generate distinct explanation output as an XAI class. XAI class can also answer similar users' questions (Carvalho et al., 2019, Liao et al., 2020).

While many studies emphasised the requisite of explainability to support trust calibration (Lai and Tan, 2019, Naiseh et al., 2021, Zhang et al., 2020), several recent studies found little evidence that the XAI class has a significant impact on trust calibration (Jacobs et al., 2021). Many reasons could have contributed to this effect. One is the complex nature of trust. According to the human-computer trust model, trust is formed in two dimensions: cognition-based trust and affect-based trust. Cognition-based trust is based on humans' intellectual perceptions of AI reasoning, whereas affect-based trust is based on humans' emotional responses to AI systems. In fact, several studies suggest that XAI class (Wang et al., 2019), the level of transparency (Kulesza et al., 2013) and explanation framing (Narayanan et al., 2018) are all factors that can affect both humans' cognition and affect. Another reason for the effect of XAI class on trust calibration is the nature of humans' cognitive biases (Naiseh et al., 2021). For instance, under-trust may be resulted from anchoring bias when humans look at only salient features of XAI class and thus judge the quality of the XAI class to be untrustworthy. Similarly, over-trust may result from confirmation bias when humans favour XAI class that is consistent in its output with their beliefs and initial hypothesis.

Although, the concept of XAI class has been studied for human-AI collaborative decision-making (Jacobs et al., 2021, Zhang et al., 2020), both to prevent under-trust or over-trust. Recent advances in XAI literature have not been well-covered and evaluated in trust calibration in recent empirical studies. In one related research, Dodge et al. (2019) examined the impact of different XAI classes on calibrating AI-based decision-making tools' perceived fairness. They aimed to determine which XAI class can help participants identify fair AI decisions. They found that Local explanations seem to be more useful than Global explanations when used to explain an unfair model's decisions, thus more effectively calibrating people's fairness judgment. Another study by Zhang et al. (2020) also studied the effect of presenting AI confidence scores and local explanations on users' truth calibration. They found that presenting confidence scores

to end-users can trust calibration goals. Another limitation of recent studies that explored the impact of XAI class on trust calibration is that they frequently limited their studies on approaching participants that are unfamiliar with the Human-AI task. Although they boosted their participants' knowledge and familiarity of the Human-AI task by introducing a training task, we argue that measuring trust calibration and its relation to XAI class may require task expert users to observe the effect during a lived experience and on a fine-grained level. Different from earlier work, we explore the impact of four XAI classes (Local, Example-based, Counterfactual and Global explanations) on trust calibration during Human-AI collaborative decision-making tasks. We also study this effect using a clinical decision support system with experts from the medical domain.

In this chapter, we endorse the same postulate (Zhang et al., 2020) that calibrating user trust requires different research from the one focusing on increasing users' trust in AI. Inspiring trust can be done without necessarily improving users' mental models and beliefs of the true AI capability and limitations. On the other hand, calibrated trust may need extra effort from end-users to understand the AI system reasoning (Naiseh et al., 2021). We study the effect of four distinct XAI classes (Local, Example-based, Counterfactual and Global explanations) on trust calibration during a collaborative Human-AI task. We hypothesise that different XAI classes can affect trust calibration differently when using an AI-based decision support system. Our research method is based on a within-subject study followed by semi-structured interviews. The study is supported by scenarios, mock-up interfaces and rating and decision-making exercises. Our application area is medical decision-making, where users can be doctors and pharmacists, representing a high stakes application domain. Taking clinical decision support systems as an exemplar and focusing on the four state-of-the-art model-agnostic XAI classes, our contribution has two parts:

- An evaluation of how explanations belonging to the different XAI classes affect users' trust calibration during the Human-AI collaborative decision-making task.
- An elicitation of main requirements of users from Explainable AI to help enhance trust calibration processes.

The findings of this study are intended to provide a richer understanding of the main needs of users from explanations in their different classes. We also aim at broadening discussions on explainable AI for collaborative decision-making and paving the way for more research on how to customise and contextualise explanations so that they fit users' needs and expectations of each of their different classes.

5.2 RESEARCH METHOD

This chapter investigates the effect of XAI class on users' trust calibration during a collaborative decision-making task. Through both quantitative and qualitative research, we aim to answer the following questions:

- RQ1. How do different XAI classes affect users' trust calibration?
 - RQ1.1 Are XAI classes seen differently in their capability to affect (increase or decrease) users' trust?
 - RQ1.2 Are XAI classes seen different in enabling trust calibration? i.e., Are different XAI classes seen different in affecting the performance of the Human-AI team?
- RQ2. What are users' requirements to support trust calibration during the Human-AI task?

5.2.1 HUMAN-AI TASK DESCRIPTION

Screening prescription is a process that practitioners in a clinic follow to ensure that a prescription is prescribed for its clinical purpose and fit the patient profile and history. We chose the use case of screening prescription as it reflects an everyday decision-making task performed collaboratively between the human and the AI. The main workflow of the AI-based prescribing system is shown in Figure 14.

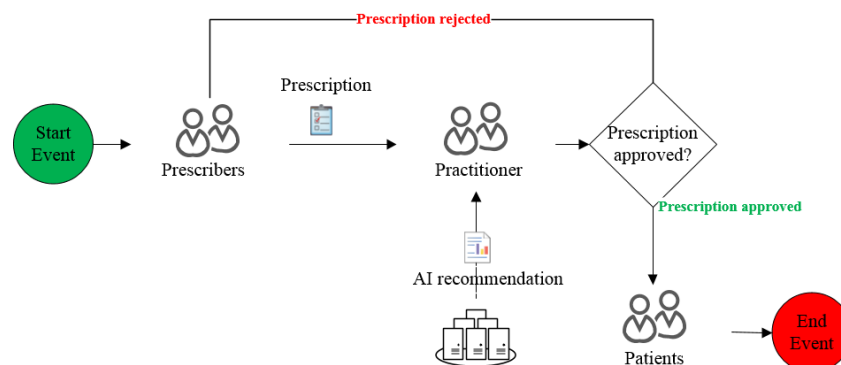


Figure 14 Workflow for prescription screening aided by AI-based decision-making tool.

To help our investigation, we designed an AI-supported decision-making mock-up that classifies prescriptions into confirmed or rejected. We designed the mock-up based on templates and interfaces familiar to our participants in their everyday decision-making tasks (See Figure 15).

Patient: Jack (Mr) Gender: Male, Born: 11-June-1969 (48Y) SAP Number: markte st007

Palliative Gemcitabine and Paclitaxel Frequency: 21 days

Height: 165 Weight: 74 Surface Area: 1.81 Performance Status: 0

☒ Add drug(s) ☐ Re-schedule ☐ Cancel appointment ☒ Add cycle(s) ☒ Add course ☐ Remember Cycle

Prescribed - Approved by AI tool.

Prescription Pathology Results Toxicities Notes(1) Documents(1) Allergies Diagnosis History

Time	drug	Dose	Administration	Frequency	Route	Duration
Day 1 (14/02/2020) Chemotherapy Day Unit						
T=30mins	Chlorpheniramine	10 mg			IV Bolus	1 minutes
T=1hr	Metoclopramide	10 mg		Three times a day	Oral	3 days
T=0	Sodium Chloride 0.9%	318mg	volume depend upon requirement	IV flush		

☒ Day 1 ☒ Day 8

Figure 15 A sample of prescribing system interface supported with AI recommendations.

Our mock-ups mimic a web-based tool and are meant to simulate user experience when working on an actual system. As the practitioner clicks on a prescription, the tool shows the patient profile, the recommendation from the AI-supported decision-making tool (accepted or rejected) and an explanation. The explanation can help the practitioner understand the AI rationale of why the prescription should be accepted or rejected. All material used in this study can be found in **Appendix 2**.

5.2.2 EXPLANATION CLASSES

We conducted a literature review in model-agnostic interpretable machine learning to categorise XAI methods into classes based on their distinct informational output. The full taxonomy resulting from our literature review can be found in Appendix 1. These XAI classes were also used in our mock-up tool:

- **Local explanations:** The explanation justifies the AI reasoning at the recommendation level; this can be done either by quantifying the contribution value for each input data feature to the recommendation (Ribeiro et al., 2016) or generating local rules or decision trees of a recommendation (Guidotti et al., 2018).
- **Example-based:** Given the AI-based recommendation, the AI justifies its decision by providing examples from that dataset with similar characteristics (Cai et al., 2019). For example, AI suggests rejecting this prescription because patient A history was similar to patient B.
- **Counterfactual:** Given the AI-based recommendation, the AI answers the users' questions "what-if" to observe the effect of a modified data feature on the recommendation (Sokol and Flach, 2019). For instance, the AI suggests rejecting the

prescription because Platelet Count= 60; however, if the Platelet Count were ≥ 75 , the prescription would have been confirmed.

- Global explanations: The explanation of an AI-based recommendation attempts at explaining the overall logic of the black-box model (Henelius et al., 2014, Wu et al., 2020). This includes presenting the weights of different data features as decision trees, rules, or ranking styles.

5.2.3 STUDY DESIGN

The study was a within-subject design followed by follow-up interviews. We manipulated the XAI class (No explanation, Local, Example-based, Counterfactual, Global explanations) and the recommendation outcome (correct and incorrect recommendations). As a result, we had ten different conditions. Each XAI class appeared in two conditions: correct AI-based recommendations and incorrect ones. We developed ten patient scenarios and interfaces to cover the study conditions. For each participant, XAI classes were randomly assigned to patient scenarios to eliminate the carryover effect (Louthrenoo et al., 2007) and the potential effect of accidental bias, i.e., having an XAI class with a specific patient scenario. We choose to manipulate the recommendation outcome to simulate a diversity of conditions that the practitioners could face in the real-world scenarios where trust calibration errors could happen, e.g., imperfect AI due to the dynamic nature of the application. Each participant completed ten different Human-AI tasks. We used a random sequence of completing ten human-AI tasks during the study. This controlled the learning effect of participants expecting a correct or incorrect recommendation and an XAI class during the study. However, we ensured that the first Human-AI task included a correct recommendation to avoid losing participants' trust in an early stage of the study (Lee and See, 2004, Marshall, 2003).

The patient scenarios presented to our participants were hypothetical scenarios designed with collaboration with a medical oncologist, i.e., there was no actual AI model. We designed the scenarios to be clear and challenging, and not trivial so that recommendations and explanations and trust calibration were all substantial processes. This ultimately helped to put our participants in a realistic setting: exposing them to an imperfect AI-based recommendation and its explanations where trust calibration is needed and where errors in that process are possible. We validated the material with a medical oncologist focusing on the border cases that need an investigation from the participants in the actual study. We tested the material and activities with two participants and refined them to optimise their fulfilment of these criteria.

5.2.4 STUDY PROCEDURE AND DATA COLLECTION

To answer our research questions, we conducted a study of two phases (Quantitative and Qualitative). We also provide an overview of the study phases in Figure 16.

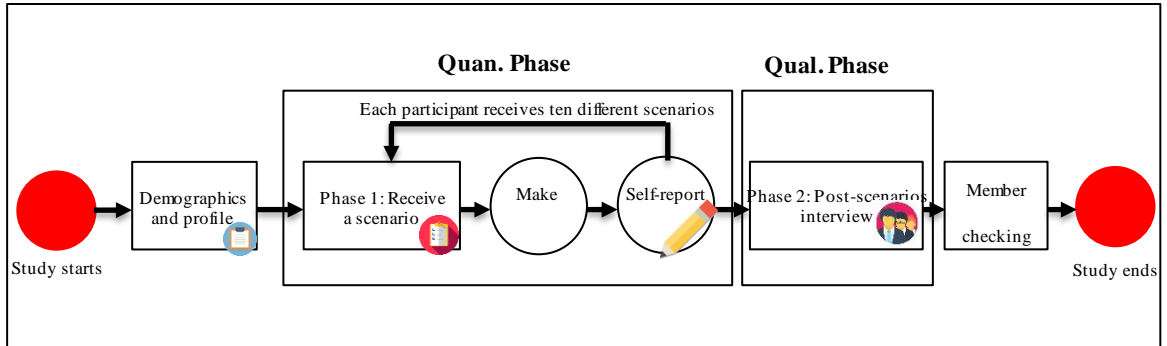


Figure 16 Study workflow

5.2.4.1 QUANTITATIVE PHASE

This phase was meant to answer RQ1. The phase involved 41 medical practitioners who have experience in screening prescriptions (Doctors and pharmacists). First, the participants were briefed about the study through a participant information sheet. They were then asked to sign a consent form. Participants were also asked several questions about themselves, such as their experience. Personal attributes, including skills and years of experience diversity, should help in covering different types of issues. For example, novice medical experts might raise more questions than those who have more experience in the task since they faced similar cases in the past and can cross-check with other sources. Although we recognise that personal differences, e.g., in agreeableness and conscientiousness (Barrick and Mount, 1991), can play a role, this was beyond the scope of our study. Then, each participant had to complete ten different screening prescription Human-AI tasks. Participants were asked to make decisions considering the patient profile, the recommendation and the explanations and whether to follow the AI-based recommendation if they see it as correct or reject it if they see it as incorrect. Each participant spent 15-20 minutes completing this stage. The study workflow is described in Figure 3. After completing a Human-AI task, participants were asked to rate their trust in the AI. Each of the participants completed ten Human-AI tasks, which resulted in 410 completed ratings. The following sections describe our measures to answer RQ1.

Trust calibration measurements.

Trust calibration during a Human-AI collaborative decision-making task can be measured by self-reporting trust (Yang et al., 2020) and behavioural indicators (Zhang et al., 2020, Wang et al., 2016, Bussone et al., 2015). We used both self-reporting trust and trust calibration behavioural indicators to answer our RQ1. Self-reporting trust measures were used to observe whether an XAI class was perceived differently in its capability to affect users' trust (RQ1.1).

On the other hand, behavioural indicators were followed to measure whether a particular XAI class can contribute to a successful trust calibration (RQ1.2).

Self-reporting measures. To measure the contribution of XAI in affecting (increasing or decreasing) users' trust, we followed Madsen and Gregor (2000) conceptualisation of human-computer trust into two main components: cognition-based components and affect-based components. The main difference in their effect is that cognition-based trust is crucial for establishing appropriate trust, whereas affect-based trust is developed as the relation continues (Nah and Davis, 2002). Furthermore, previous research showed that in critical decision-making scenarios, it is highly likely that cognition-based trust components significantly affect trust calibration (McAllister, 1995, Ng and Chua, 2006). Therefore, measuring affect-based components was not achievable during our study, and it needs more longitudinal and observational studies. Participants were asked to rate their trust at the end of each Human-AI task. Participants rated their trust based on cognition-based components: perceived understandability perceived reliability and perceived technical competence. Following (Madsen and Gregor, 2000), we used five points Likert scale to measure the effect of each XAI class on users' trust. The rating ranged from a 1 (Completely Disagree) to a 5 (Completely agree) Likert Scale. Accordingly, after each of the ten Human-AI tasks, participants were asked to rate their agreement with three statements to answer RQ1.1:

- Perceived understandability: *"I will use the AI because I can understand how the system behaves"*.
- Perceived reliability: *"I can rely on the AI in the task properly"*.
- Perceived technical competence: *"The AI has sound knowledge and accurate about this recommendation"*.

Behavioural indicators. We also collected trust calibration behavioural indicators for each XAI class condition so that we also answer RQ1.2. We looked at trust calibration behavioural indicators introduced in similar studies (Zhang et al., 2020, Wang et al., 2016, Bussone et al., 2015) and used three behavioural indicators to measure trust calibration:

- Agreement percentage. The percentage of Human-AI tasks in which participants followed the AI recommendation.
- Switch percentage. The percentage of Human-AI tasks in which the participants disagreed with the AI recommendations.
- Human-AI performance. The percentage of Human-AI tasks when following AI recommendations or switching from AI recommendations led to a correct decision.

5.2.4.2 QUALITATIVE PHASE

The second phase of our study was to answer RQ2. We conducted semi-structured interviews following the guidelines stated in (Oates, 2005). We used the interviewing method to delve into the details and understand the reasoning process and issues faced during the Human-AI collaborative decision-making process (Ericsson and Simon, 1984). We asked our practitioners to elaborate on their experience during their interaction with our mock-up tool to gain as many insights as possible. The interviewer discussed the benefits and the drawbacks of each XAI class in calibrating their trust and experience during the decision-making process. Also, the interviewer asked the participants to express their concerns and difficulties to make an informed decision, given explanations belonging to four XAI classes. 16 participants were able to participate in the follow-up interviews. No more participants were approached because, during the data analysis, themes and codes resulting from the analysis became eventually repetitive. We followed the principles of reaching the saturation point in the qualitative data collection introduced by (Faulkner and Trotter, 2017). This was a reasonable assurance that further qualitative data collection would introduce similar results and would confirm the existing themes.

5.2.5 PARTICIPANTS

Our sample consisted of 41 medical practitioners coming from three different organisations in the UK. All participants were medical experts recruited through email based on their experience in the case study. Details about the population are provided in Table 18.

Variable	Value	N=41	%
Age	20-30	22	54%
	30-40	11	28%
	40-50	8	18%
Gender	Male	26	65%
	Female	15	35%
Role	Doctors	26	62%
	Pharmacists	15	38%
Diagnosis Experience	<5	15	35%
	5-10	12	30%
	10-15	9	21%
	>15	5	14%

Table 17 Population details

5.2.6 DATA ANALYSIS

Two sets of data were collected and used to answer our research questions in this study. The first reflected the trust measurements (self-reporting and behavioural indicators). The second consisted of transcripts of the audio files of the follow-up interviews. We used several statistical

tests such as repeated One-way ANOVA to answer our first research question. For qualitative data, we followed content analysis method with the support of the Nvivo tool. To increase the trustworthiness of our qualitative analysis, we applied a member checking approach and reviewed the analysis of their interviews with them. Member-checking aims to review the analysis report by study participants to ensure that data, interpretations and themes are valid and applicable (Birt et al., 2016). Member checking technique was applied with three participants to validate the analysis and clarify the analysed data where clarification was needed. Each step of the analysis included several meetings among the authors to ensure the correct interpretation of each category and the evidence that supports them. These meetings led to splitting, modifying, discard or adding categories to ensure that all responses and their contexts were well represented.

5.3 FINDINGS

In this section, we report on our analysis results answering RQ1 that concerns whether XAI class differs in influencing user trust and trust calibration during a Human-AI collaborative decision-making task. In total, 41 medical experts completed 410 Human-AI tasks.

5.3.1 EXPLANATION CLASSES IMPACT ON USER TRUST (RQ1.1)

Each participant rated their trust (perceived understandability, perceived reliability, and perceived technical competence) during the interaction with different XAI classes. The descriptive statistics of participants ratings are shown in Figure 17. As a general observation, three trust components were seen differently by our participants, i.e., they were not mutually dependent. For instance, No explanation scenario was not perceived to be understandable but it has a higher rating in terms of perceived reliability and perceived technical competence. This means that trust indeed is not only about one dimension, and XAI interface design may need to consider each of its components when supporting appropriate trust of the AI. In the following section, we present our results based on each dimension of trust.

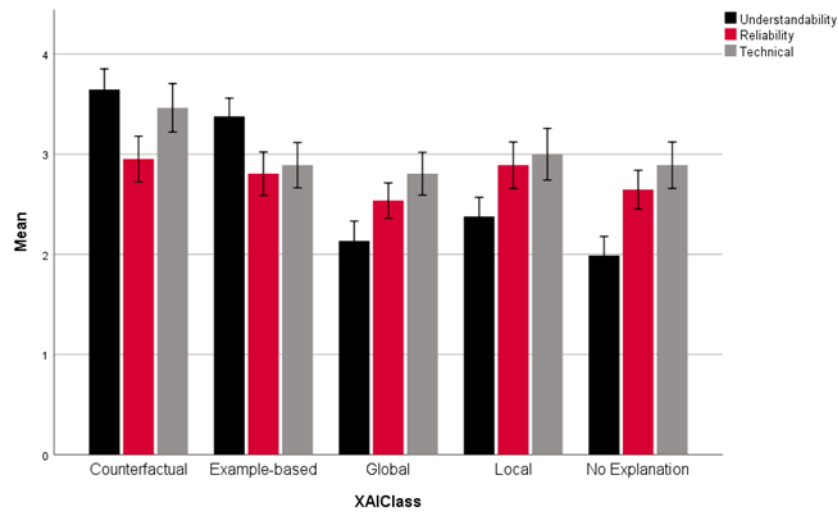


Figure 17 Mean cognition-based trust components rating per explanation class. Explanation cognition-based trust ratings range from Strongly Disagree (a rating of 1) to Strongly Agree (a rating of 5).

Perceived understandability. We applied a repeated measure ANOVA to compare the mean of participants understandability of the AI between different XAI classes conditions (No explanation, Global, Local, Counterfactual and Example-based). Our results show that participants rated their perceived understandability significantly different between the XAI classes conditions [$F(4,324) = 63.483, p < 0.001$]. Post hoc comparisons using the Tukey honestly significant difference (HSD) test indicated that the mean perceived understandability score in Example-based scenarios ($M = 3.383, SD = 0.830$) and Counterfactual scenarios ($M = 3.646, SD = 0.935$) was significantly higher than in No explanation ($M = 1.988, SD = 0.868$) Local ($M = 2.357, SD = 0.873$) and Global ($M = 2.150, SD = 0.901$) scenarios.

Perceived reliability. As shown in Figure 17, clearly, participants rating of their perceived reliability was steady across multiple XAI classes conditions [No explanation ($M = 2.646, SD = 0.880$), Global ($M = 2.537, SD = 0.804$), Local ($M = 2.890, SD = 1.054$) Example-based scenarios ($M = 2.805, SD = 0.987$) and Counterfactual scenarios ($M = 2.951, SD = 1.041$)]. Repeated measures ANOVA confirmed this observation and showed no significant difference in participants perceived reliability of the AI between different XAI classes scenarios. We further elaborate on the potential reasons for these indifferences in Section 5.3, where we report on the issues and needs expressed by the participants considering their perceived reliability of the AI. We also discuss the implications of this finding in Section 5.4.

Perceived technical competence. Results show that the XAI has a significant effect on the users' perceived technical competence of the AI [$F(4,324) = 4.815, p < 0.001$]. Post hoc comparisons using the Tukey honestly significant difference (HSD) test indicated that the mean perceived technical competency score in Counterfactual scenarios ($M = 3.463, SD = 0.1.102$) was significantly higher than in No explanation ($M = 2.890, SD = 1.054$) Local ($M = 3.00, SD = 1.176$), Global ($M = 2.805, SD = 0.974$) and Example-based scenarios ($M = 2.890, SD = 1.030$).

Our analysis of the interviews data had provided an explanation of that findings when participants mentioned that Counterfactual explanations provided meaningful knowledge for them to conclude “*Well, I can see how the AI is reasoning about this case, I would say yes the AI is reasonable*” [P12]. Although participants perceived Example-based explanations as an understandable explanation, they did not rate Example-based explanations as significantly competent. P5 commented in that context, “*Examples are beneficial and similar of what we do in the clinic, but it is not a proper explanation ... I mean it could be supportive to other explanations ... I would expect more casual or correlation relationship between the patient and the AI decision*”. In summary, participants considered explanations as competent when the explanation was understandable and showed the rationale behind the specific recommendation and those providing casual patterns at the recommendation level.

5.3.2 EXPLANATION CLASSES IMPACT ON TRUST CALIBRATION PROCESS (RQ1.2)

To answer RQ1.2, we looked at three behavioural indicators to examine the effect of XAI class on trust calibration, i.e., whether different XAI classes affected participants collaborative decision-making.

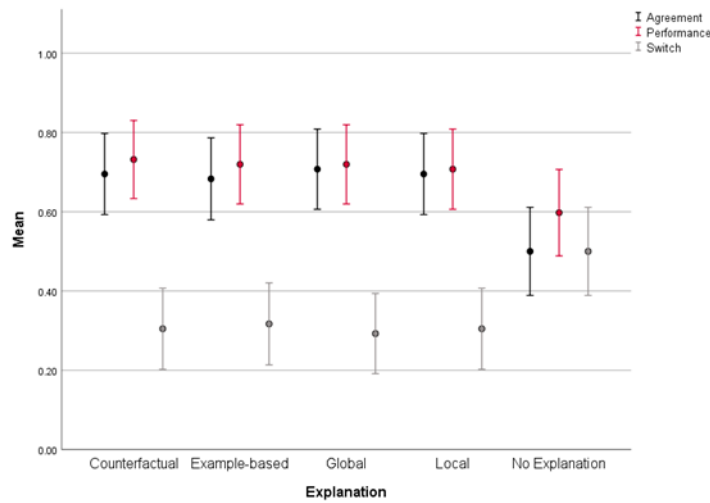


FIGURE 18 MEAN OF TRUST CALIBRATION BEHAVIOURAL INDICATORS

Different XAI classes improved the overall Human-AI performance. We plot the comparison of participants overall performance, average agreement and average switch between XAI classes conditions in Figure 18. Visually, it is clear that when participants were not provided with an explanation, they were more likely to make mistakes and explanations slightly improved the overall collaborative decision-making task. Friedman test confirmed this observation and shows that XAI class significantly affects Human-AI performance [$\chi^2(4) = 19.524, p = 0.001$]. Post hoc comparisons using the Wilcoxon signed-rank test indicated that Human-AI performance in Counterfactual ($M = 0.732, SD = 0.446$) Example-based ($M = 0.720, SD = 0.452$), Local explanations. ($M = 0.707, SD = 0.458$) and Global ($M = 0.720, SD = 0.452$) were significantly

different than No explanation ($M = 0.598$, $SD = 0.793$) scenarios. Our results appeared to be consistent with Lai and Tan (2019) conclusions, which showed that pairing AI recommendations with explanations can improve the Human-AI team.

All XAI classes increased participants agreement with AI recommendations. It is clear from Figure 18 that participants agreed more with the AI recommendations when explanations with their different classes were provided. Friedman test confirmed this observation and show that XAI class has a significant effect on Human-AI performance [$\chi^2(4) = 57.444$, $p = 0.001$]. Post hoc comparisons using the Wilcoxon signed-rank test indicated that Agreement percentage in Counterfactual ($M = 0.695$, $SD = 0.463$) Example-based ($M = 0.683$, $SD = 0.468$) Local ($M = 0.695$, $SD = 0.463$) and Global ($M = 0.707$, $SD = 0.458$) were significantly different than No explanation ($M = 0.50$, $SD = 0.503$) scenarios. Our findings seem to contradict Zhang et al. (2020) findings when they found that there is no significant effect between users' agreement in three conditions: No explanation, Uncertainty score and Local explanation. However, a closer look at their findings shows that presenting uncertainty scores to participants in their settings could interpret such differences. In other words, the explanation alone with its different classes could have increased our participants' reliance on the AI, and the uncertainty score could moderate the over-reliance effect. Our results are consistent with previous research (Bussone et al., 2015), which showed that presenting explanations to end-users could increase participants over-reliance and facilitate confirmation bias.

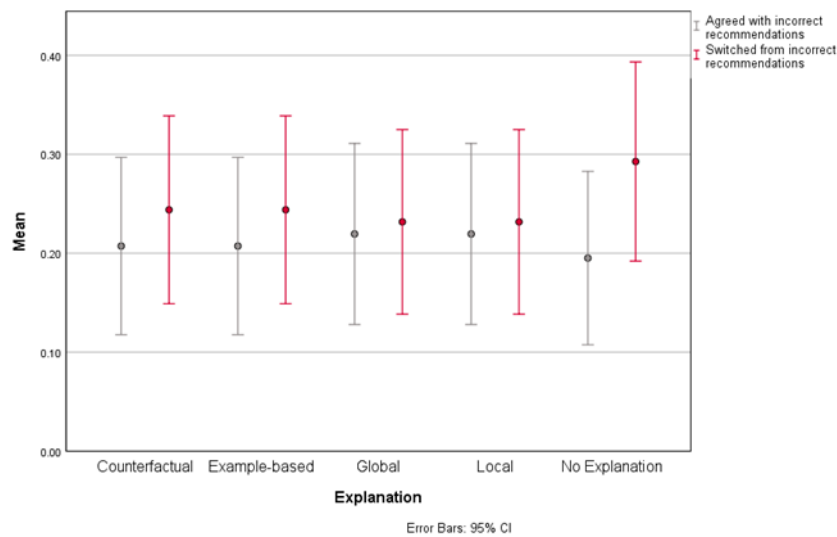


FIGURE 19 SWITCH PERCENTAGE AND AGREEMENT PERCENTAGE FOR INCORRECT RECOMMENDATIONS ACROSS DIFFERENT XAI CLASSES

XAI classes did not help participants recognise incorrect recommendations. Although XAI classes significantly improved the overall Human-AI team performance, the data suggests that XAI classes did not help participants recognise the incorrect recommendations compared to No explanation scenarios. The average of mistakes during the incorrect recommendation scenarios, i.e., agree with incorrect recommendations, made by participants during XAI classes conditions

and No explanation condition was not significantly different [$\chi^2(4) = 5.640$, $p = 0.231$]. Figure 19 compares participants responses during incorrect recommendations scenarios (Agreeing and switching from incorrect recommendations). It is clear that when the AI recommendation was incorrect participants struggled to decide whether to follow or reject AI recommendation across all XAI conditions.

We summarise the results from the quantitative analysis in three main points.

- Example-based and Counterfactual explanations had a higher users' perceived understandability than Global, Local and No explanations scenarios and helped our participants to reason about the AI recommendation.
- Users perceived technical competence seemed to be affected by explanation understandability and to provide casual patterns at the recommendation level.
- Explanations with its different classes increased the overall performance of the Human-AI team, however, we did not find a significant difference in Human-AI performance when facing incorrect recommendations.

5.3.3 USER REQUIREMENTS FOR ENHANCED TRUST CALIBRATION

In this section, we discuss our analysis results answering RQ2 concerning the user requirements from XAI interfaces in their different classes so that they enhance trust calibration processes. Upon completing the rating sheets, participants were interviewed to discuss with them the main issues they faced during their decision-making. Our focus is to elicit users' needs and requirements from XAI interfaces during their Human-AI tasks. Table 19 summarises our interview analysis results.

TABLE 18 TABLE 2 ISSUES AND NEEDS APPLICABILITY TO EXPLANATION CLASSES; (X) APPLIES (-) DOES NOT APPLY

Issues and Needs	Local	Example	Counterfactual	Global
Task-centred explanation	x	x	x	x
Usability	x	x	x	x
Assurances	-	-	-	-
Guidance and attention	x	-	-	x
Tailoring	x	x	x	x
Multi-step explainability	x	x	x	x

Guidance. Participants commented that they needed guidance to interpret Local and Global XAI classes. For instance, P9 mentioned, *“I think it is unfair for AI to explain this way because it just does whatever it was designed to explain for, so it does not give us to see the big picture ... I would like to know what it means to have a patient age with 35% influence on the AI*

decision? and how this could be interpreted for this patient?''. One interpretation of such comments is that Local and Global explanations may require previous technical knowledge to interpret them (Kaur et al., 2020). Our results also revealed that participants misinterpreted these explanations when the interviewer reviewed the explanations with them. For example, P8 commented on the Global explanation encountered during the study, *“I saw that blood test is the influential factor, and I was wondering we should screen prescriptions on that factor only?”*. Also, P12 commented on the Local explanation presented during the study: *“I feel this could be biased in some way, so that means the majority of the decisions will be made based on the tumour size”*. Participants’ interpretations of the potential bias and selectiveness in the explanation could be justified because of participants’ unfamiliarity with AI explanations even though they were given a presentation and examples before the study. Furthermore, our results showed that participants who gave low understandability and reliability ratings to Local and Global XAI classes mentioned that the design might need more attention and guidance to help them in interpreting the explanations, e.g., P3 responded, *“I think a tutorial on how to use these explanations would be beneficial”*. The previous comments signify that unfamiliarity with some XAI classes can be a significant issue that may decrease users’ trust in the AI and potentially lead to trust calibration mistakes. Our results are consistent with previous research that showed XAI methods are mainly designed and used by data scientists and provide little value to other end-users (Kaur et al., 2020, Ras et al., 2018).

Usability. Although participants identified the usefulness of the explanations to understand the AI rationale, some mentioned that they would not use these explanations in everyday scenarios. One reason for that was related to their concerns of fitting explanations in their workflow, mainly due to the need to process too much information, e.g., P6 stated, *“.... sometimes we are so busy; I won’t have that time to validate the AI through its explanation; in my opinion, a simple explanation targeting main patient issues would be enough with an option to investigate more when needed”*. In addition, participants felt that explanations could be a burden in their workflow and might cost extra cognitive efforts, e.g., P12 mentioned, *“Does this mean that I have to look at all these factors each time I make a decision?”*. Participants suggested that explanations could be better fitted when presented in textual formats as narratives or templates to reduce such a cognitive load and task impediment. For example, P13 wanted to have a generated narrative that summarises multiple XAI classes. In HCI, this happens when the tool's design leads to a mismatch between the user mental model and the conceptual mental model of the system (Carroll and Olson, 1988). Participants also mentioned that explanations sometimes contained redundant information, and they recommended different ways to make their trust calibration require fewer efforts. For instance, P9 asked to customise the number of data features in Local explanations, *“The average pharmacist does not need to see all these factors that the AI is considering, some of them are just simple rules”*. Our observations are aligned with recent studies that showed that long and redundant explanations made participants skip

them (Naiseh et al., 2021) and decreased participants' satisfaction with the explanation (Narayanan et al., 2018). In summary, explanations easiness of use and their modalities could be critical to engage users with them when participants time constraints and the difficulty of the task are primary issues. Also, limited time and unaffordable need for cognition (Petty and Cacioppo, 1986) can be both seen as a contextual disability for people who work on pressuring domains. Future work needs to consider the trade-off between effectiveness and usability of the explanation to optimise the Human-AI team performance, e.g., adaptive and personalised user interfaces could be a potential solution direction (Naiseh et al., 2020).

Task-centred explanation needs. Participants described that all XAI classes did not consider the task needs and constraints, and they were expecting an explanation of the context before being presented the explanation itself. Participants who identified these needs provided low technical competency and reliability rates. One example of task constraints needs was encountered in the Counterfactual explanation scenarios, where the explanation only provided information about changes that could be made to an AI recommendation to change the decision. In our case study, Counterfactual explanation scenario explained the recommendation by modifying the value of the patient blood test but also failed to specify that this modification could trigger risk or require another change in another data feature. Participants raised a concern regarding potential risks behind the explanation as it did not meet their task constraints. For example, in Counterfactual explanation scenarios, explaining the recommendation through a change in a data feature value without identifying the correlation with other data features is perceived as a risky explanation in their decision-making process. This means that making hypothetical scenarios without considering the domain and task constraints, such as the medical domain and patient cases, can result in unrealistic cases; cases that can be incoherent in the values of their variables and leading to additional explanation needs beyond the explaining recommended decision. The main reason for this drawback could be that the development of such an explanation method is to help data scientists debug the ML model (Kaur et al., 2020, Ras et al., 2018), where the context of the task for such explanations in real-world scenarios is still a new area to discover. In psychology, such explanations with a lack of consistency and context are usually ignored by people (Keil, 2006). Overall, participants wanted explanations that are reflective of their task, i.e., task-centred explanations, for these explanations to be meaningful and reliable. In other words, explaining the logic of the algorithm shall be done in a way that is tightly coupled with the subject, i.e. the task for which the recommendation and the explanation are given. We recognise here that this can pose much more effort to systems engineers. Approaches to auto-generate and instantiate the algorithm-level explanation to a version that is also task-specific are still needed.

Assurances needs. Assurances in HCI literature is a design property that also applies to the intelligent tool so that they help users' trust calibration (Israelsen and Ahmed, 2019).

Assurances are indicators and performance metrics to indicate the actual capabilities of the intelligent tool. Interestingly, participants described assurances in terms of the XAI class as well as the AI assurances. They also discussed how such assurances could support their intention to engage with AI explanations. Our data revealed two categories' assurances: XAI class validity and XAI class capability. Regarding the XAI class validity, participants described that they were unable to have guidance about trusting the provided explanation validity and correctness. Some suggested knowing the source of the data to assess the credibility of the explanation, with mentioning that it would also need to be up to date with the current changes in the task domain. P7 stated, *"As far as that is concerned, I cannot tell whether this explanation is right or wrong without knowing it is up to date"* and P13 added, *"from time to time we get emails to tell us the treatment x got recognised for diagnosing breast cancer. We need to ensure that the system knows this information"*. Others also asked whether the explanation is generated based on training the AI on multiple data sources and references. P2 declared, *"reliable explanation should cover multiple medical sources and knowledge"*. On the other hand, XAI class capability was related to the metrics used to evaluate the explanation in both AI and task domains. Participants argued that explanation verification with a medical expert should be performed according to accepted standards, incorporating best practices related to expert selection, elicitation protocols, bias avoidance, documentation and peer review. Moreover, participants raised questions that could be answered through clarity about the evaluation metrics used in the XAI models as introduced in a previous survey on interpretable machine learning models (Carvalho et al., 2019). Participants demanded information about the XAI class itself such as a) accuracy of an explanation on unseen cases, b) fidelity that shows how well the explanation is consistent with the underlying AI model, c) stability that represent how similar are the explanation for similar cases d) representativeness that describes how many cases could be covered by given explanation. For example, P1 mentioned, *"I was wondering if this explanation could be generalised for another patient"*. We also argue that such assurances can be essential to repair user trust in an XAI class when the explanation generated from the XAI model is incomplete. Incomplete explanation refers to failure in the explainable model to generate complete meaningful information (Malhi et al., 2020).

Tailoring needs. Participants sought to tailor and customise the explanation output and its presentation to help them in contextualising and interpret the XAI explanation. This was also to meet their decision-making behaviour. For instance, participants were asked to set thresholds for similarity parameters in Example-based explanations. Participants identified Example-based explanations as a useful XAI class for their decision-making process and a way to calibrate their trust. P1 commented, *"I think this is crucial when I am sitting in the clinic and I need to make a decision, examples allow me to ask a whole range of questions even it is one that what will your prognosis be what will the outcome be what how should I treat the patient how can I tell what events would be"*. However, providing examples of similar cases confused participants in terms

of the similarity definition, *“similarity is very hard to determine I am curious how the machine defines similarities”* [P5]. Therefore, participants wanted to control the explanation output by defining their similarity measures, *“I would like to ask for examples based on all the features in the system or subset features of the system find the similar patient for this recommendation”* [P14]. Another example of tailoring was encountered in Local and Global explanations when participants wanted to group a set of features to generate a group feature importance value. P8 described *“I think it was easier to read and recognise when this explanation [Local] groups patient history information in one value”*. Rather than providing a static explanation, our findings suggest integrating a possibility to configure and tailor explanations and what they shall contain and how they are computed. This requirement could also be due to the need to fit the explanation in task workflow by finding, accessing, and focusing on intended information while minimising unrelated information (Petty and Cacioppo, 1986).

Multi-step explainability. It refers to users’ explainability needs after utilising the main presented explanation. During our interviews, participants discussed that they had follow-up questions after reading AI explanations. They mentioned that such information would support them in validating the AI recommendation. For instance, participants figured out potential correlations between different features in the Global explanation in which they could not be able to validate their hypothesis, e.g., P2 commented, *“There is a lot of correlation between the treatment cycle and the patient history when you are aware that the system considering this correlation, I will be able to tell if the explanation is accurate”*. Another example encountered in Counterfactual explanations scenarios when participants wanted to explore the effect of a specific patient feature on the recommendation. P4 commented, *“for this patient scenario, I wanted to observe the AI decision if other toxicities have Grade 4”*. A potential interpretation for a multi-step pattern in our data could be linked to previous learning and cognitive development literature (Kurkul and Corriveau 2018). Humans are likely to ask follow-up questions when they did not receive a casual response in the explanation. This theory provides an interpretation of our participants’ needs during the study for further information. Overall, explainability is a social process between the explainer and the explainee (Miller, 2019), and this may need to follow a multi-step interaction approach between the Human and the XAI interface. P3 commented in that context, *“I would have more accurate judgment if the AI ... if a maybe I always can ask for an explanation about explanation”*. Developers of such explainable interfaces may need to collect data from end-users regarding the users’ information needs after presenting the main explanations. Design considerations for the modalities of such multi-step explainability are also required to balance explainability and usability, e.g., chatbots.

5.4 DISCUSSION

Consistent with previous research (Yang et al., 2020, Cai et al., 2019, Lundberg et al., 2018, Lai and Tan, 2019), we found that AI explanations with their different classes improved the overall

performance of the Human-AI team. In this study, 41 medical practitioners performed 410 Human-AI tasks where these tasks were diverse in their XAI classes and the AI recommendation accuracy, i.e., we included correct and incorrect recommendations. While this work reports on results from a study in the medical domain, our findings are likely applicable to other contexts in collaborative Human-AI decision-making applications for expert users, primarily when high stakes decisions are implemented. In this section, we provide a discussion of the potential implications of our results and design guidelines for enhancing trust calibration.

Perceived understandability of an XAI class can contribute to increasing or decreasing users' engagement with XAI interface. Across different XAI class conditions, XAI class was seen differently by our participants to affect their trust (increase or decrease) based on their perceived understandability. Our results showed that Example-based and Counterfactual explanations were more understandable to our participants than Local, Global, and No explanation conditions. The primary reason might be that these explanations are easy to understand by humans than local and Global explanations that require technical knowledge. According to psychological research, humans are more willing to engage with explanations when they are familiar, simple, and casually relevant (Keil, 2006, Colombo et al., 2017). This means that participants willing to engage with AI explanations may have correlated with their perceived understandability of the explanation during the study. Many participants discussed during the follow-up interviews that they skipped explanations when they could not interpret these explanations in their domain task. During the follow-up interviews, these observations were also confirmed when participants clearly suggested ways of making the explanation interpretable. For instance, participants discussed providing tools and modalities, e.g., interactive and dialogue explanations, with end-users to understand and contextualise explanations that they could not interpret. We argue that perceiving the explanation to be understood as a dimension of trust is crucial to increase participants engagement with AI explanations and therefore support appropriate trust. A previous study by Cai et al. (2019) used an onboarding technique to guide users' understanding of the actual AI capabilities and limitations and ways of using it. They aimed to familiarise AI-based decision-making tool users with the AI and help users build appropriate trust. Our results extend their view and argue that users of XAI systems shall be familiarised with its explanations, e.g., usage scenarios or tutorials, to avoid potential misinterpretability and avoidance behaviour. Future work could explore how guiding users' understandability and interpretability of AI explanations could help calibrated trust. Also, approaches like Participatory Design (Schuler and Namioka, 1993) and Co-Design (Sanders and Stappers, 2008) that involve users early in the process will lead to more acceptable and interpretable explanations that fit their target groups.

Users' perceived reliability of the AI does not seem to be correlated with AI explainability. Importantly, we identified no significant change in the reliability of the AI between the baseline

conditions and different XAI classes. The lack of difference in reliability scores suggests that participants' perceived reliability might not be related to showing explanation, which could be aligned with other AI components such as overall performance. For instance, Dietvorst et al. (2015) found that people are more likely to rely on AI when they can control the algorithmic output. The reliability dimension on trust could be related to another line of research that aims to increase trust with the AI not to calibrate users' trust (Yin et al., 2019, Yu et al., 2019).

Human cognitive biases could have triggered overreliance during the study. Explanations are a common approach for supporting trust calibration in a Human-AI environment. Despite their benefits, recent studies showed that explanations could also be misused by participants (Naiseh et al., 2021). Our research helps to unpack the complicated influence of explanation on behaviour, demonstrating how different XAI classes can affect human behaviour during Human-AI collaborative decision-making tasks. Consistent with prior research, explanations with their various classes (Example-based, Counterfactual and Local explanation) improved the accuracy of the Human-AI task compared to No explanation conditions. However, our results showed that when the AI was not accurate and provide incorrect recommendations, explanations with its different classes did not help our participants recognise incorrect recommendations. Furthermore, our results showed that participants agreed more with the AI across all XAI classes than the No explanation scenario. These observations could be interpreted as participants over-relied on the AI when explanations were provided. We argue that the dual-process theory offers a valuable lens to understand why the explanations may contribute to over-reliance. According to dual-process theory (Groves and Thompson, 1970), humans regularly operate on System 1 thinking, which follows heuristics and shortcuts when making decisions. The settings of our study were under an everyday human-AI collaborative decision-making task which probably made our participants follow system 1. On the other hand, System 2 is infrequently triggered as it is slower and effortful. System 1 probably made our participants vulnerable to cognitive biases during the study, which results in their inability to recognise incorrect recommendations. These results are aligned with previous research (Naiseh et al., 2021, Buçinca et al., 2021), which showed that designers of the XAI interface often assumed that users would engage cognitively with AI explanations and use them to calibrate their trust. Also, some studies showed that XAI users perceive explanations as an AI competency feature rather than individually assessed for their content (Bansal et al., 2021). We argue that calibrating users' trust would require extra effort from both the XAI interface designers and XAI users. XAI designers to debiasing users' behaviour and XAI users to read and engage cognitively with AI explanations. We also acknowledge that it is also possible that following incorrect recommendations by the AI was because of the visual design we adopted, specifically, the Local and Global explanation where participants perceived them to be less understandable. Nonetheless, our results highlight the importance of considering cognitive biases when designing the XAI interface for the trust calibration goal.

One explanation does not fit all users' needs during Human-AI collaborative decision-making task. One explanation does not fit all users' needs during Human-AI collaborative decision-making task (Sokol and Flach, 2020). Our qualitative phase showed that users require XAI modalities and interaction techniques to help trust calibration. Participants viewed the XAI interface as a new interactive system that needs to be customisable to their needs and task requirements. An effective XAI interface needs to answer multiple users' questions and help users adjust the explanation accordingly. Participants posed several requirements while interacting with the XAI interface, such as tailoring, multi-step explainability. This aligns with learning literature that shows that the learning process is personalised and is achieved via an explanatory dialogue (Jérivelík, 2006). Following this explanatory process for XAI would make it engaging and available to a wide range of users. Furthermore, allowing the user to customise explanations extends their utility beyond AI transparency. For instance, the explaineer can steer the explanatory process to inspect errors, e.g., identify errors and biases, and validate a hypothesis, e.g., for counterfactual explanations, users defined constraints on the number and type of features that can or cannot appear in the explanation. Finally, we argue that the case of calibrating users trust may require presenting multiple XAI classes in the XAI interface depending on the task its self and users' needs. Recent studies showed that different XAI classes could be useful to support various human reasoning methods and mitigate potential cognitive biases (Wang et al., 2019, Lim and Dey, 2010). For instance, Wang et al. (2019) discussed that counterfactual explanation is useful to mitigate anchoring bias which occurs when humans form a skewed perception and limit the possibility of exploring alternative options. Counterfactual explanations by its design determine what input features could change the AI recommendation and help humans expose to alternative decisions and scenarios.

5.5 LIMITATIONS OF THE STUDY

We note several limitations to our work. First, the study was conducted online using hypothetical patient scenarios and focused on screening prescription as an ML classification problem. Future studies shall also examine these settings in real-world clinical scenarios with different ML problems. Second, although our sample size met the requirements of a power analysis, a larger sample size across the quantitative phase would have made for more conclusive results. Further, as our sample was recruited from the mailing list and included participants from the UK, it was not fully representative. Second, the measure used for people's cognition-based trust to rate their reliability, understandability and technical competence, thus participants' expectations could have been influenced by the experimental items. Finally, the current study cannot fully provide a definitive explanation for why explanations do not always facilitate calibrated trust.

5.6 CHAPTER SUMMARY

In this chapter, we studied the effect of XAI class on trust calibration during Human-AI collaborative decision-making task. We found that Example-based and Counterfactual explanations were perceived significantly understandable by participants. On the other hand, interpreting Local and Global explanations might require additional design considerations and interactive approaches to operationalise these explanations for end-users. Our study findings contribute to the growing literature that examines how to make the collaboration between human decision-makers and AI safer and more efficient. Finally, our results showed that the presence of explanation with its different classes could introduce over-reliance on the AI, i.e., participants were more likely to follow AI recommendations when explanations were presented. These results pose future challenges for future work to explore XAI design modalities and principles to mitigate potential over-reliance risk when explanations are provided.

6. CHAPTER 6: SYSTEMATIC USERS' ERRORS WITH AI-BASED EXPLANATIONS

The increased adoption of collaborative Human-AI decision-making tools triggered a need to explain the recommendations for safe and effective collaboration. However, evidence from the recent literature showed that the current implementation of AI explanations is failing to achieve adequate trust calibration. Such failure has led decision-makers to either end up with over-trust, e.g., people follow incorrect recommendations or under-trust, they reject a correct recommendation. In this chapter, we explore how users interact with explanations and why trust calibration errors occur. We take clinical decision-support systems as a case study. Our empirical investigation is based on think-aloud protocol and observations, supported by scenarios and decision-making exercises utilizing a set of explainable recommendations interfaces. This study involved 16 participants from the medical domain who use clinical decision support systems frequently. Findings showed that participants had two systematic errors while interacting with the explanations either by skipping them or misapplying them in their task. Such errors limited XAI interface goal to calibrate users' trust. This chapter is published in the Computer Journal, Special Issues of Explainable AI (Naiseh et al. 2021).

6.1 RESEARCH RATIONALE

Current advances in machine learning have increased the enactment of human-AI collaborative decision-making tools in safety-critical applications such as medical systems and military applications (Arrieta et al., 2020). Researchers have identified trust calibration as the main requirement for safe and responsible implementation of such tools in everyday scenarios (Naiseh et al., 2020b, Zhang et al., 2020a). Trust calibration is the process of successful judgment of the main components of trust: cognition-based trust and affect-based trust (Lee and See, 2004). Trust is calibrated when the human operator can understand and adjust their level of trust to the current state of the AI [3]. This adjustment is crucial due to the dynamic and uncertain nature of AI-based applications. When users fail to manage their trust, they either end up with over-trust, e.g., people follow incorrect recommendations or under-trust, and they reject a correct recommendation. Previous research (Lee and See, 2004) identified five primary contexts where trust calibration errors in automation occur, their reasons for occurrences and potential design solutions. Overall, trust calibration errors can happen when users do not understand the system functionality, do not know its capability, overwhelmed with the system output, lack situation awareness or feel a loss of control the system. Such faulty in design has shown critical safety issues (Lee and See, 2004).

Research in eXplainable AI (XAI) showed that augmenting AI-based recommendations by explanations can enhance trust calibration as it can give human decision-makers insights and transparency on how the AI arrived at its recommendation. Explanations are supposed to

support users in developing correct mental models of the AI, identifying situations when recommendations are correct or incorrect, and mitigating trust calibration errors (Naiseh et al., 2020c, Zhang et al., 2020a, Cai et al., 2019b). However, recent evidence suggests that explainable AI-based systems also have not improved a successful trust calibration as users' still, on average, end up in situations where they over-trust or under-trust the AI-based recommendations (Zhang et al., 2020a, Bussone et al., 2015). In the context of XAI and trust calibration, previous work has typically focused on evaluating explanations in trust calibration context (Bussone et al., 2015) and identifying explanation types (Zhang et al., 2020a) and presentation formats (Tomsett et al., 2020) for improved trust calibration. In general, the work often assumed that people would engage cognitively with each explanation and use its content to build a correct mental model and improve trust calibration. However, this assumption can be incorrect; humans are often reluctant to engage in what they perceived as effortful behaviour (Kool and Botvinick, 2018) resulting in less informed trust decisions.

Indeed, some studies demonstrated situations where explanation failed to enhance users' trust calibration, e.g., explanations were perceived as an information overload (Naiseh et al., 2020b). Others also related the failure of explanation to improve trust calibration errors to human behaviour and cognitive biases, e.g., cognitive laziness of humans to read explanations (Wagner and Robinette, 2020). Despite the emerging need to design effective XAI interfaces to calibrate users' trust, there is a need for more knowledge about situations and contexts in which explanations do not enable adequate trust calibration, i.e., what kind of scenarios or errors could happen in real-time.

To this end, we aim to explore people interaction behaviour with explanations in Human-AI collaborative decision-making tasks. Such knowledge would ultimately inform future design affordances and aid researchers and designers in developing effective calibrated trust XAI interfaces. In this study, we pose the following research questions:

- How do users interact with explanations during their Human-AI collaborative decision-making task?
- What are those situations where users fail to calibrate their trust in the presence of explanations?

To answer these questions, we conducted a two-stage qualitative study that involved 16 participants (doctors and pharmacists) who use AI-based decision-support tools frequently in their clinical settings. Our results include a qualitative investigation of people interaction behaviour with AI explanations that revealed two systematic users' errors, leading to trust calibration flaws and their reasons.

6.2 RESEARCH METHOD

We conducted a think-aloud protocol where participants were asked to perform Human-AI collaborative decision-making tasks. We then conducted follow-up interviews to gain more insights and discuss our observation on participants' experience during the task. To help our investigation, we designed an AI-based decision-support mock-up tool that is meant to support medical practitioners in classifying the prescriptions into confirmed or rejected. Prescription classification is a process that medical experts in a clinic follow to ensure that a prescription is prescribed for its clinical purpose and fit the patient profile and history. We designed the mock-up based on templates and interfaces that are familiar to our participants in their everyday decision-making tasks. The scenarios simulated a diversity of conditions and explanation types that the decision-maker could face in the real-world scenarios where trust calibration errors could happen, e.g., imperfect AI recommendation due to the dynamic nature of the application. Hence, we included both correct and incorrect recommendations for each class. We chose a prescription classification case study as it reflects a high-cost decision-making task performed collaboratively between the human expert and the AI. All supporting material for this study can be found in **Appendix 3**.

6.2.1 RECRUITMENT AND PARTICIPANTS

We approached three hospitals in the UK by sending an email invitation and got a positive response from 16 individuals. No more participants were approached because during the data analysis, resulted in themes and codes became eventually repetitive. We followed the principles of reaching the saturation point in qualitative methods (Faulkner and Trotter, 2017). This was a reasonable assurance that further data collection would introduce similar results and would confirm the existing themes. Details about the population are provided in Table 20. A study protocol was developed, and pilot tested with two practitioners, one medical academic and one AI expert.

TABLE 19 STUDY POPULATION

Variable	Value	N=16	%
Age	20-30	5	31.25%
	30-40	7	43.75%
	40-50	4	25%
Gender	Male	10	62.5%
	Female	6	37.5%
Role	Doctors	4	25%
	Pharmacists	12	27%
Experience	<5	4	25%

	5-10	8	50%
	10-15	3	18.75%
	>15	1	6.25%
Hospital	A	6	37.5%
	B	6	37.5%
	C	4	25%

6.2.2 CONSENT PROCEDURE

First, the participants were briefed about the study, verbally and through a written participant information sheet. They were then asked to sign a consent form. Participants were also asked several questions about themselves, such as their experience. For enhancing the validity of the collected data, we designed the study to avoid promoting participants to think about explanations and trust calibration as a main objective of the study. We initially demonstrated the study purpose, describing it as an investigation on how medical practitioners use AI-based tools in their work environment. We also mentioned that AI-based tools can explain why a recommendation has been made. Participants were told they could discontinue the study at any point. We debriefed the participants after the study about the detailed purpose of the study.

6.2.3 STUDY PROCEDURE

We gave each of our participants ten scenarios that included AI-based recommendations. Each scenario was accompanied by an explanation. We used five explanation types revealed from **Chapter 4** literature review: Local, Global, Example-based, Counterfactual and Confidence explanations. The scenarios presented to our participants were hypothetical scenarios designed in collaboration with a medical oncologist. We designed the scenarios to be clear, challenging and not trivial so that recommendations, explanations and trust calibration were all substantial processes. This ultimately helped to put our participants in a realistic setting: exposing them to an AI-based recommendation and its explanations where trust calibration is needed and where errors in that process are possible. The 16 participants were asked to make decisions considering the patient profile, the recommendation and the explanations and whether to follow the AI-based recommendation if they see it as correct or reject it if they see it as incorrect. For each scenario, participants were encouraged to think aloud during their decision-making process. They were asked to think freely and encouraged to make optimal decisions. Each of the participants completed ten scenarios representing two cases (correct and incorrect) of each of the five explanation classes. This resulted in 160 completed decision-making tasks. The researcher observed, audio-recorded the sessions and took notes. Finally, we invited our participants to a follow-up interview about their task and explainability experience. Figure 20 summarizes the study workflow.

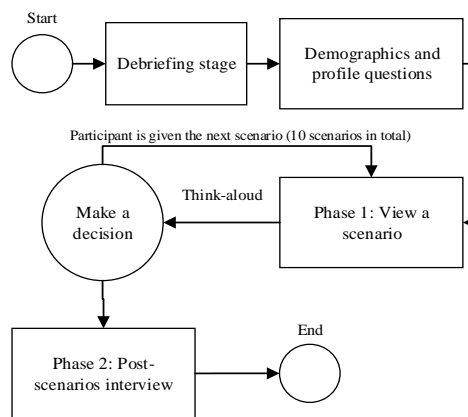


FIGURE 20 STUDY WORKFLOW

6.2.4 DATA ANALYSIS

Two sets of data were collected and used to answer our research questions in this study. The first consisted of the transcript of audio files of both of the study stages (the think-aloud and the follow-up interviews). The second is the researchers' notes, which contained their observations of participants' behaviour and interaction style with the XAI interface. For qualitative data, we performed a content analysis with the Nvivo tool support. The authors had an initial meeting where they agreed on common grounds and analysis scope and style. The analysis was mainly done by the first author. The analysis was reviewed iteratively by the others through frequent meetings which led to split, modify, discard or add categories to ensure that all responses and their contexts were well represented and categorized. No more participants were approached because, during the data analysis, themes and codes resulting from the analysis became eventually repetitive. Principles of reaching the saturation point were followed in the qualitative data collection introduced by (Faulkner and Trotter, 2017). This was a reasonable assurance that further qualitative data collection would introduce similar results and would confirm the existing themes.

6.2.5 STRENGTHS AND LIMITATIONS

Scenarios in combination with the think-aloud approach are valuable for gaining insight into decision-making mechanisms (Wolcott and Lobczowski, 2020). An additional strength of this study was the variety of explanation types used in scenarios, which triggered different responses from participants. All participants were shown the same ten scenarios. Since our sample included three different hospitals in the UK, the results are not limited to a specific practice. Furthermore, participants differed in experience, age and gender, making the sample diverse within this specific field of expertise.

Although scenarios were created to reflect daily practice, practitioners often emphasized additional steps that they would normally take before reaching a decision, such as discussing with colleagues and meeting with patients. These options were not available in the study in which practitioners could only express their desire to know more information about the scenarios. This caused practitioners to work with their knowledge and the available explanations instead of offering them the possibility to investigate their uncertainties. Furthermore, the think-aloud methodology does not ensure that all thoughts behind a decision are explicit. Some decision-making steps might have been applied implicitly, i.e., as tacit knowledge (Howells, 1996) which might have been the case for user's interaction behaviour with explanations. We tried to mitigate that through follow-up interviews. Finally, the study is qualitative involving a relatively small sample and our results are yet to be tested for generalizability. Our main purpose is to shed light on important design considerations when designing XAI for calibrated trust goals. Longitudinal studies and more objective measures, possibly through experimental design, are still needed to validate our results and map them to explanation types.

6.3 RESULTS

In this section, we report on our studies' results that are related to systematic users' errors when interacting with explanations. Through observations and think-aloud, we investigate reasons why explanations may not improve trust calibration focusing on the cognitive dimension of trust within a sample of professionals in the medical domain. Our results indicated that for this trust facet and sample, users' errors were the main source of errors in trust calibration leading to making an incorrect decision. However, these users' errors may not be exclusive to calibrated trust design goal. In addition, such errors could also be linked to other design goals for XAI interfaces, such as perceived fairness of the AI (Lai and Tan, 2019) and explanation usability (Narayanan et al., 2018).

Our analysis showed three main themes of users' behaviour: skipping, applying and misapplying. Within the scope of this paper, we only focus on skipping and misapplying themes that relate to errors in trust calibration in the presence of explanations. We considered an error as systematic if it happened for all explanation types. We also required that these errors crosscut all scenarios to avoid a case where an issue stems from one or a few scenarios and designs. Figure 21 shows a frequency analysis of behaviours when interacting with 160 interactions with explanation interfaces and the emerged themes (ten were shown to each of 16 participants).

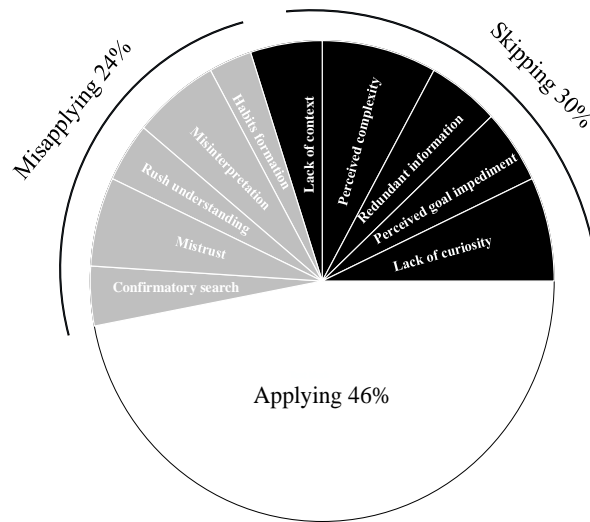


FIGURE 21 PARTICIPANTS' INTERACTION BEHAVIOUR WITH AI-BASED EXPLANATIONS

6.3.1 SKIPPING EXPLANATIONS

Explanations might fail to support the trust calibration process when they are skipped. We observed that some participants made decisions collaboratively with our AI-based decision-making tool without thoroughly reading explanations. In the following sections, we describe the main reasons for errors in the Skipping category.

Lack of curiosity. Curiosity describes the desire to know, learn, or experience an explanation (Miller, 2019). During the study, participants showed a lack of curiosity to seek an explanation from the AI-based tool. Participants did not feel that the explanation motivated them to learn new ideas, resolve knowledge and solve problems. P5 mentioned, “... *to be honest with you, I was not really interested in reading the explanation ... I mean I did not feel that could add something new to me*”. Previous research showed that humans are selective when being curious to seek for explanation and depend on the context and individual characteristics (Wrobel et al., 2013). For example, people might be more curious to read an explanation when the recommendation does not meet their expectations. Furthermore, in scenarios when the explanation contained too many features and information, participants' degree of curiosity was low, and participants were silent during these scenarios. Such situations led participants to skip explanations and discourage them from engaging in what they perceived as effortful processing behaviour.

Perceived goal impediment. Participants skipped the explanations that they perceived as a goal impediment. During the study, several participants were focused on finishing the task and making decisions with an AI-based tool rather than reading the explanation. According to reversal theory (Apter, 1997); individuals in a serious-minded state have a high goal orientation, while those in the playful-minded state have low goal orientation. People in high critical decision-making environments are likely to be in a serious-minded state, where additional

information might be prone to be perceived as a goal impediment. Furthermore, perceived goal impediment could be related to factors such as time constraints and multi-tasking. P12 mentioned, “... *that [explainability] experience was good in general ... but I doubt that it could work in real-world ... doctors and pharmacists are too busy to validate each decision with an explanation*”. Similarly, P6 added, “... *I cannot see how these explanations will work in everyday prescriptions screening*”. Such interruption into users’ tasks leads to psychological reactance and results in users’ avoidance (Van Doorn et al., 2010). Previous research used the theory of psychological reactance to explain users’ avoidance of online advertisement content (Van Doorn et al., 2010). This theory shows that people tend to be psychologically aroused when they perceive their freedom to be threatened by others. This tendency leads individuals to restore threatened freedom by reacting to the threat. In the field of communication, the theory of psychological reactance offers an explanation for why persuasive messages, including explanations, can sometimes produce odds with their intent. Humans reject or move away from a message if the message threatens or attempts to reduce his or her personal freedom of the decision. We argue that increasing users’ perceived value of the explanations would make them less likely to be skipped. For instance, the explanation design might bind into regret aversion bias (Caraban et al., 2019), e.g., people might become more careful in reading explanations when they are informed about a certain level of risk from skipping them.

Redundant information is another cause of skipping as participants mentioned that in certain scenarios, explanations contained information that is simplistic and common sense for them. For instance, P9 stated, “*The average pharmacist does not need to see all these factors that the AI is considering, some of them are just simple rules*”. Also, P6 criticized Counterfactual explanation and stated, “... *mentioning the AI could change its decision if age was 29 does not consider as a useful explanation in our setting ... I mean we all know that ... explanations should be smart enough*”. Research in cognitive science and explanations showed that people tend to avoid circular and redundant explanations (Keil, 2006). For example, people refuse an explanation such as “*this diet plan works because it helps people lose weight*”. Such repetition of facts and no additional substantiation would make users lose their trust in explanations and even avoid further explanations. In general, people evaluate the meaningfulness of the explanations based on three main dimensions: Circularity, Relevance and Coherence (Keil, 2006). To address this issue, previous research (Miller, 2019) proposed the theory of mind to suggest a design solution for achieving meaningful explanations to users in explainable AI applications. The research argued that intelligent agents should keep track of what has already been explained to users and evolve explanations over time. The adoption of adaptive and personalized user interfaces (Naiseh et al., 2020c) would also be a potential solution direction. In summary, techniques to construct a user model, either explicitly or implicitly are required in future Human-AI collaborative decision-making tools to avoid repetitive explanations.

Perceived complexity. Participants ignored explanations because they thought it would take too much time to understand them, e.g., long explanations. In contrast, shorter explanations such as Counterfactual explanations caught participants' attention. For instance, P11 ignored a Global explanation but read and engaged with Counterfactual explanation, and mentioned: *"It could be useful, but I won't bother digging what does that mean"*. Participants discussed making quick judgments whether they would interact with explanations or completely skip them based on explanation length. For instance, P12 stated during Global explanation scenario, *"I would usually look for the first three or four values"*. Explanation's variables such as their size, number of chunks and lines showed to confuse users and made explanations less acceptable (Narayanan et al., 2018). Such long explanations require more processing time and contribute to lower user satisfaction. Another factor that contributes to avoiding long explanations may involve the order in which people receive the explanations (Fischer et al., 2008). People tend to rely on the information presented at the beginning when they try to form an intention to read (Oppenheimer, 2004). Therefore, the order of the explanation chunks could be crucial to engage users with explanations and avoid skipping long explanations.

Lack of context. Participants ignored explanations that they could not contextualize to their everyday decision-making tasks. We found that participants were often expecting explanations to be task-centred and reflective of their domain knowledge and terminology. In a Counterfactual explanation scenario, P8 stated, *"I find this irrational, the explanation is saying the prescription would have been prescribed if the patient age is 50 ... I mean patient age is not something we can change ... I expected something like a blood test or any other variable that we can do something about it"*. Another case of skipping explanations due to lack of context was when participants asked for additional contextual information to contextualize the explanation to their medical practice. P9, who skipped a Global explanation mentioned, *"I would like also to see correlations between patient information to judge whether this is valid information in this case"*. Overall, participants were more motivated to engage with explanations that are reflective of their task characteristics rather than understanding the reasoning of the AI. User-centred iterative design with collaboration with domain experts, e.g., medical doctors, to identify task-centred explanations is needed.

6.3.2 MISAPPLYING EXPLANATIONS

Even when participants engaged with explanations and paid attention to them, we observed that they also misapplied them in their tasks. In the following sections, we discuss the main situations that led to misapplication errors.

Misinterpretation. Some participants misinterpreted our presented explanations and that led to incorrect conclusions about explanations and recommendations. For example, P2, who is a pharmacist, mentioned that the AI-based tool is biased based on his interpretation of the Global

explanation. The explanation in those scenarios gave a high importance value for patients' blood tests in the recommendation. P2 stated: "... *so shall we screen all prescriptions only on blood results?*". Such misinterpretation led to distrust in the AI-based tool. Similarly, P9 had a false interpretation of a Confidence explanation and stated that "*44% certainty in a diagnosis is a good value*". Participants depended on their previous knowledge to interpret the available explanations, which led to building the wrong conclusion. It may be useful to accompany the AI-based tool with an onboarding feature that allows users to understand and familiarize themselves with explanations and their interpretations. Such a technique has been used in the literature of Human-AI interaction by Cai et al. (2019) to familiarize medical practitioners with AI-based cancer prediction tools. This offered a way to aid users in building correct mental models regarding the actual capabilities and limitations of the tool. For example, videos tutorials or FAQs could serve that goal.

Mistrust. Although our participants often assume that explanations are cooperative, they were also well prepared to mistrust them. Some participants felt that explanations were deceptive or untrustworthy to follow. Participants quickly assessed that explanation and voiced scepticism about the correctness and validity explanations. P8 noted, "*I am wondering if an experienced pharmacist has looked at this before*". Sometimes scepticism about the explanation content was combined with scepticism about the source of the explanation. For example, P5 wondered if Local explanation considered data coming from different hospitals, "*we have got to know which hospital this explanation covers, this could completely change my opinion about this explanation*". Our participant required several meta-information about the explanation to judge its correctness and solve mistrust issues. People might mistrust an explanation based on what they know about the motivations and abilities of its sources (Oppenheimer, 2004). Given the well-known phenomena in the psychology literature, addressing such suspicion in the XAI interface can be detrimental for user mistrust correction.

Confirmatory search. Participants did not read the full explanation and searched for information that confirmed their initial hypothesis, i.e., they were selective in what to read and rely on. When shown an Example-based explanation, P4 who is a pharmacist stated, "*Well, I would look for the examples that I've already experienced in the past*". During the study, participants did not take into consideration disconfirming their hypothesis to correct their mental model but found confirming evidence to further strengthen their hypothesis. They completed their explanation analysis with the overconfidence of their initial insights and ended up with trust calibration errors. Several variables can facilitate confirmatory search tendency during the decision-making, such as the increased number of the available information, sequential information presentation, or negative mood (Fischer et al., 2008). XAI research is to look for design techniques that encourage them to read the full explanation and avoid bias.

Rush understanding. Participants incorrectly held a belief that they understand the explanation deeper than they actually did. This effect was obvious in the interview stage, e.g., P4 stated, “*Well in many cases I could predict how the AI work after reading the explanations in first two cases*”. Likewise, P7 mentioned, “... *I would say that I have a confidence to tell how it worked*”. However, they failed to answer our follow-up questions that delved into the details and conclusions. Such miscalibration of their understanding is another case of the overconfidence effect (Keil, 2006). Furthermore, rush understanding could also be related to the explanation itself, e.g., being incomplete or reduced, which made it difficult to have much practice in assessing ones’ understanding. One design solution could be by slowing the users down to enable reflection over their actions.

Habits formation. As job actions and decisions are typically repetitive, users collaborating with an AI-based decision-making tool are prone to develop habits (Wood and Rüniger, 2016). During the study, participants became gradually less interested in the details of an explanation and overlooked it altogether. Such behaviour is associated with the development of peoples’ expectations about the behaviour and the performance of the environment (Van Doorn et al., 2010). P4 who showed similar behaviour mentioned, “*I think this is similar to the previous explanation*”. Such habits could damage the explanation goal to support trust calibration. For instance, doctors with a successful diagnosis experience with an AI may fail to notice a minor change in the AI accuracy and the explanation output. The continuous pairing of collaborative diagnosis with positive outcome may in time cause the act to become automatic, triggering an unconscious response which is no longer linked to the explanation output (Wood et al., 2005). Habits might be also triggered by prior interaction in a chain of responses, by environmental cues, such as time of the day or location, or by the particular internal state such as moods (Van Doorn et al., 2010). XAI design is to monitor such habits formations and try to prevent them, e.g., when a user agrees excessively, an adaptative design approach can change the explanation interface structure so that it triggers fresh thinking.

6.4 DISCUSSION

One main goal of communicating explanations in Human-AI collaborative decision-making is to enhance the trust calibration process. This chapter has examined the role of explainability in enhancing Human-AI collaborative decision-making and the trust calibration process in particular. One of the key findings is that explanations failed to support users in their trust calibration process due to two primary users’ errors: skipping and misapplying. We argue that building XAI interfaces that consider these errors and develop design constraints to limit them can support the explanation goal of enhancing trust calibration. For instance, we observed a high frequency of skipping explanations when participants perceived explanations as an impediment to their task. As a corollary, a design that fits the explanation in the task workflow

can limit such errors and may support the trust calibration process as users would read the explanation and understand the AI reasoning.

Also, the relationship between failing to calibrate trust and user errors could be further investigated through the lens of human decision-making processes. According to the Elaboration Likelihood Model (ELM), humans process information in two different routes: a central route in which information processing is slow and reflective and a peripheral route in which information processing is fast and relies on mental shortcuts (Evans, 2008). It has been suggested that individuals have the disposition to use the peripheral route as it saves time and effort, and this type of processing is especially relevant to medical settings where time constraints exist. While mental shortcuts are usually effective in decision-making, their unconscious and automatic nature make them prone to cognitive biases. Overall, implementing AI-supported decision-making tools with explanations could be a way of mitigating biases that people might have in their everyday decision-making tasks as such explanations could activate central route processing (Keil, 2006). However, human biases could also influence the processing of explanations and this can lead decision-makers to either end up with under-trust or over-trust. For example, under-trust may result from anchoring bias when participants look at only salient features of AI explanations and consequently judge the quality of information to be untrustworthy. Similarly, over-trust may result from confirmation bias as mentioned before when participants favour explanations that are consistent with their initial hypothesis. In this light, the presentation of explanations has the risk of further reinforcing biases that decision-makers may already have. This highlights the necessity to address cognitive biases in the design of explanations.

Finally, either skipping or misapplying explanations could be resulted from the fact that participants did not seek an explanation. Such behaviour limited users' learning process of the AI reasoning and its underlying logic, so their trust was not calibrated. It has been found that despite the availability of explanations, people might utilize a small amount of them or avoid seeking explanations, even when they need them (Oppenheimer, 2004). Thus, if the goal of explanations is to calibrate users' trust, effective explanation seeking behaviour may contribute to improving users' learning and trust calibration processes. Our results pose a new requirement for XAI interfaces to focus, especially at the earlier stage of interacting with the AI, on increasing explanation-seeking behaviour. This could be potentially implemented by applying principles of persuasive design (Oinas-Kukkonen and Harjumaa, 2009) and persuasive learning (Aleven et al., 2003), e.g., showing users' level of knowledge about the AI.

6.5 CHAPTER SUMMARY

Designing explanations for trust calibration has been identified as one of the main goals for safe and effective AI-supported decision-making tools. However, it is often remaining unclear in the

literature why explanations were not always supporting users' in their trust calibration. This motivated our work to explore how people interact with explanations in their Human-AI collaborative decision-making task. We focused on particular situations where explanations did not effectively support users to calibrate their trust. As a general conclusion, explainability for trust calibration might conflict with usability: trust calibration requires extra efforts from the users, e.g. read and interact with the explanation. Thus, integrating explanations in Human-AI collaborative decision-making environments needs to analyse and explore the costs and benefits of favouring between explainability and usability.

7. CHAPTER 7: XAI INTERFACE DESIGN PRINCIPLES TO HELP TRUST CALIBRATION

AI-based decision-making tools are being increasingly applied in critical domains such as healthcare. These tools are often seen as closed and intransparent for human decision-makers. An essential requirement for their success is the ability to provide explanations about themselves that are understandable and meaningful to the users. While explanations have generally positive connotations, studies showed that assuming that users would interact and engage with these explanations can introduce trust calibration errors such as facilitating irrational or less thoughtful agreement or disagreement with the AI recommendation. In this chapter, the research explores how to help trust calibration through AI explanations design. The research conducted a think-aloud study with 16 participants aiming to reveal the main trust calibration errors concerning explainability in AI-Human collaborative decision-making tools – Chapter 6. Then, two co-design sessions were conducted with eight participants to identify design principles and techniques for explanations that help trust calibration. As a conclusion of the two stages, five design principles are provided: Design for engagement, challenging habitual actions, attention guidance, friction and support training and learning. These findings are meant to pave the way towards a more integrated framework for designing explanations with trust calibration as a primary goal. The chapter is structured as follows. In Section 7.1, theoretical background and related work are provided. In Section 7.2, a description of the research method followed in this stage, including the sample, material and instruments used, and our analysis is also provided. In Section 7.3, results and findings from both qualitative stages are presented. Finally, in Section 7.4, the chapter discusses design principles meant to guide designing XAI for trust calibration.

7.1 THEORETICAL BACKGROUND AND RELATED WORK.

The decision-making process in collaborative environments is based on conveying information between the collaborative members (Parayitam and Dooley, 2009). Decision-makers in such environments may need to process various information to make the final decision (Ashby, 1961). The modality in which the information is communicated between the members has a significant role in decision-making quality (Galbraith, 1973). Also, the increase in the complexity of the decisions triggers an increase in the information needed to explain the underpinning logic and, hence, the effort needed for effective usage of the available information. According to (Parayitam and Dooley, 2009), people at the time of making the decision are highly influenced by individual and affective factors of trust; some are related to the entity conveying the information, e.g. whether a human or a computer system.

In Human-Computer trust literature (Madsen and Gregor, 2000), two distinct trust dimensions are identified: cognition-based trust and affect-based trust. These dimensions

distinguish between the cognitive components of trust from the emotional components. Cognition-based components include perceived understandability, perceived reliability and perceived technical competence. Cognition-based trust enables people to use their intellectual and reasoning skills. On the other hand, an affect-based trust includes personal attachment and faith. These components of trust refer to the emotional bond, in our case, between the human and the AI, which does not result from reasoning and understanding but feeling, sense and previous experience. Previous research showed that both affect-based trust and cognition-based trust have an impact on the decision outcome (Zucker, 1986). Cognition-based trust is crucial for establishing appropriate trust, whereas affect-based trust is developed as the relation continues (Nah and Davis, 2002). Furthermore, previous research showed that in critical decision-making scenarios, it is highly likely that cognition-based trust components are more significant for trust calibration (McAllister, 1995, Lewis and Weigert, 1985). Table 21 defines explanation characteristics in terms of cognition-based trust components adapted from Madsen and Gregor (2000).

TABLE 20 HUMAN-EXPLANATION COGNITION-BASED TRUST COMPONENTS DEFINITIONS ADAPTED FROM (MADSON AND GREGOR, 2000)

Perceived understandability	The degree to which the decision-maker considers the explanation as helpful in forming a mental model about the AI-based tool.
Perceived reliability	The degree to which the decision-maker sees that the explanation provides an argument discussion and well-reasoned theory based on strong evidence
Perceived technical competence	The degree to which the decision-maker considers the explanation has been successful in analysing the recommendation accurately and correctly based on the available information.

Calibrated trust is the process of successful judgment of the main components of trust: cognition-based trust and affect-based trust (Lee and See, 2004, Muir, 1994). People evolve their level of trust, in both dimensions of cognition and affect, considering the current state of the AI and their experience with it. As the underlying nature of AI applications is inherently uncertain and dynamic, decision-makers working with an AI face difficulty in calibrating their trust in an agent. The uncertainty and complexity may lead them to over-trust the system and follow an incorrect recommendation or under-trust in rejecting a correct recommendation. Previous research (Lee and See, 2004, Wickens, 1995) identified five primary contexts where trust calibration errors in automation occur, their reasons of occurrences and potential design solutions. These errors can happen when users do not understand the system functionality, do

not know its capability, are overwhelmed with the system output, lack situation awareness or feel a loss of control over the system. Such faulty design has shown critical safety issues (Robinette et al., 2016).

One goal of explainable artificial intelligence is to mitigate trust calibration errors (Yang et al., 2017). The approach aims to help users of intelligent systems build an appropriate trust level by showing users the rationale and reasoning behind an agent recommendation. Although, many studies showed that explanation could indeed improve trust calibration (Tomsett et al., 2020). However, such studies often assumed that users would engage cognitively with explanations and calibrate their trust. Recent studies showed that even though explanations are communicated to people, trust calibration is not improved (Zhang et al., 2020a, Bussone et al., 2015). Such failure of XAI systems in enhancing trust calibration has been linked to factors such as humans' cognitive biases, e.g., people are selective of what they read and rely on (Naiseh et al., 2020b). Also, others showed that XAI failed to improve calibrated trust because of undesired human behaviour with AI-based explanations, e.g., human laziness to engage in what they perceived as effortful behaviour (Wagner and Robinette, 2021). Overall, users of XAI systems fail, on average, to calibrate their trust, i.e., human decision-makers working collaboratively with an AI can still be notably following incorrect recommendations or rejecting correct ones. This raises a question on how to design explanations to improve or operationalise trust calibration in XAI interfaces. In this study, we aim to explore how XAI interface design can improve the role of explanations in calibrating users' trust and enabling a successful judgment for trust components. Our deriving framework to discover protentional design solutions was digital nudging (Caraban et al.) and also the principles of de-biasing (Soll et al., 2014). Testing whether our results would make that impact would require further testing, possibly within experimental settings. For example, we can hypothesise that a technique like nudging through shuffling the options has the potential to break through the status quo bias and make users more receptive to a different route of thinking, more reflective than automatic. This, however, will also depend on several variables and require extensive research to fine-tune. For example, personality traits like openness to a new experience (John and Srivastava, 1999) can play a role in people decision making and hence respond to nudging in the way mentioned earlier. Furthermore, research on de-biasing (Soll et al., 2014) suggested that to encourage more reflective thinking and avoid automatic thinking .

7.2 RESEARCH METHOD

Our study design and analysis of the data are situated within a two-dimensional space: *everyday Human-AI collaborative decision-making tasks where trust calibration errors are possible, and AI-based explanations to support trust calibration*. Through multi-stage qualitative research, we aim to answer the following questions:

RQ: How to design for explainability that enhances trust calibration? What design techniques could be implemented, and what are suitable principles to guide the design?

To this end, the research method of this paper included two phases: Exploration and Co-design. The exploration phase aimed to explore how users of everyday Human-AI collaborative decision-making tasks interact with AI-based explanations and why explanations are not improving trust calibration. The co-design phase goal was to investigate how users of XAI systems would like to integrate AI-based explanation in their everyday decision-making tasks. The co-design phase helped us to understand how the solution would look from the users' perspective. The materials used in the two stages can be found in the published technical report (Naiseh et al. 2021). The following sections describe the research method. All material used in this study can be found in **Appendix 4**.

7.2.1 USE CASE AND UNDERPINNINGS

Screening prescription is a process that medical experts in a clinic follow to ensure that a prescription is prescribed for its clinical purpose and fit the patient profile and history. The main workflow of the prescribing system is shown in Figure 22.

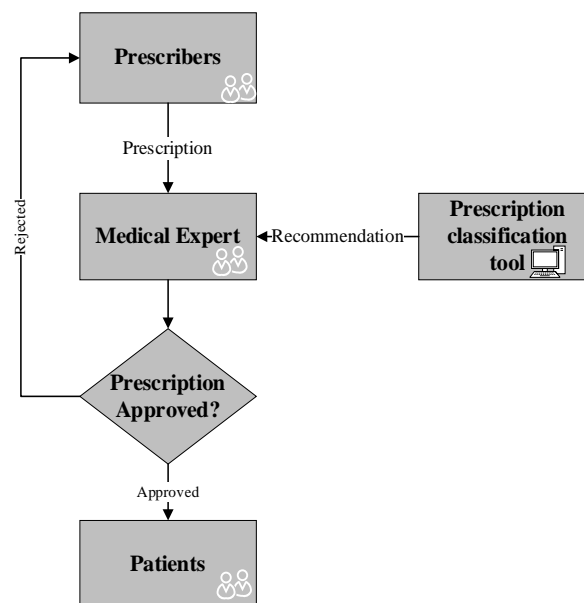


FIGURE 22 SCREENING PRESCRIPTION CLASSIFICATION AI-BASED SYSTEM CLASSIFICATION

To help our investigation, we designed an AI-based decision-making mock-up meant to help classify the prescriptions into confirmed or rejected. We chose this case study to reflect an everyday Human-AI collaborative decision-making task where trust calibration errors are indeed possible. We designed the mock-up based on templates and interfaces familiar to our participants in their everyday decision-making tasks (See Figure 23). Our mock-ups mimic a web-based tool and are meant to simulate the user experience when working on an existing

system. As the medical expert clicks on a prescription, the tool shows the patient profile and the recommendation from the AI-supported decision-making tool (confirmed or rejected). The user has access to AI-based explanations to understand the AI rationale of why the prescription should be confirmed or rejected.

Patient: Jack (Mr) Gender: Male, Born: 11-June-1969 (48Y) SAP Number: markte st007

Palliative Gemcitabine and Paclitaxel Frequency: 21 days

Height: 165 Weight: 74 Surface Area: 1.81 Performance Status: 0 [Update]

☒ Add drug(s) ☒ Re-schedule ☒ Cancel appointment ☒ Add cycle(s) ☒ Add course ☒ Remember Cycle

Prescribed - Approved by AI tool. [Explain Why?]

Prescription Pathology Results Toxicities Notes(1) Documents(1) Allergies Diagnosis History

Time	Drug	Dose	Administration	Frequency	Route	Duration
Day 1 (14/02/2020) Chemotherapy Day Unit						
T=30mins	Chlorpheniramine	10 mg			IV Bolus	1 minutes
T=1hr	Metoclopramide	10 mg		Three times a day	Oral	3 days
T=0	Sodium Chloride 0.9%	318mg	volume dependent upon requirement	IV flush		

[Confirm Prescription] [Cancel authorisation (return to prescriber)] [View PDF prescription] ☒ Day 1 ☒ Day 8

FIGURE 23 A SAMPLE OF PRESCRIBING SYSTEM INTERFACE SUPPORTED WITH AI RECOMMENDATIONS

7.2.2 PARTICIPANTS

We recruited twenty-four participants primarily through an email invitation. Sixteen participants were involved in the exploration phase and eight participants in the co-design phase. The email was sent to two oncology departments in three hospitals and faculty of Health and Social Sciences at Bournemouth University in the UK. We designed a pre-screening survey to get participants' demographic information and their experience in screening prescription as this was the domain we choose for recommendations. Table 22. shows the demographic information of our participants. We anonymise the organisations with A, B and C for data confidentiality purposes. The inclusion criteria for both phases included medical experts who used clinical decision support systems before and have experience in the selected use case. We have chosen the scenarios and recommendations in a way that equally apply to all of our participants in terms of needed expertise and skills. However, the choice of a different role was motivated by the nature of the task, i.e., screening prescription Human-AI task can be followed by doctors and pharmacists in the clinic.

TABLE 21 PARTICIPANTS DEMOGRAPHICS FOR EXPLORATION AND DESIGN PHASES

Phase	Participant	Gender	Age	Role	Year of experience	Organisation
	P1	Male	20-30	Medical Doctor	1-5	B

P2	Male	20-30	Medical Doctor	1-5	B
P3	Female	20-30	Medical Doctor	5-10	D
P4	Female	20-30	Medical Doctor	5-10	B
P5	Female	20-30	Medical Doctor	5-10	D
P6	Male	30-40	Medical Doctor	5-10	B
P7	Female	30-40	Medical Doctor	10-15	C
P8	Male	40-50	Medical Doctor	15-20	B

7.2.3 CO-DESIGN

We conducted two co-design sessions with eight participants, i.e., four participants in each session. The main aim of this stage was to explore how the design can play an effective role in enhancing users' trust calibration during a Human-AI collaborative decision-making task. We used the same inclusion criteria employed in the exploration stage, i.e., expert users in the studied task. We chose to recruit different participants to avoid the learning effect (Lazar et al., 2017) and increase the credibility of our findings as existing users already learned the objective of the study and were part of the underpinnings for this next study. The co-design method enables users who might be potential users in future AI-supported decision-making tools to reflect their experience in the design process, and this is supposed to increase the acceptance of the proposed solutions (Poole et al., 2008). Co-design can lead to a better understanding of the end-user needs, which enhances the possibility of the designs' acceptance (Song and Adams, 1993). In this phase, we discussed and negotiate how to embed AI explanations to serve users' needs, task workflow and trust calibration. Together with the participants, we conceptualised and sketched design features to support users in utilising AI explanation and reduce trust calibration errors revealed from the exploration phased. This was achieved by giving the participants initial prototypes or mock-ups (Clement et al., 2012) of the problem to help them visualise the idea and then provoke brainstorming related to the research problem. All these dynamics were hard to capture during the exploration phase. Therefore, co-design method helped us to come up with innovative designs of how the solution should look from a user perspective.

Participants were divided into two design sessions based on their availability. Due to the COVID-19 situation, we chose to conduct the study online using FreeHand tool from Invision¹. Also, it has been shown that online tools for co-design can make the process easier, cheaper and flexible for participants (Näkki and Antikainen, 2008). To mitigate any potential issues that

¹ <https://freehand.invisionapp.com/freehand/new>.

could arise from using online platforms, e.g., readability of the instructions and the tool usability issues, we conducted a pilot study with two post-graduate researchers and one academic in an interdisciplinary research group residing in the departments of Computing and Psychology in Bournemouth University. This also helped us in the preparation of the training and induction stage for the participants in the real study. All participants attended a training session to familiarise themselves with the tools' functionalities and how they can communicate online. The training session lasted for 15-20 minutes. Then participants were invited to try the tool till they felt all capable of using it. They could ask questions and one of the authors answered them.

We adopted four techniques during the co-design sessions to reach the goal of our study (See Figure 24); researcher presentation, participants discussion, sketching-up exercise and focus groups. This also helped to enhance the credibility of the study and to ensure that data bias was eliminated. Each of the sessions lasted for around 2 hours. Both sessions, including the four main steps, were audio-recorded and transcribed. Audio recording for the design session helped the authors analyse main design needs and issues revealed from participants discussions. The following sections describe each technique that we used in our design sessions.

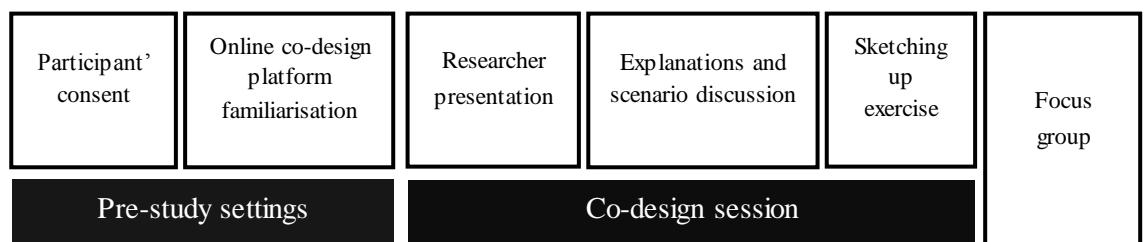


FIGURE 24 CO-DESIGN SESSION WORKFLOW

- A. Researcher presentation (10 mins). The researcher gave a 10-minute presentation on AI-based decision-making tools and an overview regarding the first phase findings, particularly those about different types of errors that emerged during the exploration study. This helped to immerse the participants in the research problem, and it involved a warming-up activity in getting the participants involved in the design sessions.
- B. Explanation and scenario discussion (25 mins): In this stage, participants started by introducing themselves. We then asked each participant to talk about how AI-based tools could help their everyday decision-making process. Then, we defined explainability methods introduced in previous interpretable machine learning surveys (Adadi and Berrada, 2018). We provided different e-cards describing different explanation types in simplified examples. This was meant to illustrate the explainability definition and potential uses of these explanations. To answer our research question, the participants needed first to immerse in a fictional problem as recommended (Buskermolen and Terken, 2012). In our

study, the fictional problem was collaborative decision-making between the medical expert and the AI. Specifically, a screening prescription using an AI-based tool. The researcher invited participants to discuss the designed scenario of an AI-based collaborative decision-making tool of a screening prescription and its generated explanations. We used a random forest classifier as an ML algorithm to train our model. We then generated explanations from current state-of-art model-agnostic explanations to examine how users would like to receive these explanations and develop prototypes for effective utilisation for such explanations in real-world scenarios. This stage was meant to scope the discussion and facilitate focused conversations using the provided scenario. This was also meant to immerse the participants with the research problem and facilitate their understanding of the researcher presentation. Our participants discussed a wide range of trust calibration scenarios using the explanation interfaces through the provided material in this stage. This stage provided a sense of realism to the problem and encouraged careful consideration of solutions to cater to different contexts and usage styles.

- C. Sketching-up exercise (40 mins): Participants were then encouraged to start sketching up their designs using the FreeHand tool from InVision. We gave each participant a blank e-page to sketch up designs considering five explanation types (Local, Global, Example-based, Counterfactual and Confidence explanations). The online platform provided several creation tools (e.g. coloured pens, shapes and sticky notes). The participants were also asked to not limit themselves to the given explanation classes and consider any extra features they would like to see in XAI interfaces to help them in utilising the explanation during a collaborative decision-making task. We deliberately asked our participants to work individually, think outside of the box, and consider different kinds of potential solutions. In this stage, our participants designed their explanations and provided multiple usage scenarios for them. They created a wide variety of usage scenarios covering different purposes and task requirements, e.g., grouping data features in Local explanations to reduce the explanation complexity.
- D. Focus group (45 mins). After each participant completed the sketching activity, each participant presented their ideas to the group. This was meant to critically analyse and evaluate the ideas by the participants to formulate robust solutions. This activity allowed our participants to explore and discuss various ways of using AI explanations in their work environment, considering trust calibration as the primary goal.

7.2.4 DATA ANALYSIS

Three sets of data were collected and used in the analysis stage: (i) audio-recording transcriptions, (ii) researcher notes and observations, and (iii) sketching data. First, we analysed the collected data from the exploration stage, including participants' audio-recording transcriptions files and researchers' observations and notes. The analysed data included themes

for trust calibration errors that need to be addressed in the XAI interface. Exploration analysed data used as an input for the co-design phase. Second, we analysed the data collected during the co-design phase, including audio-recording transcriptions, sketching data and researcher notes. Both stages were analysed using content analysis. Content analysis is used to search for themes and concepts that emerge with a description of the qualitative data study problem. We first familiarise our self with the data. Then the first author coded the data. The other authors mentored the process and verified both the coding and the conclusions made. We followed an iterative process across several research meetings to formulate, combine, and conceptualise the emerged concepts. This iterative process was meant to examine and ensure the codes were interpreted and assigned to the correct themes.

7.3 RESULTS

Results that emerged from the design phase showed that to achieve trust calibration and mitigate exploration stage errors; users require several design techniques to engage and interact with AI-based explanation during their everyday Human-AI collaborative decision-making task. Our results showed that users view explanation interfaces as a new interactive system that needs to be customisable to their needs and task workflow. Across the design phase stages, we identified four main themes characterising these techniques: abstraction, cues, control, and adaptation. These categories help to illustrate different explanation presentation and delivery techniques that need to be considered during the development phase of explainable interfaces for the trust calibration goal. Our proposed design techniques are not mutually exclusive as the explainable interface design could contain one or all of them based on the nature of the task and its requirements. In the following sections, we present the four categories of such design techniques that resulted from our analysis. They can be considered high-level design requirements to support users' in utilising the explanations. Table 22. provides a summary of the suggested designs during the co-design sessions and their definitions.

TABLE 22 THE FOUR MAIN THEMES THAT EMERGED FROM THE CO-DESIGN PHASE

Design technique	Definition
Abstraction	It refers to extracting and generating main features from the explanation and making it possible to present them at multiple abstractions and granularity levels.
Control	It refers to providing customisation functionality to control the information presented in the explanation (e.g., grouping, ordering)
Cues	It refers to additional elements that can draw users' attention and help guide them in the process utilising the explanations.
Adaptation	It refers to varying the explanations characteristics, e.g., information,

abstraction level, cues, order and modalities, in response to an interaction context, i.e., the ability to communicate explanations differently in different settings.

7.3.1.1 ABSTRACTION

The abstraction design technique refers to extracting and generating main features from the explanation to be presented at multiple abstractions and granularity levels in the XAI interface. Abstraction is intended to make it easier to read an explanation and recall the meaning at an appropriate level of details for the user profile and expertise and interest. Numerous studies have shown that the amount of presented information affects users' decisions quality, cognition and trust (Schmell and Umanath, 1988, Plaue et al., 2004, Te'eni and Sani-Kuperberg, 2005, Schaffer et al., 2018). Several participants sketched up explanations that had an abstraction feature according to their ability to understand the explanation and their interest in the details. P19 described his design as *"in this way, I can easily know and remember what is happening"*. One of the produced designs that implemented an abstraction technique can be seen in Figure 25. In this design, P22 combined Local and Counterfactual explanations into one interface and reduced them into two levels of abstraction. The higher level of abstraction contained a narrative summary for Local and Counterfactual main characteristics. The next level of abstraction was to observe in detail the two explanations and compare them together. Participants seemed to be interested in multiple levels of details to minimise the possibility of becoming overwhelmed during their decision-making tasks. P23 commented on the abstraction design feature: *"This is quite useful ... it is good to have different levels of details ... sometimes I do not need to look at all patient information"*. Other participants also agreed on this sentiment: *"it is the quickest and easiest to see at a glance the information you want"* and *"It is informative but also easy"*.

1 Explanation higher level of abstraction

The probability of accepting the prescription is highly affected by schiller=1 and number of smoking years=9 with 60% of its overall confidence. However, the prescription may be rejected if her age=54 with 88% confidence.

2 Explanation lower level of abstraction

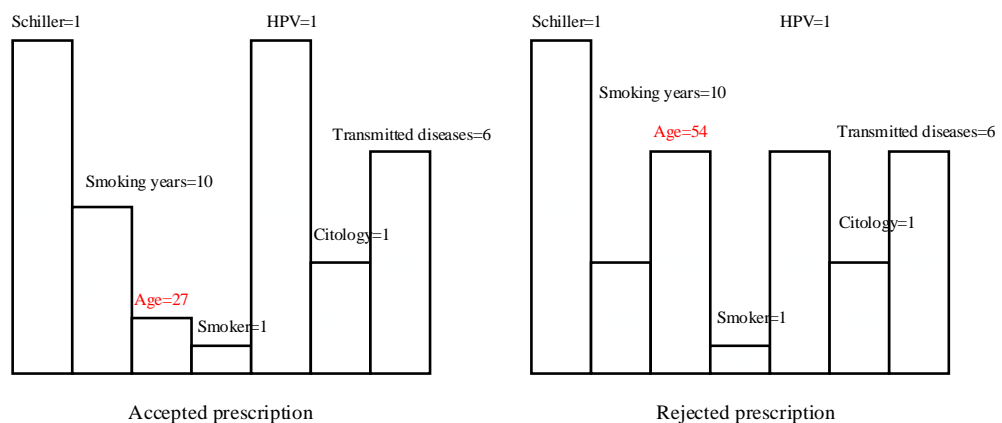


FIGURE 25 AN EXAMPLE OF TWO LEVELS OF ABSTRACTION DESIGN PRESENTED IN OUR CO-DESIGN STUDY

In general, when an AI explanation was presented to our participants, they sought to simplify the explanation into multiple levels of abstraction. This can be interpreted as humans process information at a different level of abstraction to cope with their complexity and optimise their cognitive effort (Cowan and Lucena, 1995). For instance, people usually start reading the main title of the article and then the headlines and then, potentially, the article text. Our participants were more inclined to spend more time on a specific abstraction level more than reading the explanation as one chunk. Abstraction design has been widely accepted and implemented in human-computer interaction (Shneiderman et al., 2016). Te'eni et al. (2005) showed that people usually focus on a particular level of information at a given stage of the decision-making process.

In summary, the abstraction design technique is to facilitate users' concentration to understand a particular level of explanation abstraction, and then shift to another level. It also could increase users' engagement with a usable and user-friendly explanation. This technique could be implemented at the design level using colours, fonts, folding-unfolding technique and multiple views. Designers of explainable interfaces need to consider breaking the explanation into multiple levels to have a better understanding, even for complex decisions and explanations. For instance, decision-makers may not need to go through and process all the abstraction levels to understand the current recommendation and the explanation in cases familiar to them. Also, understanding when and under which conditions the users shift from one level of abstraction to another helps produce more effective trust calibration explanation interfaces. We need further research to design the abstractions level and the navigation between them on the one hand, and the trust calibration process on the other. For example, a question to ask is whether viewing an abstraction at only a higher level of abstraction means a flawed trust calibration process and whether familiarity with the case can always be seen as a moderating factor.

7.3.1.2 CONTROL

Control refers to the customisation techniques provided in the explanation interface to allow users to contextualise and personalise the explanation content and its delivery method. Trust in automation research showed that trust could be developed inappropriately when the explanation does not match the user experience (Lee and See, 2004). Also, previous research showed that static explanations often fail to provide a satisfying explanation among all users (Sokol and Flach, 2020b). During our qualitative phases, our participants wanted to tailor explanations for better understandability and adjust the explanation to assess its accuracy, e.g., in the case of example-based explanation, the participants wanted the AI to regenerate the explanation in a specific time frame. In the following sections, we present a group of control techniques proposed during the design sessions and meant to enable the generation and delivery of a contextualised and personalised explanation and, ultimately, help a better trust calibration process (Figure 26).

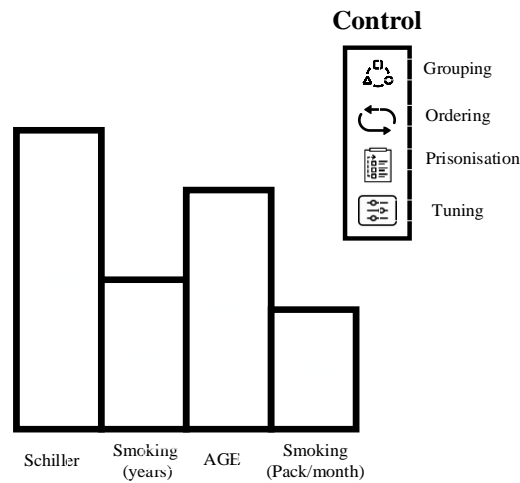


FIGURE 26 AN EXAMPLE OF SUGGESTED CONTROL TECHNIQUES DESIGNS FOR LOCAL EXPLANATIONS IN THE CO-DESIGN SESSIONS

Ordering. It refers to enabling the order customisation for the explanation content whether to meet users experience or also to meet their reasoning process. Order effect is a well-studied phenomenon in the literature of behavioural decision-making. It studies the influence of the information order on decision-making outcomes in a more or less systematic fashion (Strack, 1992). The information offered at the beginning of a sequence might be more influential when people are to make decisions, an effect known as primacy (Tubbs et al., 1993). Also, recency refers to the significant influence the information presented at the end of a sequence has (Bergus et al., 1995). Whether primacy and recency, this effect of information ordering have been experimentally validated in decision-making environments, e.g. (Curley et al., 1988). In human-computer interaction literature, controlling the order is considered an effective technique to meet the evolving user experience (Garcia-Lopez et al., 2015). Similarly, in collaborative AI-based decision-making tools, decision-makers trust judgments are based on several explanation types and information presented in the explanation. The order of an explanation in the explainable interface or a piece of information in the explanation could affect users' decisions, e.g., an explanation can be designed to tell the influence of patients' age before patients' smoking behaviour on AI cancer prediction or put them together in the same order. Several participants sketched different orderings of explanation types and explanation content in their designs in which the order met their decision-making strategy. This pattern also emerged in the focus group stage when participants discussed the benefits of controlling the order of the information in the explanation, e.g., P17 stated: *"Well, usually I would look for first three or four values ... so it would be useful to present them first in the graph"*. Our analysis shows that this ordering feature could be user-controlled or system-adaptive (Section 4.2.4.). For user control feature, users can control the order of the explanation types and the order of the explanation content, so it meets their reasoning process of the explanation interface.

Grouping. With the large numbers of data features that could be presented in explanations such as Local and Global explanations, our study showed a trend where six participants grouped a set

of relevant features to calculate group-feature importance. P19 mentioned: *“It makes more sense to have a value that represents the importance of the transmitted diseases”*. They argued that grouping similar data features in the explanation could reduce its perceived complexity and enable more contextualised and informative explanations. P21 stated: *“... doctors sitting in the clinic will see that complex. Gathering all patient smoking information in one value is more informative”*. Our participants suggested two ways of grouping the features: default groups, e.g., patient history, laboratory tests, and user-created grouping, i.e., ability to create a group, e.g., features that the user thinks are relevant to each other. Our results complied with the previous survey by (Biran and Cotton, 2017) who found that enabling the data feature grouping is vital for non-AI experts’ understandability. Such a technique was essential for our participants to contextualise the explanation in their environment and assess its correctness. P22 commented: *“... combining all the patient history values makes more sense and helps to draw proper conclusions”*. Given this feedback from our participants, it would be essential to allow participants to amend the way that they would like to present the AI explanations, mainly in grouping some data features to convey meaningful interpretation from their perspective.

Prioritisation. It refers to enabling users to prioritise the explanation interface components based on the users’ perspective; this includes choosing their preferred *explanation type* and *explanation content*. Participants argued that this feature would help them in reducing the complexity of the explanation interface and tailoring it to fit their reasoning process. For *explanation type prioritisation*, we observed that several participants used different explanation types in their designs and ignored the other explanation types. For instance, P19 centralised her design on Local and Example-based explanations. P19 stated, *“In my opinion, other explanations are complementary information”*. P17, on the other hand, wanted the explanation interface to include Local and Counterfactual explanations. The discussion about the expressive power of explanation types and disagreement about it, made our participants suggest including an explanation type prioritisation feature where users can choose their preferred type. On the other hand, participants discussed two scenarios to prioritise the *explanation content*: *manual* and *automatic* prioritisation. Manual prioritisation can be seen in P22 Local explanation design when he mentioned: *“I do not need to look at all these patient data effects on AI decision, five or six important ones should be enough ... it is good to have a feature to choose and prioritise amongst them”*. Furthermore, automatic prioritisation was discussed when participants were overwhelmed with the number and explanation size. Automatic prioritisation was mainly about setting thresholds for the importance of the explanation content to be presented in the interface with respect to the AI recommendation. For instance, two participants in Global explanation wanted to show the features which had at least a 15% impact on the AI reasoning. P24 stated, *“I only need to see reasons that have a 15% or more influence”*. Previous research explained this result when they showed that the explanation interface’s complexity varies between humans and affects willingness to interact with the explanation (Narayanan et al., 2018, Bofeng and Yue,

2005, Milliez et al., 2016). Some users prefer a detailed explanation, while others need a brief explanation based on their curiosity level (Miller, 2019). Overall, allowing users to guide and customise the explanation interface might benefit trust calibration goal by making the interface output more suitable to the users' experience and the reasoning process. However, there is also a risk that this freedom of choice leads to a biased trust calibration process. Balancing between the user experience and the goal of calibrating trust objectively is a question to address in future research.

Explanation tuning. It is a control feature in which users can tune or change specific explanation engine parameters or configurators to generate different explanation instances. We observed that lack of verification techniques had motivated our participants to propose several designs with a tuning functionality. P18 stated: *“Well ... it is not fair for an AI to explain in this way ... it would be useful to try different scenarios from explanations”*. As a general theme, the explanation output was not satisfying to our participants. They required an interactive and verification technique to motivate them in using the XAI interface. Our study identified three categories of explanation tuning:

1. Degree of similarity. This category was witnessed in Example-based explanation. It refers to tuning the example similarity to a preferred threshold. For instance, P17 commented on his design: *“In medicine, it is hard to define similarity. I would like to define it and amend it so I can judge the explanation in a better way”*.
2. What-if analysis. This category was witnessed in all explanation types. It refers to allowing users to ask what-if questions and directing them to verify the recommendation. In one scenario, when counterfactual explanation explained a recommendation of a patient to reject the prescription: *“The prescription would have accepted with 67% confidence if the smoking years=15 and Hormonal Contraceptives (years) = 13”*. P21 asked, *“Could the AI ignore Hormonal Contraceptives and re-explain how the prescription could be accepted”*.
3. Time frame. This category appeared in all explanation types. It refers to a technique to regenerate the explanation output based on data gathered in a specific time frame. Several participants discussed their concerns regarding the effect of outdated data on the explanation. For instance, in Example-based explanation, P20 asked to regenerate examples excluding data before 2010 as old examples might be misleading, e.g., healthcare and lifestyle change rapidly, and timely data about this is essential to judge the explanation and recommendation it explains.

Participants requests served a need for an interactive technique to help them in guiding the explainability process and meet different thresholds of their recommendation judgement. Our results are consistent with human-automation interaction literature which identified users' input into the intelligent system as an essential requirement for achieving trust calibration (Sokol and

Flach, 2020b, Lee and See, 2004). They argued that users' involvement could be crucial in the trust calibration process, especially when the underlying reasoning is dynamic and changing over time.

7.3.1.3 CUES

Cues refer to additional elements in the interface that can draw users' attention and guide them in understanding and reading the explanation (Aigner and Miksch, 2004). Users interact with a large amount of information every day and may occasionally fail to recognise and detect important details in the explanation. Also, in the long-term, explanations might become peripheral or checking them becomes habitual but not necessarily conscious and effective in drawing the users' attention to important details and nuances (Verplanken and Wood, 2006). Our participants sketched interfaces with two cues categories (visual and information) to guide them in the process of reading and quickly judging the explanation. For instance, for explanation reliability assessment, participants used an informational cue to denote the generalisability aspect of the explanation (See Figure 27). In the following sections, we present two categories of visual cues that resulted from our study.

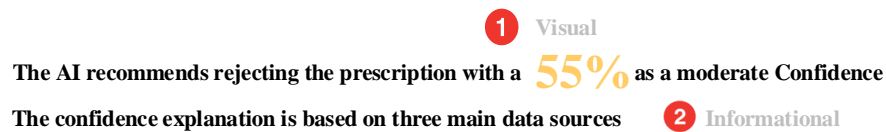


FIGURE 27 SUGGESTED VISUAL CUES RESULTING FROM THE DESIGN PHASE

Visual cues. Colour coding, font coding, shape coding, and directional cues were among the visual techniques used to help quickly read relevant information. Participants used visual cues to focus their attention on relevant information in the explanation interface and guide them in utilising it. For example, our participants used green/yellow/red colour schemes for faster perceptual judgment of the Confidence explanation value (“high”, “moderate” and “low”) (See Figure 6). Another example is when participants' sketches for Local and Global explanations revealed visual cues aligned with data features to provide more contextualisation and understandability. Three participants used a red/green colour scheme with a local explanation to distinguish between data features that positively influenced the recommendation (green) and those that negatively influenced it (red). Participants in Counterfactual explanation designs also used directional cues such as arrows and pointers to point out the direction of the changed data feature, i.e., in either increasing or decreasing the changed data feature value. In HCI literature, visual cues have been used to reduce overlooking specific parts of the interface for high critical tasks (Paterno et al., 2009, Chaffee and Zimmerman, 2010). Overall, visual cues could be an assistive design technique for explanation utilisation. It would help quickly understand the explanation and direct the user to different sections in the explanation interface. It aids their

trust calibration process by finding the relevant information and reducing fatigue and time needed. Visual cues should be defined in the design process with collaboration with end-users and HCI experts to identify key visual cues to be added to parts of the explainable interface.

Informational Cues. It is an indicator of the explanation quality. Decision-makers in uncertain and dynamic environments might lack the ability to process all the available information, e.g., explanations, so they use informational cues to assess its quality before reading it, e.g., data sources. Informational cues have been used in the literature as a signal of information quality (Grewal and Monroe, 1995) and increase users' ability to detect errors (Bass et al., 1996, Wickens, 1995). In the context of our study, participants used multiple information cues in their explanation designs to help them in judging the explanation quality and identify the extent to which they can trust it. P21 discussed using a timestamp informational cue with Example-based explanation and stated, *"doctors cannot use these examples in this way ... they would need to see when each example case has been prescribed"*. Participants discussed that they would be more willing to read the explanation when it can show its own quality. For instance, P24 used an expert validation informational cue with Local explanation to indicate that the explanation has been designed with medical expert validation. P24 mentioned, *"I thought this [informational cue] would be handy to make people trust the explanation"*. Likewise, P17 integrated an informational cue to the interface to inform about potential missing information in the patient profile, which might lead to an incomplete explanation. P17 argued: *"... in this way I would know that the explanation is incomplete"*. This is a common approach used in high-stakes decision support systems to increase the safety, speed and accuracy of the decision-making (Chaffee and Zimmerman, 2010). Our participants who had previous clinical information systems experience included informational cues in their explanation designs for effective trust calibration utilisation. Our findings are consistent with previous research in information processing to build trust (Zhang et al., 2009) showing that people subconsciously look for informational cues when deciding whether they can accept the presented information or reject it. Consequently, exploring additional informational cues for each explanation could enable better users' assessment explanation quality and facilitate effective explanation utilisation. Overall, excluding such information cues in the explainable interface development can limit the trust calibration process.

7.3.1.4 ADAPTATION

Adaptation refers to varying the explanation characteristics based on the interaction context. Adaptation is itself an intelligent decision and requires the development of the variability space of explainability options that are paired, either directly or through following some inference rules, to the task and personal states. As a general theme, participants felt that a well-designed adaptation technique could increase their perception of the explanation technical competence and encourage them to utilise the explanation. P11 stated: *"AI should be smart enough to know*

when to explain and how to explain”. We noticed two categories of explanation adaptation characteristics among participants' adaptation designs: interface complexity and interface content order.

Interface complexity. It refers to varying the complexity of the explanation interface to meet the collaborative decision complexity. P19 and P20 discussed situations where increasing the decision itself' complexity should trigger the need for a higher level of details in the explanation interface. P19 mentioned: *“I think it is unfair to explain this way, AI should consider presenting sophisticated examples and correlations in such complex case”*. Then P20 added: *“... with some patients, I don't need that much information to determine if the AI is right or wrong”*. Moreover, participants discussed varying the level of explanation interface complexity based on the certainty of the AI in the provided recommendation. Interestingly, they expressed a low level of trust in the absence of adaptation, in this case, P5 described: *“I would trust the AI more when it is not certain and provide a short explanation or just say I cannot explain ... no enough evidence”*. Overall, our participants expected the AI-based tool to understand the decision complexity and generate an appropriate explanation.

Interface content order. It refers to an intelligent decision to change the order of the explanation type or explanation content based on the decision context. Several participants discussed situations that required a variation in the explanation type order in the explanation interface in different contexts. For instance, P24 stated that when the patient has Human papillomavirus viruses (HPV), Example-based explanation has a higher impact on the decision. P24 stated: *“HVP is the most important risk factor for this type of cancer some of them can cause a type of growth called papilloma ... in such cases, I usually look for similar patients first”*. Our analysis also showed a pattern where three participants discussed scenarios for explanation content order adaptation. P20 discussed presenting the smoking effect in Local explanation first when the patient has a smoking habit. P20 mentioned: *“smoking is a high cause of accepting or rejecting such treatment when a patient does have that ... it should be at the top of the graph”*.

Our results are consistent with previous research (McCall and Trivedi, 2007) which showed that transparency that conflicts with the user experience and requirements of the task at hand might trigger cognitive confusion, mental overload and trust calibration errors. This means, explanation alone is not enough but rather its delivery and presentation should adapt to the user and task context to avoid that conflict. Also, our previous literature review revealed that explanation shall be adapted to users' level of knowledge and expertise in a given Human-AI collaborative decision-making task (Naiseh et al., 2020c). Similarly, previous studies showed that adaptive information increases collaborative Human-Robot performance (Torrey et al., 2006). Overall, the developers of explanation interfaces might require applying usability techniques like task analysis (Mills, 2000) to determine different contexts and scenarios where

adapting explanation is required. Incorporating task characteristics and contexts into the explanation delivery methods would likely increase the efficiency of explanation utilisation, and thus, improve trust calibration. Adaptation with XAI interfaces can also be used to meet the continuous development of the users' experience and meet their learning curve about the AI.

7.4 DISCUSSION

Human-AI collaborative decision-making tools are usually built to support trust calibration through the recommendations interface design (Berner, 2009, Scheepers-Hoeks et al., 2013, Huang et al., 2018). For example, users are typically enabled to control the level of assistance they want to get from an intelligent agent, so it provides important benefits to trust calibration, including improved situation awareness and more accurate Human-AI performance (Miller and Parasuraman, 2007). In this section, we expand the previous research by discussing five design principles that consider the interaction between the user and the XAI interface for improved trust calibration. We derived these principles as a reflection following a series of qualitative studies. This included the previously described studies of think-aloud and co-design sessions. The active involvement of XAI interface users and their lived experience facilitated identifying novel, interactive design concepts that would develop XAI interfaces that help trust calibration. As a conclusion of our qualitative studies, we provide more ecologically valid support for previous work on Human-AI collaborative decision-making, specifically when explanations are introduced. We tie our findings into a border discussion and principles on designing XAI interfaces to help trust calibration in the following sections. Although our principles provide guidelines for designing XAI interfaces to help trust calibration, calibrated trust XAI interface design may require a continuous improvement and monitoring in which the design is fine-tuned according to the feedback from the user, whether its explicit from the users or implicitly through behaviour indicators, such as a user is skipping explanations all the time. Hence, the XAI system will be able to infer the correlation between the required techniques, the explanation and the recommendation, which by time can become principles and lessons learned.

Design for engagement

During our Exploration phase, several participants skipped the explanation presented to them due to factors such as lack of curiosity and motivation. Participants did not feel that the explanation motivated them to learn new ideas, resolve knowledge and solve problems. Also, for participants who were interested in reading the explanation, they applied heuristics to interpret the explanation in their task. These findings provide a possible interpretation for why exposure to explanations did not improve trust calibration (Bussone et al., 2015, Zhang et al., 2020a). Furthermore, during the design phase, participants sketched up explanation structures that motivated them to read the explanation, e.g., abstraction and grouping. They were more motivated to engage with an explanation structure that represents a reduction principle of the

Persuasive Systems Design model (Torning and Oinas-Kukkonen, 2009), i.e. structures which present information in pieces and stages.

These results can be interpreted according to the Elaboration Likelihood Model of persuasion (Petty and Cacioppo, 1986). People follow two cognitive processing pathways: the peripheral route, which is fast and automatic and the central route, where people follow a slow and deliberative approach. People in everyday decisions tend to follow a peripheral route that employs heuristics and shortcuts to make decisions (Kahneman, 2011). This can limit the role of XAI to help trust calibration as engaging cognitively with AI-based explanation is triggered rarely. Such behaviour might be the main reason for skipping or misapplying the explanation. Indeed, successful trust calibration would require users to think and cognitively engage with the explanation. Therefore, an effective calibrated trust design may require XAI interface designers to increase users' tendency to cognitively engage with the explanation and trigger the central route. Such engagement is determined by individual factors such as the need for cognition or peoples' interest in the information (Petty et al., 1983).

Several approaches can be followed from other domains to analytically engage people with the provided explanations. One of the promising ways appears in applying the principles of herd theory (Banerjee, 1992). Herd cues can support users' engagement with the explanation by providing information about other users' interactivity actions with the explanation. For instance, a herd cue message can be represented as the following "several experts have used the explanation to judge the AI recommendation". Herd theory suggests that individual engagement will be influenced based on other actions in imitation-based behaviour (Sun, 2013). Understanding users' herd behaviour plays a critical role in influencing users' engagement in information systems and can be effective in promoting the desired user behaviour and interaction style (Tucker and Zhang, 2011). For instance, Barlow et al. (2018) showed that a herd cue message about the compliance of other users in changing their passwords increased the likelihood that users would comply with the same behaviour. The messages are to be framed carefully to avoid a situation where users become more inclined to follow the recommendations instead of following the behaviour of others of utilising the explanation to calibrate trust.

Design for challenging habitual actions

A series of similar former responses might form habitual actions (Ouellette and Wood, 1998). Habits can be triggered by environmental cues, such as time of the day; by internal states, such as individual mood; and by series of interactions with the same partner. Habits reduce sensitivity to minor changes in the explanation interface, curtail explanation utilisation, and reduce assessment and reflection about the decision (Cooper, 2002). For example, pharmacists who are in the habit of making decisions using AI-based screening prescription tools may fail to recognise incomplete explanations (Schrills and Franke, 2020). Previous experiment (Fazio et

al., 2000) exposed participants repeatedly to pictures of people to form well-practised reactions toward them. Participants were subsequently given the same pictures again but in a slightly amended version. Participants who had seen the faces repeatedly in the first part of the study had higher difficulty identifying the amendments and relied on their expectations formed during the prior exposures. These findings suggest that people with strong habits hold trust and expectations about the environment, which reduces their capacity to utilise the explanation and calibrate their trust.

During our co-design phase, participants frequently discussed taking an active role in controlling the explanation output to meet their reasoning process of everyday decisions – including the presentation and phasing, e.g., P12 mentioned: “... *I would like to exclude patient age from future explanations*”. Although such techniques could help users in their trust calibration and meet their user experience, however, they might be prone to develop habitual actions, which results in a failure in utilising explainability for trust calibration goal. Furthermore, in the exploration phase, several participants were gradually less interested in the explanation details and started to overlook the explanation. Hence, an effective design for a calibrated trust may need to consider challenging users’ from developing with the explanation interface. Research in psychology suggested two different habits challenging approaches which are *downstream* and *upstream* (Verplanken and Wood, 2006). The downstream approach focuses on the individual level of intervention and adopts strategies such as education, stimulus control and other behavioural modification strategies. One example of applying a downstream approach in XAI interface could be through developing educational material to show the benefits of explainability in calibrated trust. Educational material can also be used to inform users about the costs of undesired behaviour with the explanation and increase self-efficacy to perform the desired behaviour. In contrast, the upstream intervention approach targets more extensive structural conditions where peoples’ behaviours are embedded. For instance, the upstream approach might change the structure of the XAI interface where the explanation is presented before the recommendation. This approach aims to provide a structure that promotes desired behaviour, i.e., presenting the explanation before the recommendation would increase the likelihood of reading it. These interventions in psychology literature gained their effectiveness because it renders people with strong habits open to new information (Wood et al., 2005). We recommend future work on designing XAI for long-term explainability to focus on user experience for enabling usable explanation utilisation but also combat developing habits using approaches like the downstream and upstream intervention methods.

Design for attention guidance

In both phases, we observed that participants needed the XAI interface to support them in reading the explanations. During the exploration phase, participants felt that the explanation was complex, and they were selective in what to read and rely on. This finding can be interpreted by

the fact that human visual perception is selective (De Koning et al., 2009). Participants focused their attention only on a little number of elements in the interface and those only small portions of explanation content were processed. Such observations motivated our participants, in the design phase, to suggest design techniques and requirements to support them in reading the full explanation in a usable and user-friendly way, such as abstraction and visual cues. Our results from both phases suggest that helping trust calibration in the XAI interface design could be further enhanced by applying the principles of attention guidance (Amadiou et al., 2011). The role of the attention guidance principle could be critical in the trust calibration process when users' desired behaviour is to look for relevant content in the explanation and combat overlooking it. Utilising the explanation for the trust calibration goal would also expect to guide the user attention from one element to another and support them to determine the next element in the XAI interface. The main goal of such principles is to increase the amount of the processed explanation content by the users and thus improve their understanding of AI reasoning. Such a principle has been widely used in learning environments to help learners develop an improved understanding of the presented information (De Koning et al., 2009, Bradley, 2013). When designing attention guidance, careful consideration shall be paid to whether the guidance itself lead to biased processing and becomes persuasive and lead users to neglect or follow the recommendation.

Design for friction

Across the Exploration phase, several participants who worked collaboratively with an AI were more interested in the decision-making task itself rather than the reading and understanding the explanation. This can be seen in categories such as perceived goal embedment and rush understanding (Table 3). Participants developed a negative attitude towards explanations perceived to distract them from their main task. Moreover, participants frequently discussed or designed interfaces to accelerate their task completion time, e.g., abstraction designs. These results could be interpreted as avoidance or refusal behaviour. Our observations support previous work which found that people working collaboratively with an AI agent are not willing to engage in what they perceived as effortful behaviour with the AI explanations (Eiband et al., 2019b, Duff and Faber, 2011). Overall, during both qualitative phases, the degree of willingness to read the full explanation was low, specifically when participants discussed using the explanation for their everyday decision-making tasks. Thus, the explanation might fail to support users in their trust calibration process due to factors such as avoidance or refusal.

This finding shifts from usable design to friction design for calibrated trust when designers for explainable interfaces might adopt techniques to combat explanation avoidance. Friction design is user experience defined as interactions that hinder people from painlessly achieving their goals when interacting with technology (Mejtoft et al., 2019). For instance, explanation interface designers might use anticipating possible errors technique which considers user

performance as a metric and warns whether an action might cause a problem, e.g., a user spent a short time reading the explanation (See Figure 28). Another technique could be increasing the steps of making the final decision where the explanation is presented during multiple steps. The active obstruction and forced delay of the task completion caused by such techniques are likely to generate an enhanced understanding by slowing down the speed of users' actions. These techniques have shown an effective way to increase users' level of understanding and enable consciousness when interacting with the presented information (Mejtoft et al., 2019, Campbell et al., 2017). Overall, slowing users down can facilitate a reflection on their actions and this might be crucial for effective calibrated trust explanation utilisation. However, the obtrusiveness of such techniques can lead to reactance (Brehm and Brehm, 2013), i.e., the user feels that their freedom of choice has been taken away from them and they become more inclined to reject the XAI altogether.

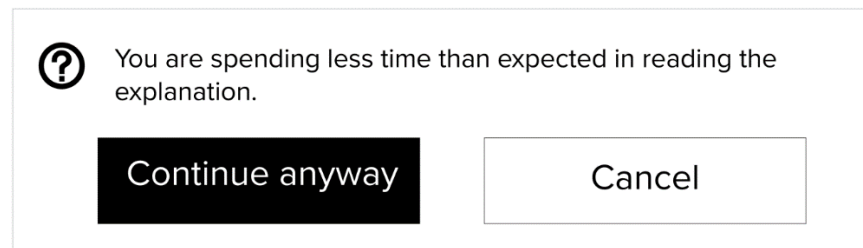


FIGURE 28 FRICTION DESIGN EXAMPLE FOR CALIBRATED TRUST GOAL

Support training and learning

During the Exploration phase, some participants failed to apply the explanation in their decision-making task due to reasons related to lack of familiarity and knowledge. They misinterpreted the explanation, mistrust the provided explanation or looked for confirmatory information. This finding relates to earlier work in Human-AI trust calibration (Cai et al., 2019b), which showed that training clinicians to use the AI-based decision-making tool reduced users' errors and increased the Human-AI collaborative decision-making performance. The training included learning optimal ways of using the AI in their settings and pointing out possible errors (Cai et al., 2019b). Thus, relative to Human-AI interaction research, human-explanation calibrated trust design may need to facilitate correct explanation interpretation and learn optimal usage scenarios before using XAI interface.

Furthermore, we observed that several participants' designs considered the explanation interface as a learning interface, where users can learn from the explanations. Participants used several interactive techniques, such as tuning and grouping, to generate several instances and observations from the explanations and compare them together. This helped them to extract new knowledge from the explanation and improve their mental model. Participants' designs tended to encourage users to look for alternative options to learn from them. For instance, P19 designed the Confidence explanation as a way of drawing her attention to missing actions, which the AI could provide to them through a large amount of processing data and processing power.

Specifically, she wanted to know how the AI thinks its certainty can achieve higher values, e.g., asking the doctor to request a further blood test to increase the AI decision accuracy. Some participants also designed explanation interfaces with hyperlinks to medical databases and research references. Their expectations of the XAI interface can be strongly anchored to their prior experience; this can sometimes require the XAI interface to provide a systematic and argumentative discussion.

We also recommend future work to apply principles of self-learning to facilitate users' learning process from the explanation interface (Hadwin and Winne, 2001). As such, self-learning can improve users' learning from the XAI interface and refresh their knowledge about AI for the most effective way to calibrate trust. It would be likely to increase the chance that they attain the desirable trust calibration behaviour over time. These principles include allowing users to write notes about the explanation, link them together, archive explanations for future comparisons and share their explainability experience with other system users.

7.5 CONCLUSION

The dynamic nature of AI-based decision-making tools poses new requirements for developing interfaces with a calibrated trust goal in mind, specifically when explanations are presented. In this paper, we have conducted a qualitative approach that provides a detailed look at explainability and trust calibration. Our approach consisted of two qualitative phases: (a) exploration phase, which aims to provide a contextual understanding of the problem; (b) design phase to reveal main concepts and designs techniques that improve the role of explanation in trust calibration. Our work presents a broad view of Human-AI collaborative decision-making tools and raises important questions for future work. In particular, our design implications point towards supporting the interface design with techniques and principles to increase users' interaction with the XAI interface to help trust calibration. For example, our results suggest that presenting explanations for trust calibration should mainly be designed to avoid undesired behaviour such as skipping explanation and habits formation. This is in line with what we have proposed although at this stage of our research we are unable to pair between the configurations of the XAI design technique and the type of bias and error. Future work shall focus on the balance between making explanation effective enough in trust calibration and, simultaneously, avoiding the potential harm to user experience and being seen as a persuasive tool instead of critical thinking aid. Such neutrality in the recommendation as well as ensuring a reflective and measured reasoning in the users can be hard to achieve, especially that other requirements like engagement and design for friction are also in place.

8. CHAPTER 8. C-XAI: A METHOD FOR DESIGNING XAI INTERFACES TO HELP TRUST CALIBRATION

The C-XAI method is developed based on the research findings of previous chapters. Four main phases representing the main design stages that shall matter to the XAI interface design team are presented. The first two phases focus on analysing and assessing the role of explanations in helping trust calibration during a Human-AI task. The first two phases help anticipate potential trust calibration risks that may emerge during the Human-AI collaborative decision-making task. The third and fourth phases represent the XAI interface design stages. The design team needs to address the issues and risks identified in the first two stages. In this chapter, a description of each phase of the design method is presented. The chapter also gives a complete list of activities and guidelines for the system analyst and design team to help the design process. Finally, good design practices to help guide the team when designing the XAI interface for calibrated trust goal are provided.

8.1 C-XAI METHOD OVERVIEW

Explaining recommendations generated from black-box models gained much interest in Human-AI interaction in which humans can understand AI reasoning and calibrate their trust. However, recent evidence from the literature showed that users, on average, still fail to calibrate their trust even though explanations are communicated to humans (Zhang et al., 2020a, Bussone et al., 2015, Naiseh et al., 2020b). In this thesis, the researcher discussed that XAI interface designers often assumed that users, during Human-AI collaborative decision-making tasks, will engage cognitively and analyse the proposed explanation to calibrate their trust. **Chapter 5 and Chapter 6** showed that XAI interface might be at risk of failing to help trust calibration due to either humans' errors or explanation technical properties, e.g., the generated explanation is inconsistent. Such findings motivated further work in **Chapter 7** to propose XAI design principles and techniques to help the design team mitigate different users' errors. **Chapter 7** also showed how the design shall consider different explanation properties to mitigate potential trust calibration risks, e.g., designers might use grouping and ordering techniques for complex and long explanations. However, several questions could arise during the design process, such as *“Are all principles and techniques equally important?”*, *“Are they compatible with each other?”* and *“Which principles are needed for a particular XAI class and method?”*.

This chapter proposes a method to help designers identify appropriate design principles and techniques relevant to a Human-AI task and its explanations. Such a method is needed due to the fact that XAI methods in the literature vary in their technical properties, and thus, their implementations to help trust calibration may vary as well (Sokol and Flach, 2020a). For instance, perceived complexity as a trust calibration risk has a minor effect considering counterfactual explanation classes. Counterfactual explanations are designed to be human-

friendly explanations and target non-expert users (Wachter et al. 2017). A method will be devised accordingly to enable the design team to select appropriate design principles based on the underlying XAI method and their corresponding trust calibration risks.

The method is for a design team that would like to augment their XAI interface with a new design layer to support users' trust calibration process. The method will follow a particular outline that would help the system analyst and the design team locate trust calibration risks that need to be addressed in the design. During the identification process, representative users should be recruited and supported to specify required explanations in a given Human-AI task. Then, the assessment process aims to assess explanations' role in supporting users' trust calibration process, including explanation technical properties and their potential trust calibration risks. The assessment process's outcome will enable the design team and system analyst to identify potential trust calibration risks for the Human-AI task explanations. For example, explanation generalisability as a social property for the LIME explainable method cannot be achieved, i.e., LIME explanations cannot be generalised beyond specific AI recommendations. Hence, users of LIME explanation might be at risk of forming incorrect interpretations and generalising the explanation among many recommendations. In such a scenario, a learning design principle shall be prioritised in the design process to mitigate the risks of explanations' incorrect interpretations. In the same context, using the learning design principle might not be appropriate for counterfactual explanations. Counterfactual explanations are human-friendly explanations and easy to interpret. The process is expected to be iterative, where subsequent meetings will be scheduled to re-evaluate the design and potential trust calibration risks. In this chapter, system analyst and analyst will be used interchangeably.

Various qualitative studies and literature reviews were performed to create the C-XAI method, such as semi-structured interviews, think-aloud protocols, and co-design sessions. The literature reviews and studies conducted in this thesis show that when users talked about explanations from AI-based systems, they talk about a set of requirements and answers to their questions in a specific context and specific task. They expect the explanation to be smart enough to analyse the context and generate an appropriate explanation. The findings from these studies and literature reviews discussed in the previous chapters of this thesis form the C-XAI method elements depicted in Figure 31. In **Chapter 5**, an investigation around the effect of different explanation classes on the trust calibration process is discussed. **Chapter 5** has argued that the explanation class choice for calibrated trust is highly dependent on the nature of the Human-AI task and users' needs. **Chapter 6** aimed to investigate XAI users' lived experience and explore how users interact with AI-based explanations. The chapter revealed several trust calibration risks that need to be addressed in the XAI interface. **Chapter 7** aimed at exploring different design principles that can be applied to help users' trust calibration process. The findings from **Chapters 5, 6, and 7** fed into the design and creation of the C-XAI phases, templates,

guidelines and supporting documents. The template elements would help the design team to assess the explanation role in the trust calibration and understand the design problem. It will also help identify appropriate design principles for the XAI interface to help trust calibration. This chapter will commence by providing an introduction to the C-XAI method. The chapter will then explain each of the stages (See figure 29) and the methods' activities (See figure 30). Finally, the chapter will describe the method workflow and activities involved in each stage of the method. The final version of the C-XAI method can be found in [<https://bit.ly/3gmdPpy>].

8.2 CHAPTER RESEARCH GOAL

This chapter aims to propose a design method to help XAI designers in developing calibrated trust XAI interfaces. Such a design process would help them identify the potential trust calibration risks. This identification would ultimately help designers in following appropriate design principles to mitigate trust calibration risks. The method will follow a particular outline that would help C-XAI users from a multi-disciplinary background such as system analysts, AI experts, and psychologists to assist the design process. C-XAI method adopts a participatory design approach to ensure that all relevant stakeholders are actively involved in the earlier stages of the XAI interface design. The C-XAI method includes four main phases (See Figure 29), supported by templates indicated by (**SHEET ID 1, 2, 3,4,5,6**). C-XAI method also provides users with supporting documents to help them complete each activity.

FIGURE 29 STAGES OF THE CAT METHOD



FIGURE 30 C-XAI METHOD ACTIVITIES

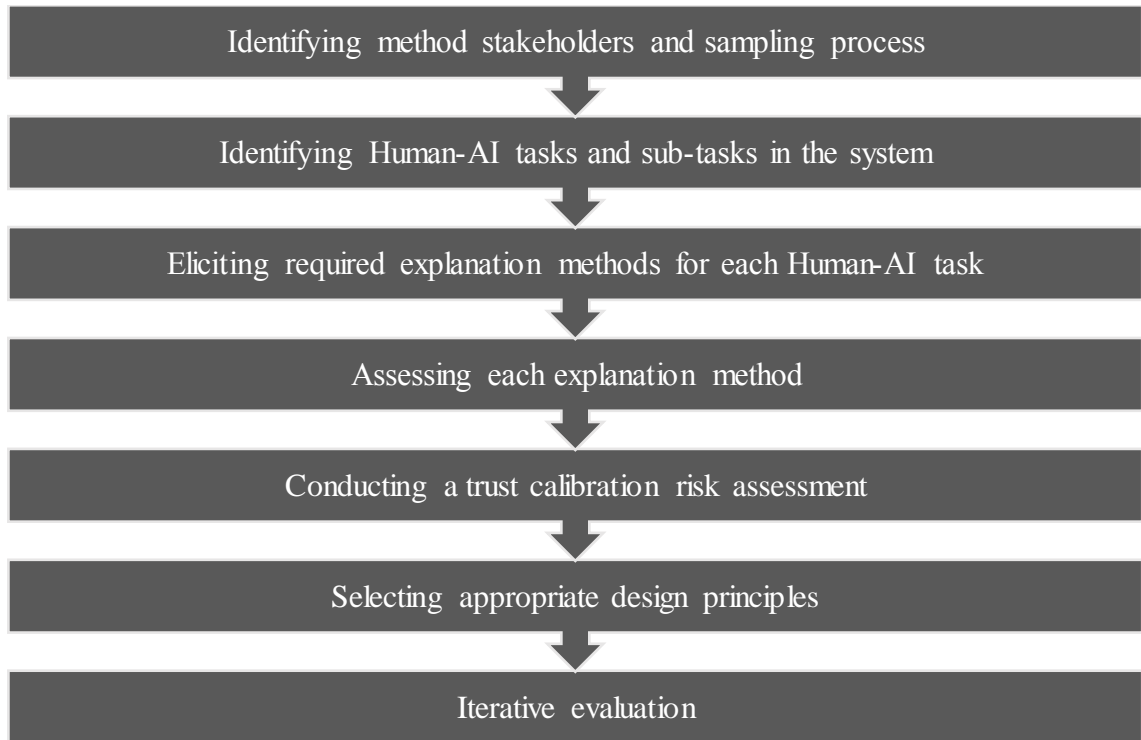


TABLE 23 DESCRIPTION FOR METHODS' ACTIVITIES

Process building blocks	Description
Identifying method stakeholders and sampling process	It refers to inviting representative users' and other stakeholders.
Identifying tasks and sub-tasks in the system	It refers to an analysis of the system to identify tasks and sub-tasks where AI recommendations are issued to Humans' decision-makers.
Eliciting required explanation method for each task	It refers to an elicitation process for the required explanation methods for each task.
Assessing each explanation method	It refers to a systematic assessment for each of the selected explanation methods.
Conducting a trust calibration risk assessment	It refers to anticipation and analysis of various risks that can limit users' trust calibration process. These risks can be anticipated based on the input from the explanation method assessment.
Selecting appropriate design principles	It takes the trust calibration risk assessment and determines principles that are capable of

	mitigating the risks.
Iterative evaluation	As each design principle is implemented, it must be evaluated in the context of the task. This might lead to iterating through the selection and evaluation phases.

8.3 PHASE 1: IDENTIFICATION PROCESS

The identification phase is the initial part of the method. It includes eliciting explainability needs for each Human-AI collaborative decision-making task in an AI-based decision-making system. This stage also includes a declaration for several sub-tasks where the AI-based recommendations are issued to end-users. Also, the identification phase requires identifying and inviting different stakeholders to participate in the XAI interface design. Such participation will provide the method with knowledge about the target explanation audience and their explainability needs in a given task. The following sub-sections discuss each sub-stage of the identification phase.

8.3.1 IDENTIFYING METHOD STAKEHOLDERS AND SAMPLING PROCESS

This sub-stage involves identifying various stakeholders to take part in the design method. The stakeholders include those who will help ensure that the XAI interface can help users calibrate their trust and mitigate any potential trust calibration risks. Before starting the design method, the system analyst is expected to identify the stakeholders who will participate in the design process. A set of stakeholders are suggested in Table 24. Some stakeholders will be directly involved in the entire design method, while others participate in some of the method phases.

TABLE 24 C-XAI METHOD STAKEHOLDERS

Stakeholder	Description	Degree of participation
Representative users	It refers to users who will use the assisted AI-based decision-making tool.	Representative users are expected to involve in the identification and evaluation phases of the CAT method.
System analyst	It refers to the person who would guide the design method and gather the collected requirements. The system analyst is expected to be an expert in software engineering and has adequate experience in AI, human factors and usability.	The system analysts' role is to guide the design process and be actively involved in all design stages.
AI experts	It refers to the person who is responsible for technically assessing the required explanation for each Human-AI task.	AI expert is expected to be involved in the identification and assessment phases.

Design team	It refers to a group of people who are responsible for designing the XAI interface.	They are involved in the assessment, implementation and evaluation phases.
Domain expert	It refers to the people who have experience in the Human-AI task, e.g., an experienced doctor for a cancer detection task.	They are responsible for analysing and evaluating potential constraints and requirements for explanations in the Human-AI task, e.g., the task requires temporal and terminological explanations.
Psychologist	It refers to people who have psychological knowledge and relevant background on cognitive biases, human behaviour and decision-making theories. They are responsible for helping the design team understand the users' psychological state and their decision-making strategies.	They are fully involved in all the stages from the design method from the identification process to evaluating the generated explainable interfaces.

8.3.1.1 RECRUITMENT PROCESS OF REPRESENTATIVE USERS

Representative users shall be approached and recruited to elicit their explainability needs during their Human-AI collaborative decision-making task. During the recruitment process, the system analyst is encouraged to look for diverse users to increase the collected needs' validity and credibility. Various methods can be adopted during the recruitment process, for example, convenience samples via organisational mailing lists. Based on the previous user studies in **Chapters 4,5,6**, and previous work in the literature (Eiband et al., 2018a, Hoffman et al., 2018), inclusion and exclusion criteria shall consider the factors presented in Table 25.

TABLE 25 CRITERIA AND RATIONALE FOR STAKEHOLDERS' RECRUITMENT

Criterion	Rationale
Role	Users' role in the system would enable the elicitation of different views or preferences around the explainability needs for XAI interfaces, e.g., medical doctors and nurses for a clinical decision support system.
AI knowledge	AI knowledge is an attribute, which would enable gathering different views and preferences that may affect the acceptance and the need for explainability, e.g., a low level of AI knowledge may require meta-information to understand the explanation.
Domain knowledge	This aspect would show the diverse explainability needs for multiple levels of knowledge in the studied task, e.g., experienced doctors in a clinical decision support system may have different explainability needs than novice doctors.
Curiosity level	The nature of seeking explanations is driven by curiosity. Considering this aspect in the recruitment phase would enable the analyst to elicit diverse explainability needs. To help address this criterion, the method

provides a curiosity checklist adopted from (Hoffman et al., 2018) in Figure 31.

Why would you ask for an explanation? Check all that apply	
	I want to know why the AI is recommending a decision.
	I want to know that I understand the AI system correctly.
	I want to understand what AI will do next.
	I want to know why the AI did not make another decision.
	I want to know what the AI would have done if something had been different.
	I want to know if I am missing any information about AI.

FIGURE 31 CURIOSITY CHECKLIST

After recruiting representative users and identifying the stakeholders, the analyst can conduct an induction session around the expected exercises during the design method. The system analyst can also use the supporting documents, templates and material. During the induction session, the system analyst is encouraged to clarify that some of the templates shall be completed through a discussion between the analyst and other stakeholders. On the other hand, some templates can be completed by the stakeholders without intervention from the system analyst. The induction sessions would help familiarise different stakeholders with the relevant documents and activities during the design session and stimulate their thinking. The system analyst can consider two options for the induction sessions. One is to conduct a short focus group separate from the design method session where all the supporting documents are presented to the stakeholder and discussed with them. The other option is to conduct the induction session at the beginning of each activity with the relevant documents, e.g., 30 minutes before the activity.

8.3.2 IDENTIFYING HUMAN-AI COLLABORATIVE TASKS

After identifying the stakeholders and recruiting representative users, the system analyst needs to identify the various tasks in the system involved in issuing AI-based recommendations to end-users. The method proposes a task analysis sheet called the Human-AI task Analysis (HAI-A) to assist this process, as shown in Figure 32. It is similar to traditional task analysis techniques (Annett, 2003, Crandall et al., 2006). Nevertheless, it is extended to support explainability needs analysis in further stages.

Task	Subtasks	XAI class	XAI method	Additional needs
------	----------	-----------	------------	------------------

Task1	Subtask 1	Class A	E1	
		Class B	E2	

FIGURE 32 HAI-A ANALYSIS SHEET

8.3.3 ELICITING USERS' EXPLAINABILITY NEEDS FOR EACH TASK

The XAI interface's effectiveness to calibrate users' trust depends on two major design decisions (Herlocker et al., 2000, Eiband et al., 2018a): *explanation content and explanation presentation*. Overwhelming the end-users with irrelevant explanations would decrease the efficiency of utilising the interface and likely decrease users' willingness to interact with the explanation (Gregor and Benbasat, 1999). Findings in **Chapter 5** showed that users explainability needs during a Human-AI task vary based on the task itself and target users. There is no standard answer for "*What to explain to users during a collaborative Human-AI task*". During this stage, the analyst is expected to seek explainability needs for each of the identified tasks in the HAI-A Sheet. This would allow the system analyst and the AI experts to choose XAI classes and methods that fit users' needs. This stages' outcome is the list of XAI classes and methods that fit users' needs to calibrate their trust for a given Human-AI task. After that, the AI expert's role is to choose appropriate explanation methods from the identified XAI classes.

The method follows the guidelines provided by Eiband et al. (2018) to identify the explainability needs in a given task. Their research showed that eliciting explainability needs consists of two main stages:

1. **Intrinsic elicitation.** This stage answers the main questions of what can be explained to end-users given a black-box model. It refers to key XAI classes and methods that can be used by ML engineers and data scientists to generate explanations.
2. **Users' needs elicitation.** It refers to users' explainability needs for a Human-AI task.

8.3.3.1 INTRINSIC ELICITATION

System analysts shall collaborate with an AI expert to identify what can be explained to end-users given a black-box model. As the C-XAI method is focused on model-agnostic explanations, it provides a supporting document that covers an intrinsic elicitation for model-agnostic explainable models. Explanation by definition is an answer to a user question (Liao et

al., 2020). The intrinsic elicitation supportive document describes the diversity of the questions that each explanation class can answer. The term explanation class refers to a family of explanation methods that generate similar explanation content. The explanation method or XAI method refers to an explanatory model that generates an explanation from a black-box model. For instance, LIME (Ribeiro et al., 2016b) is an explanation method that belongs to the Local Explanation class. This stage aims to answer a common question “*What can be explained to users in a given Human-AI task?*”. The supportive document is based on a previous systematic literature review on the model-agnostic explainable model (See Figure 33) and supported by previous work (Liao et al., 2020). System analysts can use this document as a reference point for answering what can be explained to end-users.

SHEET ID: 1		SHEET TITLE: INTRINSIC ELICITATION
Main class	Sub-class	Question – Content - Scope
Global explanations	Global feature importance	<i>Ranking the data features</i> Why – Importance - Model
		<i>Correlation between features</i> Why – Dependences – Model
	Decision tree approximations	Why – trace – Model Why not – trace – Model How – trace – Model What if – trace - Model
	Rule extraction	<i>AND-OR rules</i> How – trace – Model What if – trace - Model Why – trace – Model Why not – trace – Model
		<i>IF-ELSE rules</i> How – trace – Model What if – trace - Model Why – trace – Model Why not – trace – Model
Local Explanations	Local feature importance	Why – importance – recommendation
	Local Trees	How – trace – recommendation What if – trace - recommendation Why – trace – recommendation Why not – trace – recommendation
Example-based	Prototype	When – exemplar – recommendation What else – exemplar – recommendation
	Counterfactual	When – exemplar with small changes – recommendation What else - exemplar with small changes – recommendation
	Influential	When – abstract exemplar – recommendation What else – abstract exemplar – recommendation
Counterfactual	Feature Influence	What if – Influence - recommendation
	Counterfactual features	When – influence – recommendation Why – influence – recommendation How to – influence - recommendation

Confidence		How accurate – uncertainty - recommendation How accurate – uncertainty – model
------------	--	---

FIGURE 33 INTRINSIC ELICITATION FOR MODEL-AGNOSTIC EXPLANATIONS

8.3.3.2 USERS' NEEDS ELICITATION

This stage shall focus on users' needs during a Human-AI task and consists of two steps: (1) elicitation of the users' needs (2) comparing users' needs and intrinsic elicitation to identify similarities and differences. Table 26 presents methods that have been used to elicit users' needs in domains such as knowledge acquisition for expert systems, instructional design, cognitive psychology and educational computing. Table 25 also lists some of the methods' strengths and weaknesses. These methods have been adopted from a previous survey in users' mental model elicitation (Hoffman et al., 2018).

As it has been discussed in **Chapter 5**, the gathered requirements in this stage shall focus on identifying the required XAI classes for a Human-AI collaborative decision-making task. Also, **Chapter 5** showed the calibrated trust is a systematic design approach starting from building appropriate trust to calibrate users' trust during the Human-AI task. This means that the analyst may need to collect requirements and needs before using the XAI interface, during the interaction time, and after presenting the explanation. C-XAI method provides a list of good practices during the users' needs elicitation stage (Section 8.7).

TABLE 26 USERS' NEEDS ELICITATION METHODS

SHEET ID: 2	SHEET TITLE: Users' needs elicitation methods
1	<p>Method: Think-aloud problem solving, where participants think aloud during a decision-making task.</p> <p>Strengths: It offers rich information about the users' mental models.</p> <p>Weaknesses: The process of transcription and data analysis might be time-consuming.</p> <p>References: (Woods and Roth, 1988)</p>
2	<p>Method: Think-Aloud protocol with specific question and answering activities.</p> <p>Strengths: It enables the researcher to target specific issues during the decision-making task.</p> <p>Weaknesses: It might introduce bias as it depends on the researcher skills in designing the study.</p> <p>References: (Woods and Roth, 1988)</p>
3	<p>Method: Card Sorting based on the semantic similarity between the domain concepts</p> <p>Strengths: It enables the researcher to understand the relation between different domain concepts.</p>

	<p>Weaknesses: The collected data might be sparse about the events or processes.</p> <p>References: (Van der Veer and Melguizo, 2003)</p>
4	<p>Method: Nearest Neighbour task, where the participants select the best explanation that fits their task.</p> <p>Strengths: It can provide a quick understanding of the users' mental models.</p> <p>Weaknesses: It might be prone to the phenomena of Illusion of Explanatory Depth; people overestimate their understanding of complex systems.</p> <p>References: (Hardiman et al., 1989)</p>
5	<p>Method: Self-explanation task, in which participants are presented with several AI-based recommendations and are asked to explain these recommendations.</p> <p>Strengths: It can provide quick access to users' mental models.</p> <p>Weaknesses: It requires a clear rationale for the choice of the AI-based recommendations to be the focus of the users' mental model elicitation task.</p> <p>References: (Ford et al., 1993)</p>
6	<p>Method: Glitch Detector Task, in which participants are asked to identify the strengths and weaknesses in each of the available explanations.</p> <p>Strengths: It can support users to freely express their mental model that might be incorrect.</p> <p>Weaknesses: The glitches shall be built-in in the design. Also, knowledge shields may reduce the awareness of the glitches.</p> <p>References: (Hoffman et al., 2001)</p>
7	<p>Method: ShadowBox task, in which participants compare their understanding of the system to the expert explanation.</p> <p>Strengths: It can provide quick access to users' mental models.</p> <p>Weaknesses: Participants may not be able to understand the expert explanation.</p> <p>References: (Klein and Borders, 2016)</p>

After identifying the required explanation classes that fit users' mental models, the AI expert is expected to complete the HAI-A Sheet to indicate potential explanation methods for each task.

8.4 ASSESSMENT PHASE

C-XAI method is not only for eliciting the explainability needs for a Human-AI collaborative decision-making task but also intended for assessing the role of the explanation in supporting users' trust calibration process. Therefore, the assessment process's objective is to help the design team identify potential trust calibration risks given the output from the Identification phase. This phase involves two main steps: (i) Technical assessment of each XAI method, (ii)

trust calibration risk assessment for each XAI method. **Sheet ID 3.1-3.5 and 4.1-4.5** to be completed in this stage before the selection and implementation process.

8.4.1 ASSESSING THE EXPLANATION METHOD

XAI methods hold a diverse range of technical properties and features that may require attention at XAI interface design level (Sokol and Flach, 2020a). Neglecting such properties during the design facilitated several trust calibration risks introduced in **Chapters 5 and 6**. Following a systematic assessment for the selected XAI method would help anticipate potential trust calibration risks and select appropriate design principles in the XAI interface to mitigate them. For instance, when considering that users might develop habits with the XAI interface, i.e., people become gradually less interested in the details of explanation and overlook and perceive it to be familiar to them. This risk is critically important when XAI methods' novelty is high; high probability of generating new information to users each time. Good design practice in such cases is to highlight the new information in the XAI interface to guide users' attention and challenge users' habits.

For this purpose, C-XAI method provides assessment sheets for XAI methods based on Explainability Facts Sheets (EFS) framework (Sokol and Flach, 2020a). System analyst with a collaboration with AI experts is expected to complete **SHEET ID 3.1 – 3.5**. The assessment has five main dimensions:

1. **Functional.** This dimension can help to evaluate whether a particular XAI method is suitable for the underlying ML algorithm used in the Human-AI task.
2. **Operational.** This dimension can support designers in understanding how users can interact with an explanation.
3. **Usability.** It helps to evaluate the XAI method based on theories of explainability in social sciences.
4. **Safety.** Explanation models communicate partial information about the data set used to train the AI-based system. Safety dimension evaluates the effect of the XAI method on the security and privacy of the AI-based system.
5. **Validation.** It reveals methods of evaluation have been done in the literature to evaluate the assessed XAI method.

SHEET ID: 3.1	Functional Requirements assessment
TASK ID:	Explanation method:
Which family of machine learning approaches are applicable for the XAI method?	
<input type="radio"/> Unsupervised learning <input type="radio"/> Supervised learning <input type="radio"/> Reinforcement learning <input type="radio"/> Other: -----	
Which family of a problem type is assigned to the XAI method?	

<ul style="list-style-type: none"> ○ Binary classification ○ Multi-class classification ○ Multi-label classification ○ Probabilistic classification
SHEET ID: 3.2 Operational Requirements assessment
TASK ID: Explanation method:
What is the explanation class related to this XAI method? <ul style="list-style-type: none"> ○ Local feature associations ○ Global feature associations ○ Example-based ○ Counterfactual ○ Confidence score
How can the explanation be communicated ? <ul style="list-style-type: none"> ○ Statistical, e.g., weights for the features ○ Visualisation, e.g., Plots and charts ○ Textualisation; natural language ○ Mixture of the above
What is the explanation interaction method? <ul style="list-style-type: none"> ○ Static ○ Interactive
What are the explanation audience knowledge requirements? <ul style="list-style-type: none"> ○ AI experts ○ Task experts ○ Lay audience ○ Other -----
What is the goal of the XAI method? <ul style="list-style-type: none"> ○ Transparency of the ML model ○ Fairness of the ML model ○ Accountability of the ML model ○ Other -----
Does the explanation method have a casual nature ? <ul style="list-style-type: none"> ○ Yes

<input type="radio"/> No
Does the explanation method have an actionable nature ?
<input type="radio"/> Yes <input type="radio"/> No
Are there any other assumptions or operational aspects related to the explanation method?

SHEET ID: 3.3	Usability Requirements assessment
TASK ID:	Explanation method:
How truthful is the XAI method with the underlying ML model?	
<input type="radio"/> Low <input type="radio"/> Medium <input type="radio"/> High	
How well can the generated explanation be generalised beyond a particular recommendation?	
<input type="radio"/> Low <input type="radio"/> Medium <input type="radio"/> High	
Describe any contextual information that the explanation shall accompany; This could be any information needed by the users to interpret the explanation correctly.	
Does the XAI method allow interactive interaction?	
<input type="radio"/> Yes <input type="radio"/> No	
Does the explanation have an actionable nature?	
<input type="radio"/> Yes <input type="radio"/> No	
Does the explanation inherent time order of the event?	
<input type="radio"/> Yes <input type="radio"/> No	
Does coherence implemented in the XAI method?	
<input type="radio"/> Low <input type="radio"/> Medium <input type="radio"/> High	
What is the degree of XAI method novelty ?	
<input type="radio"/> Low <input type="radio"/> Medium <input type="radio"/> High	
What is the degree of complexity inherited in the generated explanation?	
<input type="radio"/> Low <input type="radio"/> Medium <input type="radio"/> High	
Are there any other assumptions or operational aspects related to the explanation method?	

FIGURE 36 EXPLANATION METHOD USABILITY ASSESSMENT SHEET

SHEET ID: 3.4	Safety Requirements assessment
TASK ID:	Explanation method:
What is the information leakage assessment done to the XAI method?	
Describe scenarios where the explanation can be misused ?	
Describe scenarios where the explanation method can be unstable or inconsistent ?	
Are there any other assumptions or operational aspects related to the explanation method?	

SHEET ID: 3.5	Validation Requirements assessment
TASK ID:	Explanation method:
Describe how the explanation method has been evaluated in user studies ?	
Describe how the explanation method has been validated in synthetic experiments ?	

FIGURE 38 EXPLANATION METHOD VALIDATION ASSESSMENT SHEET

FIGURE 37 EXPLANATION METHOD SAFETY ASSESSMENT SHEET

8.4.2 CONDUCTING TRUST CALIBRATION RISK ASSESSMENT

This C-XAI stage aims to help the design team point out which XAI method's properties could trigger a trust calibration risk. Trust calibration risk is defined in this thesis as a limitation in the interface design that may cause users' errors and limit users' trust calibration process. Trust calibration risk assessment is defined as the significance and the acceptability of the risk probabilities and consequences to the trust calibration. Such assessment has shown to be an effective approach to identify the design requirements in a requirement elicitation process (Boehm, 1989). This stage involves a collaboration between system analysts, domain experts, AI experts and psychologists. Conducting risk assessment would enable the analyst and the design team better to understand the applicability of XAI method in the Human-AI task. It is also for anticipating potential trust calibration risks with the XAI interface. For instance, the XAI methods' incompleteness could introduce a rush understanding trust calibration risk as users may form incorrect interpretations for the explanation.

Risk assessment is not intended to be predictions of the users' behaviour or system functionalities, but it helps to identify and assess potential risks (Armstrong, 2001). The risk assessment template is outlined in Figures 39-43 and based on risk management guidelines presented in previous research (Harding, 1998). The proposed errors and risks are based on the work presented earlier in this thesis – **Chapter 5, and Chapter 6**. In this stage, the system analysts shall complete SHEET ID 4.1-4.5 with a collaboration with domain experts, psychologists and designers. C-XAI method also provides several heuristics to map between XAI method technical properties and potential risks.

SHEET ID: 4.1	Trust Calibration Risk assessment – Functional
Dimension	
TASK ID:	Explanation method:
Provide a title for this assessment - This section depicts what trust calibration risk is.	
Risk identification. Based on the Explanation assessment sheet, what kind of risks can be identified?	
Please identify which XAI functional properties might introduce risks to the current Human-AI task?	
<ul style="list-style-type: none"> ○ ML approach applicability ○ Problem type ○ XAI method scope ○ Computational complexity ○ ML model applicability ○ ML model and XAI method relationship ○ Data features applicability ○ XAI method assumptions ○ Other: ----- 	
Risk analysis. Elaborate on what risks could be introduced and when the risk you identified could happen?	
Risk frequency and potential probability.	
Risk assessment. Please determine how acceptable is the risk on the trust calibration in the Human-AI task.	
No risks have been identified	
The risks require changing the explanation method	
The risks can be mitigated on the design level	
The risks have little effect on trust calibration	
Is there any additional information that you would like to be considered or you think is relevant?	

FIGURE 39 TRUST CALIBRATION RISK ASSESSMENT SHEET – FUNCTIONAL DIMENSION

SHEET ID: 4.2	Trust Calibration Risk assessment – Operational
Dimension	
TASK ID:	Explanation method:
Provide a title for this assessment - This field depicts what trust calibration risk is.	
Risk identification. Based on the Operational dimension assessment sheet, what kind of risks can be identified?	
Please identify which XAI operational properties might introduce risks to the current Human-AI task?	
<ul style="list-style-type: none"> <input type="radio"/> Explanation class <input type="radio"/> Explanation communication <input type="radio"/> Explanation interaction <input type="radio"/> Explanation audience <input type="radio"/> Explanation goal <input type="radio"/> Causality <input type="radio"/> Actionability <input type="radio"/> Other: ----- 	
Risk analysis. Based on the selected properties, what kind of risks can be identified?	
Skipping. <ul style="list-style-type: none"> <input type="radio"/> Lack of curiosity <input type="radio"/> Perceived goal impediment <input type="radio"/> Redundant information <input type="radio"/> Perceived complexity <input type="radio"/> Lack of context <input type="radio"/> Unfamiliarity <input type="radio"/> Other: ----- 	
Misapplying. <ul style="list-style-type: none"> <input type="radio"/> Misinterpretation <input type="radio"/> Mistrust <input type="radio"/> Confirmatory search <input type="radio"/> Rush understanding <input type="radio"/> Habits formation <input type="radio"/> Other: ----- 	
Risk frequency and potential probability.	
Risk assessment. Please determines how acceptable is the risk on the trust calibration in the Human-AI task.	
No risks have been identified	
The risks require changing the explanation method	
The risks can be mitigated on the design level	
The risks have little effect on trust calibration	
Is there any additional information that you would like to be considered or you think is relevant?	

FIGURE 40 TRUST CALIBRATION RISK ASSESSMENT SHEET – OPERATIONAL DIMENSION

SHEET ID: 4.3	Trust Calibration Risk assessment – Usability
Dimension	
TASK ID:	Explanation method:
Provide a title for this assessment - This section depicts what trust calibration risk is.	
Risk identification. Based on the Usability dimension assessment sheet, what kind of risks can be identified?	
Please identify which XAI usability properties might introduce risks to the current Human-AI task?	
<ul style="list-style-type: none"> <input type="radio"/> Truthfulness <input type="radio"/> Completeness <input type="radio"/> Contextfullness <input type="radio"/> Interactivity <input type="radio"/> Actionability <input type="radio"/> Novelty <input type="radio"/> Time ordering <input type="radio"/> Coherence <input type="radio"/> Complexity <input type="radio"/> Other: ----- 	
Risk analysis. Based on the selected properties, what kind of risks can be identified?	
Skipping. <ul style="list-style-type: none"> <input type="radio"/> Lack of curiosity <input type="radio"/> Perceived goal impediment <input type="radio"/> Redundant information <input type="radio"/> Perceived complexity <input type="radio"/> Lack of context <input type="radio"/> Unfamiliarity <input type="radio"/> Other: ----- 	
Misapplying. <ul style="list-style-type: none"> <input type="radio"/> Misinterpretation <input type="radio"/> Mistrust <input type="radio"/> Confirmatory search <input type="radio"/> Rush understanding <input type="radio"/> Habits formation <input type="radio"/> Other: ----- 	
Risk frequency and potential probability.	
Risk assessment. Please determines how acceptable is the risk on the trust calibration in the Human-AI task.	
No risks have been identified	
The risks require changing the explanation method	
The risks can be mitigated on the design level	
The risks have little effect on trust calibration	
Is there any additional information that you would like to be considered or you think is relevant?	

SHEET ID: 4.4	Trust Calibration Risk assessment – Safety Dimension
TASK ID:	Explanation method:
Provide a title for this assessment - This section depicts what trust calibration risk is.	
Risk identification. Based on the Safety dimension assessment sheet, what kind of risks can be identified?	
Please identify which XAI usability properties might introduce risks to the current Human-AI task?	
<input type="radio"/> Information leakage <input type="radio"/> Misuse <input type="radio"/> Instability <input type="radio"/> Other: -----	
Risk analysis. Based on the selected properties, what kind of risks can be identified?	
Skipping. <input type="radio"/> Lack of curiosity <input type="radio"/> Perceived goal impediment <input type="radio"/> Redundant information <input type="radio"/> Perceived complexity <input type="radio"/> Lack of context <input type="radio"/> Unfamiliarity <input type="radio"/> Other: -----	
Misapplying. <input type="radio"/> Misinterpretation <input type="radio"/> Mistrust <input type="radio"/> Confirmatory search <input type="radio"/> Rush understanding <input type="radio"/> Habits formation <input type="radio"/> Other: -----	
Risk frequency and potential probability.	
Risk assessment. Please determines how acceptable is the risk on the trust calibration in the Human-AI task.	
No risks have been identified	
The risks require changing the explanation method	
The risks can be mitigated on the design level	
The risks have little effect on trust calibration	
Is there any additional information that you would like to be considered or you think is relevant?	

FIGURE 42 TRUST CALIBRATION RISK ASSESSMENT SHEET – SAFETY DIMENSION

SHEET ID: 4.5		Trust Calibration Risk assessment – Validation Dimension	
TASK ID:		Explanation method:	
Provide a title for this assessment - This section depicts what trust calibration risk is.			
Risk identification. Based on the Validation dimension assessment sheet, what kind of risks can be identified?			
Please identify which XAI usability properties might introduce risks to the current Human-AI task? <ul style="list-style-type: none"> ○ Users' studies ○ Synthetic experiments ○ Other: 			
Risk analysis. Based on the selected properties, what kind of risks can be identified?			
Skipping. <ul style="list-style-type: none"> ○ Lack of curiosity ○ Perceived goal impediment ○ Redundant information ○ Perceived complexity ○ Lack of context ○ Unfamiliarity ○ Other: 			
Misapplying. <ul style="list-style-type: none"> ○ Misinterpretation ○ Mistrust ○ Confirmatory search ○ Rush understanding ○ Habits formation ○ Other: 			
Risk frequency and potential probability.			
Risk assessment. Please determines how acceptable is the risk on the trust calibration in the Human-AI task.			
No risks have been identified			
The risks require changing the explanation method			
The risks can be mitigated on the design level			
The risks have little effect on trust calibration			
Is there any additional information that you would like to be considered or you think is relevant?			

FIGURE 43 TRUST CALIBRATION RISK ASSESSMENT SHEET – VALIDATION DIMENSION

8.5 THE SELECTION AND IMPLEMENTATION PHASE.

The selection and implementation phase takes the risk assessment sheet from the assessment process and determines mechanisms that are capable of mitigating potential trust calibration risks. As there are almost multiple ways to address each trust calibration risk. This is a creative process that will likely depend on the design team; however, C-XAI method presents several design principles and guidelines for designing XAI interfaces to help trust calibration (Table 27).

TABLE 27 TRUST CALIBRATION RISKS DESIGN GUIDELINES

Design principle	Description	Recommended for risks	Techniques examples
Engagement	XAI interface designers to increase users' tendency to engage with the explanation.	Skipping	Third-party endorsement Herd cues
Challenging habitual actions	XAI interfaces designers for a calibrated trust may need to consider challenging users from developing habits with the XAI interface.	Habits formation Redundant information	Adaptive Effective monitoring
Attention guidance	XAI interface to promote the desired user behaviour and encourage users to look for relevant content in the explanation and combat overlooking it.	Perceived complexity	Visual cues Navigational cues Abstraction
Friction	XAI interface may adopt interactions that hinder people from painlessly achieving their goals when interacting with technology	Confirmatory search Rush understanding	Micro-boundaries Slow design Uncomfortable interaction
Training and learning	XAI interface design may need to help users to form correct interpretations for a given explanation. This may also include presenting usage scenarios, tutorials and FAQs.	Unfamiliarity Misinterpretation Mistrust Rush understanding	Onboarding technique Video tutorial Interactive explanation Self-learning tools

8.6 THE EVALUATION PHASE

This stage completes the design process and aims at evaluating the effectiveness of the prototype(s) developed in the Selection and Implementation stage. As the thesis argued in **Chapter 6** and **Chapter 7** that helping trust calibration does not only require providing effective explanations to end-users, but it also needs ways to nudge people towards applying cognitive engagement with XAI interface and help them in forming correct interpretations. A trade-off between the usability of the design and its effectiveness at helping trust calibration shall be considered to reach an acceptable solution. Many studies have proposed several measures to measure the role of XAI interface design in helping trust calibration and facilitating cognitive thinking (Bućinca et al., 2021, Chromik et al., 2021). The literature showed that evaluating XAI interfaces in trust calibration context shall be evaluated using behavioural indicators rather than self-reporting trust (Yin et al., 2019, Yang et al., 2020, Lai and Tan, 2019). They should that self-reporting trust are not reliable measurement for Human-AI collaborative decision-making tasks. C-XAI method provides four behavioural trust calibration measures used in previous work and revealed from this thesis literature review (**Chapter 2 – Section 2.2.4**). A summary of behavioural trust calibration measures is presented in Table 28.

TABLE 28 CALIBRATED TRUST BEHAVIOURAL MEASURES

Calibrated trust evaluation – behavioural measures		
	Description	Reported studies
Agreement percentage	It refers to the percentage of trials in which the participants decided to agree with the AI-based recommendations.	(Zhang et al., 2020b, Yin et al., 2019, Naiseh et al., 2021)
Compliance percentage	It refers to the percentage of trials in which the participants choose to follow the AI-based recommendation. The main difference between Agreement and Compliance measures is that the participants agreed with the AI-based recommendation and automatically made the final decision in the agreement case. In contrast, compliance only considers the case where the participants disagree with the AI-based recommendation, but they intend to comply with the AI-based recommendation.	(Zhang et al., 2020b, Yin et al., 2019, Naiseh et al., 2021)
Incorrect decisions	This is a team-performance measure, and it is extracted from incorrect decisions made between the human and the AI.	(Buçinca et al., 2021, Bussone et al., 2015, Yang et al., 2020)
Correct decisions	It measures the percentage where the collaborative decision-making between the human and the AI has led to a correct decision.	(Buçinca et al., 2021, Lai and Tan, 2019, Naiseh et al., 2021, Yang et al., 2020)

Although behavioural trust calibration measures are effective in measuring the role of XAI interface in helping trust calibration, it does not measure the effectiveness of the XAI interface in promoting cognitive thinking and debiasing users' behaviour. Hence, in addition, to trust calibration behavioural indicators, evaluating XAI interface shall consider the extent to which the XAI interface enables cognitive thinking. To help this process, C-XAI method provides several methods to measure users' engagement with the XAI interface during the evaluation process (See Table 29). These methods are based on a recent systematic literature review on users' engagement (Doherty and Doherty, 2018).

TABLE 29 EVALUATION METHODS FOR USERS' ENGAGEMENT BY (DOHERTY AND DOHERTY, 2018)

Cognitive thinking evaluation methods	
Measure	Description
Subjectivity-Oriented approaches	Subjectivity-oriented measures cover observation, questionnaires, interviews, and other forms of self-report. The choice of these methods entails compromise to gather rich data including cognition, emotion, and memory, e.g., User Engagement Scale (UES) (O'Brien and Toms, 2013) and Think-aloud protocol (Glasnapp and Brdiczka, 2009).
Objectivity-Oriented approaches	Objectivity-oriented measures attempt to deduce engagement without direct questions or the researcher involvement. This involves techniques as logging and interaction, psychophysiological measures, and audio and visual analysis. The advantage of these methods rests in their ease of use and limited disruption to participants experience. One common example of Objectivity-Oriented approaches is eye-tracking (Ishii and Nakano, 2010)

Once an acceptable solution is reached from C-XAI evaluation stage, the design is ready to undergo more traditional evaluations using human factors (Whitefield et al., 1991) and performance analysis methods (Vermeeren et al., 2010).

8.7 C-XAI WORKFLOW

C-XAI method adopted a participatory design approach to ensure that all relevant stakeholders are actively involved in the early stages and later stages of developing XAI interfaces for collaborative Human-AI interaction. C-XAI includes seven main activities (See Figure 44). The activities in (Figure 44) aim to explain four stages of C-XAI method.

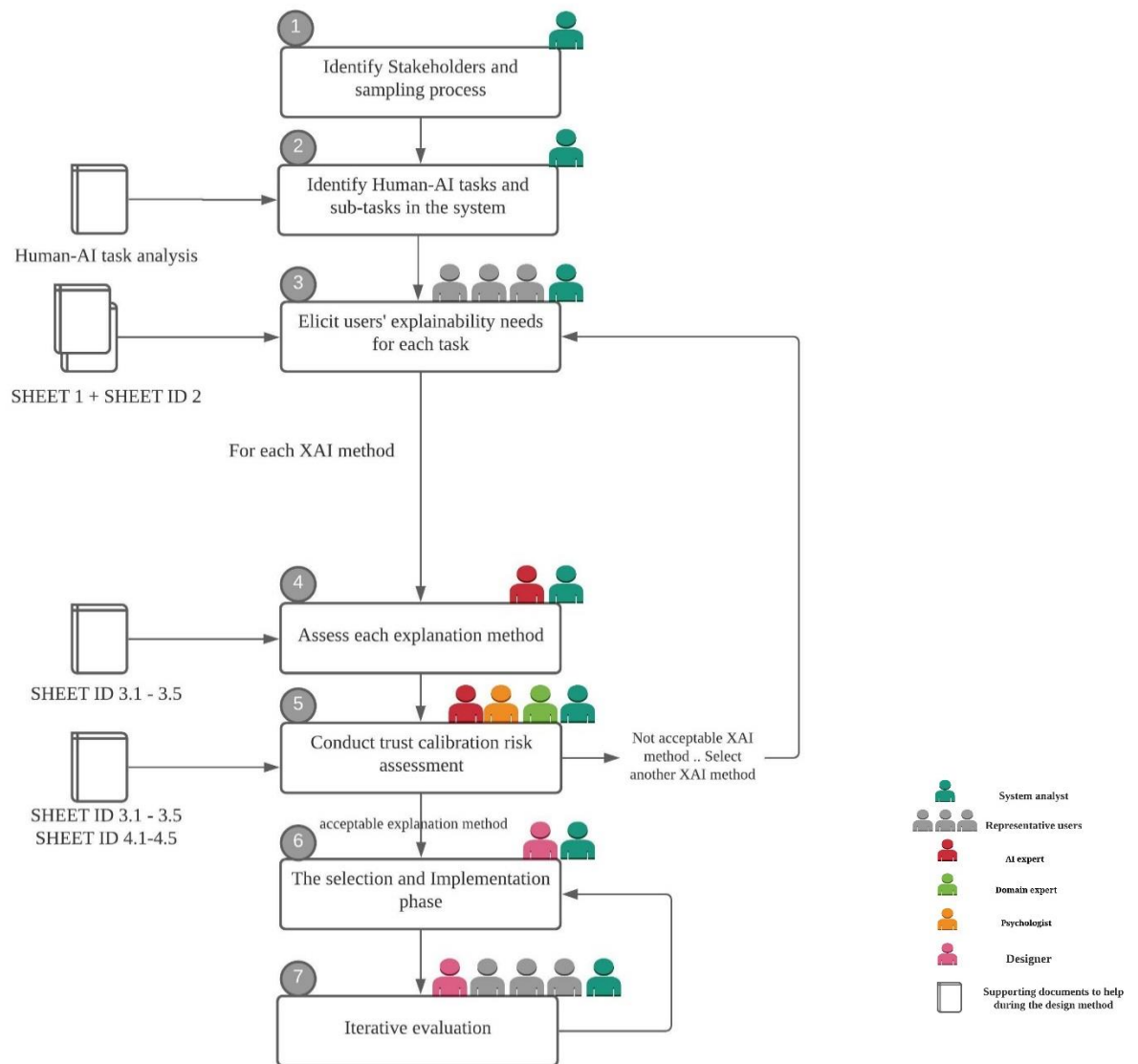


FIGURE 44 C-XAI DESIGN METHOD WORKFLOW

During the design method, the system analyst needs to consider supporting users' in judging their cognition-based trust components to inform better design decisions. To help this process, this section aims to provide some good practices for the analyst and the design team to be used as a reference point when designing the XAI interface. The design practices presented here are derived from relevant literature and previous studies in this thesis.

Facilitate guidance to interpret AI explanations.

XAI interfaces can change users' behaviour and help them form correct mental models by guiding them in using and interpreting the explanation. In general, explanations have boundless and unlimited nature, i.e., explanations are often large and far beyond the grasp of any one individual. For instance, most of the AI decisions have an enormous set of relations and hundreds of data feature contributing to the decision. People tend to be miscalibrated with their understanding- and think they understand an explanation more they do, this phenomenon is called the "*illusion of explanatory depth*". Thus, guiding the process of interpreting the explanation can change their attitude or behaviour; this might include video tutorials or FAQ tutorials accessible for users. Information such as when, how, why to use explanations can be essential for a better learning experience in XAI systems.

Facilitate usable interaction

An explanation design that makes the explanation easy to find and navigate provides opportunities to persuade users to read the explanation. Explanation design shall guide users through the process of reading the explanation. This aligns with the phenomena related to learners' readiness to learn and read. Learners usually tend to follow measures and instructions in which the instructor guide the learning process. For instance, highlighting explanations' chunks that are important to read, or using navigational cues to switch from a section of the interface to another.

Task-centred explanations

When explainability is used in a collaborative decision-making environment, it is supposed to be integrated with the task workflow and task constraints. Explaining the logic of the algorithm shall be done in a way that is tightly coupled with the subject, i.e., the task for which the recommendation and the explanation are given. One example of task constraint was encountered in our user studies in Counterfactual explanation scenarios, where the explanation only provided information about the kinds of changes that should be made to an AI recommendation to change the decision. Our participants raised their concerns mentioning that some data features in the patient profile have static values and cannot be changed, e.g., patient demographics. During the

users' mental model elicitation process, the system analyst may need to focus on eliciting the task constraints regarding the selected explanation methods.

Include assurances in the XAI interface design.

Assurances in HCI literature is a design property that also applies to the intelligent tool so that they help users' trust calibration. Assurances are indicators and performance metrics to indicate the actual capabilities of the intelligent tool. C-XAI revealed two categories of assurances: *XAI class validity* and *XAI class capability*. Regarding the XAI class validity, participants described that they were unable to have guidance about trusting the provided explanation validity and correctness. Some suggested knowing the source of the data to assess the credibility of the explanation, with mentioning that it would also need to be up to date with the current changes in the task domain. Others also asked whether the explanation is generated based on training the AI on multiple data sources and references. On the other hand, XAI class capability was related to the metrics used to evaluate the explanation in both AI and task domains. Participants argued that explanation verification with a medical expert should be performed according to accepted standards, incorporating best practices related to expert selection, elicitation protocols, bias avoidance, documentation and peer review. Moreover, participants raised questions that could be answered through clarity about the evaluation metrics used in the XAI models as introduced in a previous survey on interpretable machine learning models (Carvalho et al., 2019). Participants demanded information about the XAI class itself such as a) accuracy of an explanation on unseen cases, b) fidelity that shows how well the explanation is consistent with the underlying AI model, c) stability that represent how similar are the explanation for similar cases d) representativeness that describes how many cases could be covered by given explanation.

Tailoring needs.

Tailoring is defined in this thesis as a modification or an explanatory process to answer the user-specific question, e.g., what-if questions. A particular risk of a lack of tailoring and personalisation mechanisms is that the explanation may not always align with end-users needs, increasing the possibility of skipping or misapplying. Chapter 5 and Chapter 6 showed that allowing the users to steer and guide an explanation can benefit trust calibration by supporting users' in exploring different explainability scenarios. This approach is also used to help users' in forming correct mental models (Sokol and Flach, 2020b). Therefore, tailoring can be used in the XAI interface regarding the explainable method properties or trust calibration risks. The literature review section in this thesis derived a full taxonomy of tailoring the explanation to end-user. It is based on the six-dimension model of personalisation.

Multi-step interaction

Developers of such explainable interfaces may need to collect data from end-users regarding the users' information needs after presenting the main explanations. Humans are likely to ask follow-up questions when they did not receive a casual response in the explanation. Explainability is a social process between the explainer and the explainee (Miller, 2019), and this may need to follow a multi-step interaction approach between the Human and the XAI interface. Design considerations for the modalities of such multi-step explainability are also required to balance explainability and usability, e.g., chatbots.

8.9 CHAPTER SUMMARY

This Chapter presented C-XAI method and its supporting documents. It explained the four phases of the method and recommended the supporting documents to be used during the design phases. Guidelines for the system analyst and the design team for designing the XAI interface are also provided. The next chapter presents the evaluation of the C-XAI method, and this would help assess whether the method could help the design team to specify and anticipate trust calibration risks given an explanation method.

9. CHAPTER 9: EVALUATING C-XAI METHOD

CAT is a design method that provides a systematic approach to guide the design of XAI interfaces for helping trust calibration. C-XAIs' ultimate goal is to raise the chances of successfully building an XAI interface to help users calibrate their trust. Due to the limited knowledge on approaches for designing XAI interfaces for trust calibration and because trust calibration is a dynamic process, it would be a challenging aim to develop an effective method. Furthermore, this would require establishing a clear understanding of the users' personality and intention to use. Therefore, C-XAI method does not claim that the generated interface will eliminate potential trust calibration errors, but it follows a systematic approach to mitigating these errors. **Chapter 8** presents C-XAI method after the multistage evaluation process followed in this chapter. This thesis stated that a successful trust calibration in AI-based recommendation supported with XAI interfaces needs more than a correct and impartial explanation content and shall benefit from novel design elements to persuade and nudge users towards more cognitive engagement and correct interpretations (**Chapter 5** and **Chapter 6**). This would untimely help XAI users to understand the AI reasoning and calibrate their trust. The methods' scope involves guiding different stakeholders during the XAI interface development to elicit required explanations for Human-AI tasks, assess explanations' role in helping trust calibration and select appropriate design techniques to aid trust calibration. The method is supported with various activities, templates and supporting documents, including design guidelines and good design practices as reference material for the system analyst and the design team. A qualitative case study will be adopted to evaluate C-XAI method. The evaluations' focus will consider the usefulness and the completeness of the method as well as its ability to generate effective designs to enhance trust calibration process. This Chapter aims to achieve the second section of **Objective 5** outlined in this thesis.

Evaluating an engineering method might have different approaches depending on the context in which the evaluation is employed, i.e., having participants with diverse knowledge and understanding taking part in the evaluation study (Kitchenham et al., 1997). Evaluating an engineering methodology is categorised into three main categories: objective, subjective and hybrid evaluation. The objective evaluation focuses on identifying the usefulness of the proposed method. Such evaluation includes a quantitative assessment for accelerating the process and reducing the cost of the process. Subjective evaluation involves a qualitative assessment of the generated requirements and needs from the method. Finally, hybrid evaluation combines subjective and objective evaluation approaches in the evaluation phase.

Evaluating engineering methods can include three different activities (Kitchenham et al., 1997).

Experimental evaluation. Participants engage with the engineering method by performing various tasks related to the engineering method. Then, the collected data can be analysed following a statistical approach.

Case study evaluation. The method is applied in a real-world context. A case study takes a real-world problem and measures the role of the engineering methodology in supporting the process. Such evaluation method includes following the guidelines and procedures of the organisation.

Survey evaluation. It includes collecting data from participants who are experienced in using engineering methods and tools. They will provide statistical data about the evaluated method.

9.1 EVALUATION GOALS

This chapter aims to evaluate the **effectiveness** of the C-XAI method. In particular, this evaluation is meant to find out how the C-XAI method is useful and effective in the way it helps the design team identify and cater for the design requirements for mitigating potential trust calibration risks. The effectiveness of following a design process (Veryzer and Borja de Mozota, 2005) suggested will be used as a baseline for this evaluation study.

- i. Aid the focus of the stakeholders on the design problem
- ii. Increase the awareness of the users and their diverse contexts and needs
- iii. Better design of the product
- iv. Effective communication tools and increased engagement among the design team.

Furthermore, the evaluation study aims to evaluate the proposed templates and supporting documents based on the following criteria:

- **Completeness.** This criterion refers to the methods ability to cover all design stages to develop the XAI interface. It also considers whether the guidelines provided to aid the design method are enough.
- **Understandability.** This criterion is to determine the understandability degree of the method from the stakeholders' point of view. It also covers the evaluation of supporting documents and templates.
- **Usefulness.** This criterion aims to evaluate how the method and its supporting documents simplify and improve the XAI interface design process.

9.2 EVALUATION CONTEXT

The thesis will follow a case study evaluation approach to evaluate the C-XAI method's effectiveness in designing XAI interfaces that enhance users' trust calibration. The case study approach facilitates the exploration of a phenomenon using various data sources (Baxter and

Jack, 2008). The following case study evaluation approach investigates whether the proposed method helps various stakeholders successfully develop novel XAI interfaces that help trust calibration. A case study evaluation approach would enable the researcher to observe stakeholders interacting with the C-XAI method. Such an observational approach would provide rich data and better understand stakeholders' reactions during the design process. The case study will also help the researcher identify various difficulties and flaws in the supporting documents and templates. The outcome of the evaluation study would then be used to refine and optimise the C-XAI method. The case study evaluation method's choice is informed by the stated advantages, available resources, time, and research type.

9.3 EVALUATION CASE STUDY

The evaluation stage followed a case study approach of the Screening Prescription (SP) tool. This tool is an AI-based tool to validate if the prescription fits the patients' profile and history and can be prescribed for a specific patient. This tool was developed to help medical practitioners in their everyday decision-making scenarios. Medical practitioners interacting with the SP tool might fail to calibrate their trust due to the task's dynamic nature of the AI margin of error. The SP interface developers would like to build a safe and effective interface by making the underlying logic of the SP tool transparent and explainable. It has also been shown in this thesis even though explanations were presented to users; they failed to calibrate their trust. This was due to two main reasons: (i) users skipped important and useful explanations that support them in calibrating their trust, and (ii) users misapplied the explanation in their task, e.g., misinterpretations and unfamiliarity issues. The C-XAI method's role is to identify the required explanations SP tool and identify constraints and design principles on the XAI interface level to enhance trust calibration process. Such an approach would ultimately limit users' errors and mitigate trust calibration risks introduced earlier in this thesis (see Chapter 6).

9.4 EVALUATION STUDY PHASES

The evaluation process consists of three main phases (See Figure 45). The following sections summarise these phases. Detailed justification of these phases can be found in the following sections. Evaluation material used during these phases can be found in [<https://bit.ly/3gmdPpy>].

Phase 1. This phase included a familiarisation and induction session to the C-XAI method and the supporting materials. During this phase, the researcher explained the study's goal and introduced the participants to relevant documents and templates. The researcher also mentioned that the design process is expected to be an iterative process; this was to inform participants that follow-up meetings might be needed to evaluate the generated designs.

Phase 2. This phase was mainly about validating the C-XAI method and its supportive documents from domain experts' points of view. The evaluation was to address the

completeness, understandability and usefulness of the templates and their supporting documents from an expert point of view.

Phase 3. The evaluation in this phase was to evaluate C-XAI methods effectiveness as described in Section 9.1. This phase included two main exercises to evaluate C-XAI method. This phase was divided into two stages:

- **Stage 1.** It involves designing the XAI interface with the aid of the C-XAI method. This includes ideation and brainstorming activities to enable using C-XAI in the design process. The role of the researcher during this stage was to observe C-XAI method in helping the design process. Two focus groups were conducted with two different stakeholders to use C-XAI method.
- **Stage 2.** Each participant was invited to evaluate Calibrated trust method using an evaluation survey (Section 9.7).

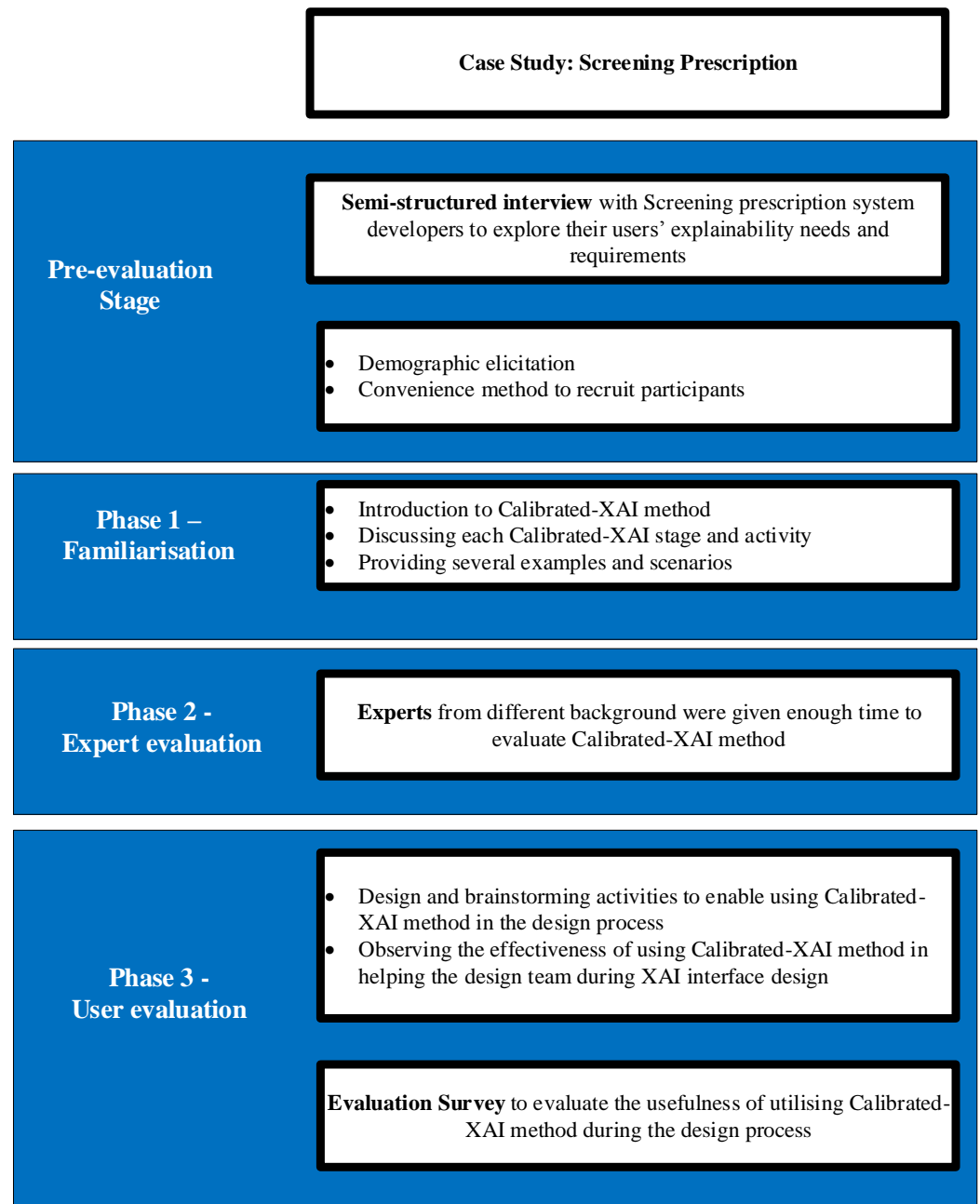


FIGURE 45 C-XAI METHOD EVALUATION PHASES

9.4.1 PHASE 2: EXPERT EVALUATION

This phase involved an evaluation of the C-XAI method and its relevant documents and templates. First, the researcher presented the templates and documents to the experts who have experience in relevant domains, e.g., AI, HCI, requirement engineering and psychology. Then the researcher provided a 20-minute presentation with a detailed justification of all the materials before starting the evaluation. This phase required experts to review and validate all the documents provided to them critically. Finally, they were asked to refine the documents and templates by freely adding or removing elements. This phase's outcome supported the evaluation study to produce refined versions of the templates and supporting documents.

The main aim of this phase was to figure out whether the templates and the supporting documents are complete, understandable and guide the XAI interface design process. Expert evaluation was also to correct any issues before commencing Phase 3 of the evaluation case study. The recruitment of the participants was based on two inclusion criteria (i) five years or more experience in relevant fields, including AI, HCI, requirement engineering and psychology and (ii) familiarity with trust or trust calibration literature. Such inclusion criteria were to ensure the credibility of the findings. As a result, six participants were able to participate in this phase (See Table 30). Participants in this phase were provided with a booklet that contains detailed information about C-XAI method. Participants were provided with guided questions to evaluate the C-XAI method (See Table 31). They were also asked to amend the booklet and add their feedback and comments.

TABLE 30 EXPERT PARTICIPANTS DEMOGRAPHICS

Participants ID	Background and expertise	Experience
P1	Machine learning expert	8 years
P2	Requirement engineering expert	5 years
P3	Interaction design and UX expert	6 years
P4	Behavioural change/ Psychology	10 years
P5	Software engineering	8 years
P6	Human-computer interaction expert	12 years

Expert feedback was collected in this phase and produced a refined version for the documents and the templates based on their recommendations. After collecting experts' feedback, the researcher further refined the templates and the supporting documents. The refinement version of C-XAI method after an expert evaluation is presented in [<https://bit.ly/3gmdPpy>].

TABLE 31 EXPERT EVALUATION GUIDED QUESTIONS

Expert questions	<ul style="list-style-type: none"> Do you think the method misses any aspect of the design? What do you think about the relevance of the methods' template? What challenges do you think the users of this method could have? What recommendations do you provide to enhance the methods' understandability and completeness? To what extent do you think the templates and the supporting material would help the method's stakeholders identify trust calibration risks and design requirements?
------------------	---

9.4.2 PHASE 3: DESIGNING XAI INTERFACE WITH C-XAI METHOD

Phase 3 was to design an XAI interface with the help of the C-XAI method and its supporting documents. First, participants were invited to a focus group to immerse themselves in the design

problem and discuss their opinions about how the XAI interface design can help users calibrate their trust. Then, participants were allocated randomly to two focus groups to design the XAI interface with the help of the C-XAI method. Participants were provided with three primary documents:

1. **Document ID 1** describes the design goal, the design problem and the Human-AI collaborative decision-making task, i.e., screening prescription.
2. **Document ID 2** describes C-XAI method documents.
3. **Document ID 3** describes explainability needs for the Human-AI collaborative decision-making task collected during the pre-evaluation stage (**See Figure 45**).

Document ID 3 was based on users' needs analysis taking prescription screening as a case study from IQemo IQHealthTech². That means Phase 3 evaluation study evaluated the C-XAI method starting from the Assessment phase, i.e., the identification process has already been performed by the IQHealthTech organisation. The available resources and time informed the choice of such an approach. Participants involved in this stage of phase 3 were not engaged in any thesis studies that helped develop the C-XAI method. The main aim of this inclusion criterion was to ensure that participants had no prior knowledge about the C-XAI method. Participants who agreed to participate in this stage collaborated with the system analyst to generate the XAI interface design.

Then, two design sessions were conducted to design an XAI interface for screening prescription AI-based tools to help trust calibration. The researcher started both sessions by briefing the session's aim and a description of the design problem. The system analysts who guided the design process was familiar with requirement engineering standards and guided different stakeholders through the design process. Various stakeholders worked in one group to evoke brainstorming and critical thinking. Participants worked together to produce the prototypes following the instructions provided in (**DOCUMENT ID 1,2 and 3**).

After completing each session, participants were given an open-ended survey to gather participants experience during the design sessions. They were also asked whether they had encountered issues or difficulties during the design process. This survey aimed to take participants overall experience into consideration when conducting the analysis. The collected data was analysed to refine the method templates and supporting documents. The researcher focused on evaluating the templates' effectiveness, i.e., to what extent the method and its template helped participants understand and analyse the design problem. The researcher also focused on looking for the templates' flaws and disadvantages and what can be added or removed from the templates.

² <https://www.iqhealth.tech/>

9.5 STUDY PROTOCOL

The researcher followed the following guidelines for conducting evaluation methods and analysing qualitative data provided in (Baxter and Jack, 2008, Pope et al., 2000). The evaluation phases followed the below workflow:

1. A pilot study was conducted to evaluate the study design within the research team and two external reviewers.
2. Recruiting participants was based on the selection criteria introduced in Section (9.6). Participants were asked to complete a pre-selection survey.
3. Participants were emailed with a consent form, together with a participant information sheet verifying the purpose of the study, and the expectations of their involvement, (See Appendices)
4. A semi-structured interview will be conducted with IQemo Developers. The interview aims to investigate their users' explainability needs and requirements.
5. Then, experts from multidisciplinary backgrounds were invited to evaluate the C-XAI method. The method booklet was emailed to the experts and they were given 15 days to return the booklet and answer the evaluation questions. To increase the evaluation validity, participants from different backgrounds and five or more years of experience were recruited.
6. The researcher made a one-month gap between expert evaluation and user evaluation to analyse the collected data and improve the C-XAI method based on experts' feedback.
7. Then, a group of multidisciplinary teams, which consists of 14 participants were invited to attend two focus group sessions. The first session was 'familiarisation', and the second session was brainstorming and design. Participants received a consent form and an information sheet before the study date. Both documents described the purpose of the study and the level of expected participation from participants. Participants also received a Zoom link to join the session online.
8. Participants joined the online session. The researcher provided all the relevant documents to all participants.
9. The researcher provided a 20-minute presentation to the participants to familiarise them with the study activities and the provided material. Also, the researcher informed the participants about their expectations and ways of completing the activity.
10. During the study, the researcher role was limited to an observer and provided guidance when needed. This helped the researcher observe main issues and difficulties with the method. The researcher will trigger C-XAI by two activities:
 - a. Design activity to use the C-XAI method to design an XAI interface.

- b. Brainstorming activity to reflect more on their experience during the design process. It also examines the effectiveness of the C-XAI method of communication.

11. At the end of the session, GAQ will be given to each participant to evaluate the use of C-XAI method during a design process to help trust calibration.

9.6 PARTICIPANTS RECRUITMENT PROCESS

The evaluation requires participants to play stakeholders' roles based on their background and experience. C-XAI method is designed based on a participatory design approach. The C-XAI method's main stakeholder is the system analyst, who is expected to collect and analyse requirements. Also, the system analyst should know about system developments and provide guidelines for the different stakeholders during the software development process. Psychologists are also expected to participate in the C-XAI method evaluation process. They are expected to know human biases and decision-making theories to support the identification of trust calibration risks and ensure that these risks are mitigated in the XAI interface design. Designers shall also have experience in designing interactive and user-centred interfaces. Representative users and domain experts shall be experts in the selected task; they shall be doctors or pharmacists who used AI-based prescribing systems before. Table 32 present the demographic information about the participants involved in the C-XAI evaluation study. The study information sheet and informed consent can be found in [<https://bit.ly/3gmdPpy>].

Participants selection criteria for the evaluation study included (i) representative users who have experience in prescription classification, and (ii) experts shall have at least 5 years of experience in their domains. The evaluation study employed a convenience sampling approach.

TABLE 32 STAKEHOLDERS EVALUATION

Participants ID	Background and expertise	Experience	Role
User-centred approach evaluation activity			
P1	Requirement engineer	9 years	System analyst
P2	Medical doctor	6 years	Representative users
P3	Medical doctor	6 years	Representative users
P4	Expert doctor	10 years	Domain expert
P5	UX designer	5 years	Designer
P6	Psychologist	7 years	Psychologist
P7	Machine learning engineer	9 years	AI expert
C-XAI approach evaluation activity			
P8	Machine learning researcher	12 years	AI expert

P9	Pharmacists	6 years	Representative users
P10	Pharmacists	8 years	Representative users
P11	Expert pharmacist	11 years	Domain expert
P12	HCI researcher	7 years	Designer
P13	Requirement engineering researcher	6 years	System analyst
P14	Psychology researcher	9 years	Psychologist

9.7 DATA COLLECTION AND ANALYSIS METHODS

In this stage, the research adopted a qualitative data approach. Methods such as focus groups, interviews and an open-ended questionnaire were followed to help to achieve the evaluation study goals. The following sections describe each of the used methods:

Phase 2 – Semi-structured interview. This method was used to collect feedback on the templates and the methods' supporting material for the expert evaluation phase. The researcher followed a semi-structured interview protocol and sought a justification for the participated experts' feedback. Before the interview, participants were informed about the study's goal, and they were inducted about the C-XAI method and its supporting documents. Then, participants had 15 days to read C-XAI feedback and provide their comments and reviews. Then, a 30 minutes semi-structured interview was conducted to review experts' feedback and clarify the analysed feedback where clarification was needed.

Phase 3 - Focus group. Participants in Phase 3 were invited to participate in a focus group study. Two focus groups were conducted to collect feedback from 14 multi-disciplinary experts who took part in the evaluation phases. Participants were randomly assigned to one of the focus groups based on their experience. Participants were introduced to the study workflow and the supporting documents before the focus group. After completing the design process, the researcher invited participants to a brainstorming activity to reflect on their experience during the design process.

Phase 3- Observation method. Phase 3 will use an observational data collection method. Observation is a qualitative research method in academic disciplines such as anthropology, sociology, education, development studies and psychology. It gives a robust way to capture and understand peoples' behaviour and reactions towards a particular goal by observing and noting their activities from a distance. Such a method helped collect data without disturbing the design process and affecting participants behaviour. As this phase uses a case study approach, the observation method will be beneficial in evaluating the effectiveness of the C-XAI method in the design process. It is possible to observe how design team members use the C-XAI method in many aspects, such as where they find it useful, challenging and understandable. It also provides

more freedom for both the observer and design team to use, think aloud and ask questions during the evaluation process.

Phase 3 – Group Administrated questionnaire (GAQ). GAQ will assess the design process both with and without the C-XAI method and get more insights into the design team. GAQ is an option for individual interviews; each participant was asked to complete the questionnaire. Participants recorded their responses individually without communicating with each other. This method is beneficial for the researcher and ensures a valid and independent view of the design process. If a participant does not understand any specific question, they can ask for clarification. The GAQ in this study included open-ended questions and built on the design methods used (with and without C-XAI). The open-ended questions will be used to evaluate the effectiveness of using the C-XAI method from the design team's perspective and is based on their experiences of utilising it during the session.

Data analysis. Four sets of data were collected and used during phase 3 evaluation. The first consisted of the transcript of audio files of both of the study stages. The second is the researchers' notes, which contained observations of participants' comments, communication and reactions. The third data set contained the collected templates and supporting documents. The fourth dataset included participants answers to the GAQ survey. For qualitative data, content analysis with the Nvivo tool's support was performed. The researcher transcribed audio files, which helped familiarise writing the data and listening to the audio files. Codes that emerged from the analysis represent particular issues and concerns for the study participants such as a situation that triggered an error or difficulty. The researcher also coded several usage patterns, emotions, and psychological states that were not expected during the interaction time and were vital to refine the C-XAI method.

9.8 FINDINGS

This section presents the findings from the evaluation study. The results will be discussed based on the evaluation study goals: completeness, understandability, usefulness and effectiveness.

9.8.1 EXPERT EVALUATION FINDINGS

As a general observation, experts rated the proposed method as a practical, comprehensive and complete method for designing XAI interfaces. Experts mentioned that the templates provide a comprehensive description of the design process and help C-XAI users identify several trust calibrations risks. However, five out of six experts mentioned several examples and heuristics that map between XAI technical properties and trust calibration risks should be provided as an additional document to the C-XAI method. They commented that such a document would help stimulate the thinking process and evoke brainstorming around the design problem. It is also to facilitate a dialogue between different stakeholders during the design process.

Considering the content of the templates, AI experts suggested changing the answers of some XAI method properties that have “Yes/No” options to “low/medium/high” options. This would represent an accurate description of the property of the XAI method. P1, a machine learning expert, mentioned: “*all explanation models have these features, but it depends on their levels ... so yes/no answers are not really descriptive*”. Five explanation method properties were changed accordingly: Complexity, Novelty, Coherence, Soundness and Completeness.

Regarding the terminology used in the templates and the method. Participants agreed that the language and the terminology are generally understandable. However, some participants have suggested adding another document to justify each explanation method's properties and trust calibration risks provided in the templates. In addition, they mentioned that some of the provided terminologies are not self-explanatory and might cause a misunderstanding to the method's stakeholders. For instance, P4 mentioned, “*Well ... misusing the explanation could be interpreted differently... it is better to provide what does that mean*”. Therefore, a glossary template was added to the methods’ supporting documents.

Regarding the technical XAI assessment sheets, participants suggested that some elements can be removed because they are repetitive. The researcher argued that this was adopted from a framework in the literature, but the experts made the case that providing such repetitive questions could confuse the stakeholders. For instance, explanation actionable feature appeared both in the operational assessment sheet and usability assessment sheet. Furthermore, experts suggested adding five different trust calibration risk assessment sheets corresponding to each XAI assessment dimension. Participants argued that it might be overfolding and confusing to combine five assessment dimensions in one template. Five Trust calibration risk assessment sheets were developed accordingly.

Experts were interested in the design guidelines and good design practice and agreed that such information is useful during the XAI interface development, “*I like the guidelines ... designers and requirement engineers would definitely need them*”. However, amendments were suggested related to the terminology, length and writing style of the guidelines. Three experts mentioned the length of the guidelines and suggested shortening them. P3 suggested, “*The guidelines need to be shortened as they contain a lot of information and it might be difficult for the designers and system engineers to follow with it*”. Two experts also described the guidelines as academic guidelines where the guidelines of the C-XAI method shall be more informative to the designers. P4 added, “*Well I can understand these guidelines, but I doubt designers would be able to follow up with this style of writing... it is more academic*”. Therefore, the guidelines and good design practices were amended and styled to reflect their feedback.

Finally, the C-XAI method provided evaluation metrics based on the thesis literature review, such as the compliance with the AI and the percentage of correct decisions as behavioural

indicators for a successful trust calibration. However, five out of six experts mentioned that the proposed evaluation metrics were not complete regarding the description of the design problem. As the goal of the design is to engage XAI users with more cognitive thinking and form correct interpretations of the provided explanations, qualitative and quantitative measures to cover both cognitive engagement and correct interpretations shall be considered during the design process. Therefore, guiding methods that measure XAI interface cognitive engagement and correct interpretations were added.

All the mentioned modifications and amendments can be found by comparing the C-XAI v1 to C-XAI v2. Table 33 summarises the main points regarding each evaluation criteria resulting from the evaluation study.

TABLE 33 SUMMARY OF EXPERT EVALUATION PHASE

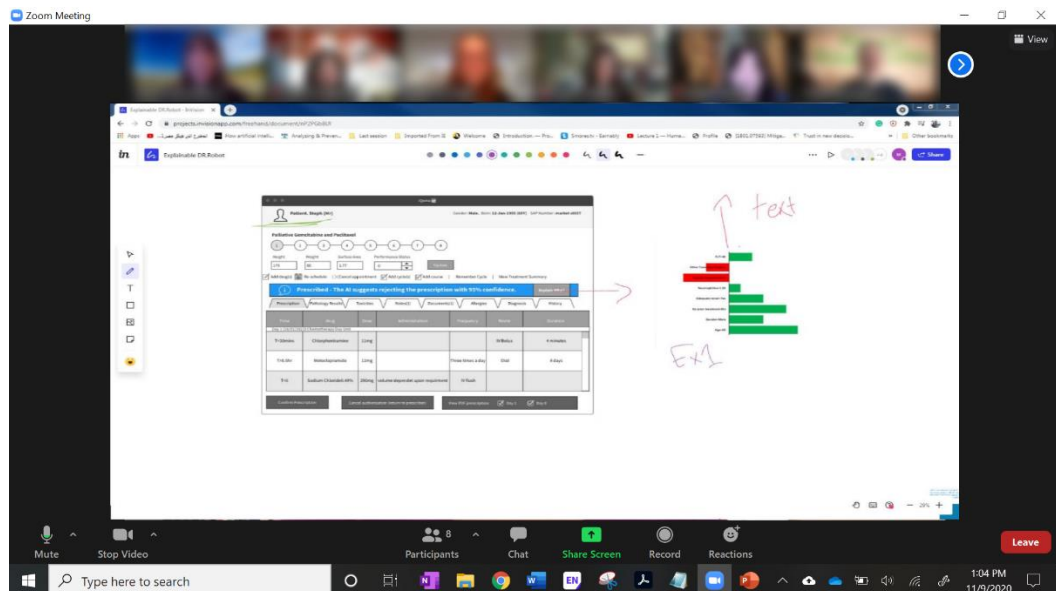
Completeness	
Templates and supported documents covered all the required assessments related to enhancing trust calibration problem “ <i>I would say it is complete, I cannot think of anything missing</i> ”. Another remarked that “ <i>The entire process was explained sufficiently in the booklet</i> ”.	+
Participants required additional spaces to improve the communication between the design stages. These suggestions are described in section 7.9.1.	-
Participants asked for additional supporting documentation that explains the main terminologies used in the design method, Participant stated, “ <i>what does explanation completeness exactly mean</i> ”.	-
The evaluation stage missed a direction to help C-XAI users measure cognitive thinking and correct interpretations.	-
AI experts suggested changing the answers of some XAI method properties that have “ <i>Yes/No</i> ” options to “ <i>low/medium/high</i> ” options.	-
Experts suggested adding five different trust calibration risk assessment sheets corresponding to each XAI assessment dimension.	-
Understandability	
The templates and the workflow provided a clear structure of the CAT method, which helped the system analyst to follow.	+
The order of the stages was logical, for instance, providing an assessment for the explanation method before the trust calibration risk assessment.	+
Participants agreed that more examples are required in each step to enhance users’ understandability of the design problem, e.g., P3 during the trust calibration risk assessment stated, “ <i>I would like to see more examples for explanation properties that may introduce trust calibration risk</i> ”.	-
System analysts mentioned that the good design practice for eliciting explainability needs pointed out several requirements to be considered. For instance, the system analyst was not aware of collecting explainability needs before and after consuming the	+

main explanation.	
Explaining ‘when’ and ‘how’ to implement the proposed design guidelines stimulated participants thinking and helped them to come with innovative design solutions.	+
Usefulness	
Experts mentioned the templates encapsulate different information in relation to trust calibration design problem.	+
Experts mentioned that the generated design would help reduce the cost of users’ errors, specifically when a high-stake Human-AI task is implemented. They mentioned that the C-XAI method will help the designers recognise what those potential problems are to be solved in the design to increase the effectiveness of the explanation. One expert mentioned, “ <i>I can see how this method would point out different user errors ... methods such as User-centred design may not be able to reveal such problems ... it takes a considerable amount of research to come up with these ideas and errors</i> ”.	+
Experts were interested in the design guidelines and good design practice and agreed that such information is useful during the XAI interface development	+

9.8.2 DESIGNING XAI INTERFACE WITH C-XAI METHOD FINDINGS

This phase included a design activity following the proposed C-XAI method. Fourteen participants joined this phase-separated in two sessions (Seven participants in each activity), and the researcher was an observer during the session. Overall, this activity lasted for 3-4 hours. During this activity, the dialogue between different stakeholders was documented by taking notes and collecting the generated material. Participants worked together to design an XAI interface for an AI-based screening prescription tool. This activity was guided by the supporting documents provided in this stage. Participants started the activity by reading the design problem description, and they had enough time to understand the problem and ask follow-up questions. Based on the CAT design method structure, participants were not fully active during this time. Only the system analyst was involved in all the design activities. In the next subsections, the findings from each C-XAI activity is discussed. Screenshot from the participants' design session is presented in Figure 46.

FIGURE 46 SCREENSHOT FROM WITH C-XAI METHOD EVALUATION SESSION



9.8.2.1 IDENTIFY TASKS AND SUB-TASKS IN THE SYSTEM

Before the session, the system analyst was given a description of the Screening prescription tool and full documentation about its functionalities and the representative users. The system analyst was also given Human-AI task analysis sheet based on IQemo Analysis to their system and explainability needs. The system analyst took enough time to read the sheet and understand the required tasks and explanations. The system analyst took a 15-minute break before the next activity of the C-XAI method.

Task	Subtasks	XAI class	XAI method	Additional needs
Screening prescription	NA	Local Explanations	LIME	Data features correlation Data features annotation Data features grouping Explain the terminology of the AI when needed

FIGURE 47 HAI-A ANALYSIS SHEET PROVIDED TO PARTICIPANTS

9.8.2.2 ASSESS EACH EXPLANATION METHOD

The system analyst and AI experts joined the assessment activity to assess each explanation method identified in the HAI-A analysis sheet. In this activity, the system analyst collaborated with an AI expert to complete **SHEET ID 3** templates. This is a crucial step for the C-XAI method as it helps identify XAI methods properties that may trigger trust calibration risks. The activity started with reading the HAI-A analysis sheet. The system analyst role was to lead the session, ensure all the templates were completed, and support the AI expert in accessing the available information and resources about each XAI method. A summary of the main findings is provided below. Samples of the completed templates are also provided (Figure 48-49).

- AI expert mentioned that classifying the XAI method assessment into five main dimensions (*functional, operational, usability, safety, and validation*) was useful as it provided a clear direction to follow. This enabled the AI expert to search for the related information about the XAI method easily.
- The AI expert mentioned that the amount of writing in the templates was adequate and did not take much effort to complete; this was not a challenging task for the expert and did not lead to a loss of interest in the activity.
- The AI expert accentuated that assessing the XAI method is crucial for successful calibrated trust design and the activity was not difficult to accomplish for one XAI method identified in the HAI-A sheet. However, AI experts mentioned that the more XAI methods required for the task, the more AI experts should participate in this activity in real-world scenarios. AI experts also mentioned that such activity might not be adequate to complete in one session, and it may require several sessions to analyse and assess the XAI method critically.
- The researcher observed that the AI expert did not follow a particular order of completing the templates, i.e., the AI expert ignored the provided order. AI expert mentioned that it would be helpful to complete the templates in any order based on the availability of the information. The researcher clarified that while the order has no effect on the final assessment outcome, ordering the templates ensures that the templates shall be completed systematically and prevent confusion and missing out on some templates.
- The researcher observed that the system analyst and the AI expert discussed and noted that functional assessment might not be useful to detect trust calibration risks, e.g., the AI expert questioned how the machine learning family assigned to LIME method could help trust calibration process. The researcher made it clear that the functional dimension is expected to assess the XAI method applicability with the task and the underlying machine learning model. Such an assessment is less likely to introduce users' errors, but it ensures that the XAI method is appropriate for the current task.
- The researcher noticed that templates questions lacked additional space. AI experts struggled to write additional information related to each explanation feature introduced in the templates (see Figure 50). Therefore, the templates are extended with extra space at the bottom of each template.
- The researcher observed that the participatory design approach adopted by the C-XAI method aided in an effective XAI method assessment. Having AI experts involved in the design process enhanced the understandability of the assessment phase and ensured the applicability of the XAI method to the Human-AI task.

- The AI expert was proactive during the assessment phase and noted several concerns to the system analyst regarding further C-XAI stages. For instance, when assessing the LIME audience, AI experts asked the system analyst to include a note to the design team. AI experts mentioned that designers should guide the user on interpreting LIME feature values and generating meaningful interpretations.

SHEET ID 3.3		Usability Requirements assessment	
TASK ID: Screening prescription		Explanation method: LIME	
How truthful is the explanation method with the underlying ML model?			
This feature needs to be empirically evaluated based on the underlying algorithm and cannot be calculated with the available information.			
How well the generated explanation can be generalised beyond a particular recommendation?			
Lime explanations are not complete and cannot be generalised beyond specific recommendations.			
Describe any contextual information that shall be accompanied by the explanation?			
Users should be informed that each explanation is unique and only applicable for the current recommendation.			
Does the explanation method allow interactive interaction?			
What is the explanation interaction method?			
<input type="radio"/> Static <input type="radio"/> Interactive			
What are the explanation audience knowledge requirements?			
<input type="radio"/> AI experts <input type="radio"/> Task experts <input type="radio"/> Lay audience <input type="radio"/> Other: users should be able to interpret the meaning of the features.			
What is the function of the explanation method?			
<input type="radio"/> Transparency of the ML model <input type="radio"/> Fairness of the ML model <input type="radio"/> Accountability of the ML model <input type="radio"/> Other -----			
Does the explanation method have a casual nature?			
<input type="radio"/> Yes <input type="radio"/> No			
Does the explanation method have an actionable nature?			
<input type="radio"/> Yes <input type="radio"/> No			
NOTE: Designer might need to use grouping feature to help users interpret the meaning of the features.			

FIGURE 49 SAMPLE RESPONSES FOR ASSESSING LIME USABILITY REQUIREMENTS

<input type="radio"/> Yes <input type="radio"/> No Could be done on the design level
What is the level of actionability in the explanation method?
<input type="radio"/> Yes <input type="radio"/> No
Does the explanation inherent the time order of the event?
<input type="radio"/> Yes <input type="radio"/> No
Does coherence implement the explanation method?
<input type="radio"/> Yes <input type="radio"/> No
What is the degree of explanation method novelty?
Low
Can the complexity of the explanation be adjusted?
Low

9.8.2.3 CONDUCT TRUST CALIBRATION RISK ASSESSMENT

This is the fifth activity for C-XAI design method. Stakeholders involved in this activity were medical domain experts, HCI expert, psychologists and system analysts. The primary goal of this activity was to identify XAI properties that may trigger trust calibration risks and identify those risks. In addition to **SHEET ID 4**, the system analyst provided stakeholders with the completed templates (**SHEET ID 3.1 – 3.5**). This was an essential input for this activity as stakeholders will analyse the completed templates. The main findings from this activity are presented below. A sample of participants answers is also provided in Figure 50.

- Participants agreed that classifying the risks into skipping and misapplying was useful for them as it provided a clear direction to follow. This prevented them from overlooking or ignoring trust calibration risks. Furthermore, some participants stated that the format helped them understand the nature of the design problem.
- Identifying trust calibration risks, such as perceived complexity, was straightforward for the participants. They were able to locate different risks based on the provided templates and documents. The system analyst and the participants were satisfied with the examples provided to them on the ‘when’ and ‘how’ a trust calibration risk could happen. They agreed that such examples would be vital in the successful identification of trust calibration risks. During the study, the researcher ensured that any help to the participants was provided.
- The researcher observed that participants discussed and noted that some of the provided risks might not be relevant considering long-term Human and XAI interaction, e.g., redundant information risk might not be a potential risk at the earlier stages of using the XAI interface. However, this risk might be essential and need to be reassessed in a long-term interaction. The researcher made it clear that the C-XAI method is an

iterative process, and re-evaluation of the XAI requirements should be done frequently to make the necessary adjustments.

SHEET ID: 4.2	Trust Calibration Risk assessment – Operational
Dimension	
TASK ID: Screening prescription	Explanation method: LIME
Provide a title for this assessment - This section depicts what trust calibration risk is.	
Complexity and misinterpretability.	
Risk identification. Based on the Operational dimension assessment sheet, what kind of risks can be identified?	
Please identify which XAI operational properties might introduce risks to the current Human-AI task?	
<ul style="list-style-type: none"> <input type="radio"/> Explanation class <input type="radio"/> Explanation communication <input type="radio"/> Explanation interaction <input type="radio"/> Explanation audience <input type="radio"/> Explanation goal <input type="radio"/> Causality <input type="radio"/> Actionability <input type="radio"/> Other: ----- 	
Risk analysis. Based on the selected properties, what kind of risks can be identified?	
Skipping. <ul style="list-style-type: none"> <input type="radio"/> Lack of curiosity <input type="radio"/> Perceived goal impediment <input type="radio"/> Redundant information <input type="radio"/> Perceived complexity <input type="radio"/> Lack of context <input type="radio"/> Unfamiliarity <input type="radio"/> Other: ----- 	
Misapplying. <ul style="list-style-type: none"> <input type="radio"/> Misinterpretation <input type="radio"/> Mistrust <input type="radio"/> Confirmatory search <input type="radio"/> Rush understanding <input type="radio"/> Habits formation <input type="radio"/> Other: ----- 	
Risk frequency and potential probability.	
Users might have these risks frequently; expert doctors might face difficulties each time to interpret feature-based values.	
Risk assessment. Please determine how acceptable is the risk on the trust calibration in the Human-AI task.	
<input type="radio"/> No risks have been identified	
<input type="radio"/> The risks require changing the explanation method	
<input type="radio"/> The risks can be mitigated on the design level	
<input type="radio"/> The risks have little effect on trust calibration	
Is there any additional information that you would like to be considered or you think is relevant?	
<ul style="list-style-type: none"> • Statistical numbers as an explanation medium might be perceived as complex by users. • LIME explanation might increase the number of data features in the interface. • The inability to generalise the explanation might cause future risks of misinterpretability. <p>We recommend the designer consider mapping the statistical numbers to a more understandable and contextual interpretation. For instance, feature contribution with the value of 0.75 can be communicated to users as a high contribution to AI-based decisions. Furthermore, for reducing the complexity, a threshold for the contributed features shall be 30%, which can be highlighted.</p>	

FIGURE 50 SAMPLES OF PARTICIPANTS ANSWERS TO TRUST CALIBRATION RISK ASSESSMENT

9.8.2.4 THE SELECTION AND IMPLEMENTATION STAGE

Designers were the main stakeholder in this activity. The completed templates (**SHEET ID 4.1 – 4.5**) from the Trust calibration risk assessment was the main input for this activity. The system analyst role was to answer design team questions related to **SHEET ID 4**. The main goal of this activity was to mitigate trust calibration risks following appropriate design principles. The design principles were provided to the design team in **SHEET ID 5**, supported by several examples. The researcher ensured that the design principles were clearly explained to the design team. The design team were asked to implement appropriate design principles in the XAI interface to mitigate trust calibration risks. The system analyst supervised the process and helped clarify any questions that the designers might have. The final decision of implementing the design principle was made collaboratively between the design team. A discussion between the design team and the system analyst was also done to clarify the task characteristics. The following bullet points summarise the findings. A summary of design techniques chosen in this stage to combat trust calibration risks is presented in Table 34.

- The system analyst and design team agreed that the design principles and the description were precise and useful. It helped them to follow each principle and evoke a brainstorming on how the design would look like. In addition, participants stated that the principles presented in SHEET ID 5 help raise their awareness of the various possible techniques that can enhance the trust calibration problem and mitigate potential risks. SHEET ID 5 also supported them to focus on interface constraints and persuasion techniques to engage XAI users in a more reflective thinking process.
- The participants recognised that providing details of how to implement selected principles and their recommended risks gave them a direction and focused on what is required. Furthermore, design techniques and prototypes provided in the supporting documents were valuable and helped them think about further techniques. It was noted that designers spent time reflecting on the design's role to mitigate trust calibration risk regarding the task characteristics. When asked why this was the case, P8 replied: *“it is easy to understand what is required regarding this risk but elaborating on the task characteristics can be time-consuming”*. Also, they emphasised that the re-evaluation of the initial design in the task context and re-expressing the requirements will take less time as they become familiar with the various aspects of the task. The design team wanted a more descriptive analysis of users' behaviour during their everyday decision-making to select appropriate interaction design techniques. This information was not available during the selection and implementation activity.
- During the prototyping activity, the researcher noticed that participants struggles to implement the Learning principle in the XAI interface design. SHEET ID 4 indicated

contextual information shall be accompanied with LIME explanations but did not specify exactly this information.

- It was observed that the design team were mainly familiar with attention guidance and persuasive design principles. They have little understanding of challenging habitual actions and learning principle.

TABLE 34 SAMPLES OF IMPLEMENTED DESIGN TECHNIQUES USED BY PARTICIPANTS

Identified trust calibration risk	Explanation method	Selected decision principle	Selected design technique
Misinterpretation	LIME	Learning and training	Onboarding technique
Solution description	Participants wanted to solve the misinterpretation risk of LIME explanation by implementing an onboarding design technique. Onboarding design technique contained detailed information about the explanation limitations and capabilities. Participants also included examples of how to interpret the explanation and undesired ways of interpreting the explanation. They incorporated this technique as a constraint on the AI-based tool where participants need to complete this stage before using the AI-based tool.		
Perceived complexity	LIME	Attention guidance	Abstraction contextualisation
Solution description	Participants divided the explanation into multiple abstraction levels to make it easier to read and recall the meaning of the explanation. They also added contextual information to the data features names and values.		

9.8.2.5 EVALUATION STAGE

The evaluation goal of this stage is not to evaluate the generated prototype but to evaluate C-XAI evaluation guidance and the communication between different stakeholders. Prototypes created from the selection and implementation stage were evaluated with three participants who used AI-based screening prescription before. The design team used a think-aloud protocol to evaluate the prototypes. The design team were supported with ten different screening prescriptions scenarios to implement their designs and validate the prototypes with participants. The design team and system analyst encouraged participants to explore the prototype and think-aloud (1) What they say on each screen, (2) What they thought could be improved, and (3) Whether the design fits their workflow. The design team also used two calibrated trust behavioural indicators measures: *Compliance and correct decisions*. The design team also noted participants behaviour and asked participants about their understanding of each AI-based recommendation. The design team goal was to know whether participants perceived the XAI interface as useful, understandable and engaging.

The collected observations from this stage were analysed. As a general observation from the first iteration, participants engaged with the XAI interface. They noted that the XAI interface helped them identify relevant information about the AI-based recommendation. However, participants were confused with the Onboarding technique. They did not want to spend time

learning about the explanation before using it. The design team solved this issue by including an additional view in the interface, which included the learning material that helps users of XAI interpret the explanations (See Figure 51). The overall evaluation needed two design iterations to be completed. While these results indicate the effectiveness of the generated prototypes in terms of explanation effectiveness and XAI interface engagement, room for improvement regarding integrating the explanation in the task workflow was found. For example, participants wanted to generate a textual explanation from a predefined template to store it in the patient profile. The evaluation allowed the design team to demonstrate the effectiveness of the developed prototypes in the task workflow and their ability to engage the users with the XAI interface.

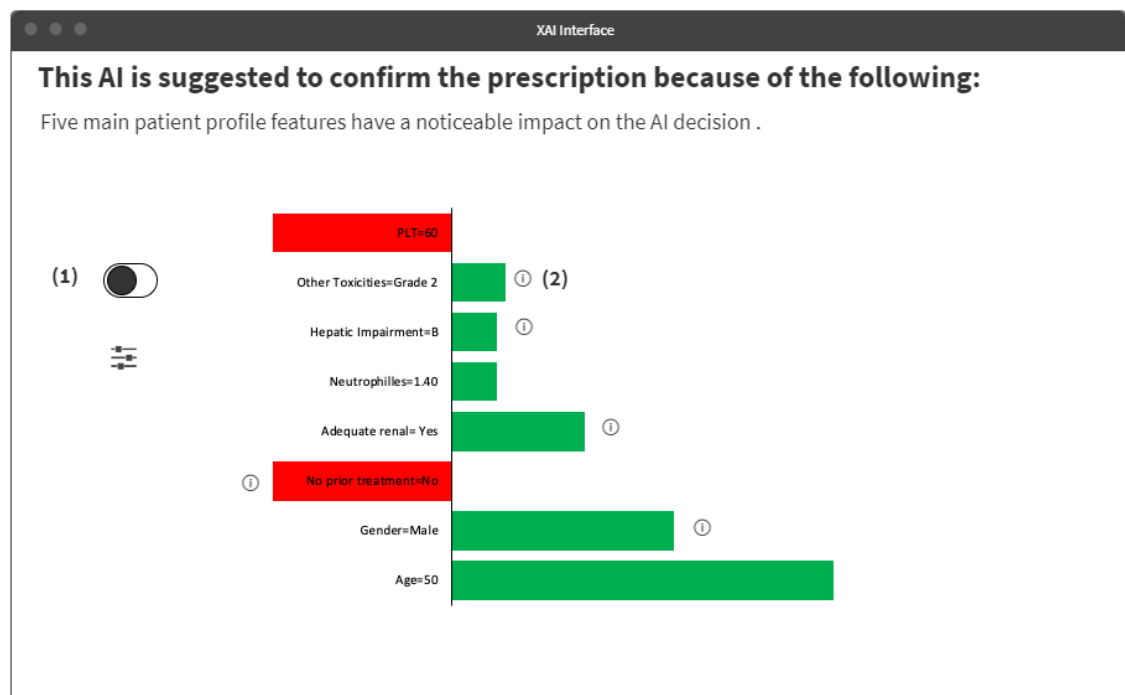


FIGURE 51 FINAL XAI INTERFACE GENERATED FROM C-XAI METHOD

The following bullet points summarise researcher observations from this activity.

- The system analyst and design team agreed that the provided measurements and methods were clear and helpful to achieve the design goal. C-XAI evaluation stage gave the participants a clear direction to follow and assess which methods and measurements are needed. One of the design teams noted that “*Each measurement has sufficient description that helps us to apply and to identify the best one*”.
- Participants noted that the available time and the number of participants were the main two issues for a comprehensive evaluation. One participant mentioned, “*well ... we still cannot assure that these designs would work in real-world scenarios*”.
- Participants noted that task analysis lacked a clear direction of guiding the design team to choose the measurements. They noted that cost-benefit analysis and trade-off

between the measurement shall be considered at the earlier stages of the system analysis.

9.9 BENEFITS OF C-XAI IN THE DESIGN PROCESS

Participants utilised C-XAI in the design process to design an XAI interface for screening prescription tools. Participants completed the GAQ consisting of two sections, the first section related to evaluating the effectiveness of C-XAI during the design process. While the second section is related to the potential improvements that can be performed to the method. In general, all participants emphasised that the C-XAI was useful and effective in the design process. The following sections discuss the benefits and risks of using the method that emerged from the GAQ analysis. The following points also discuss the final amendments made to the C-XAI method based on researcher observations and GAQ analysis.

9.9.1 EFFECTIVE COMMUNICATION AND INCREASED ENGAGEMENT

C-XAI are evaluated in terms of facilitating effective and clear communication between different stakeholders. Adopting participatory design and specifying stakeholders' roles and tasks during the design process provided a clear direction to each participant. Due to the fact the literature lacked a straightforward method to guide the XAI design for calibrated trust goal, C-XAI led to an increase in the engagement of different stakeholders during the design process. However, some communication issues were revealed during the evaluation based on the researcher observations and participants' feedback to enhance the ease of using the method.

Regarding the time needed to complete the explanation method assessment sheets, it was observed that AI experts required more time to provide complete and comprehensive feedback to each explanation method. In this study, one explanation method assessment activity was completed. This was not a challenging task to complete for the AI expert, but the system analyst had to provide some help by searching for relevant literature and providing access to research databases. Consequently, it is recommended that the AI expert in this stage may need to complete this stage in multiple days or multiple sessions. This would ensure that the explanation properties are assessed precisely. P8 mentioned, *"assessing the explanation method need more time so the AI team in any software development would need multiple sessions across several days to collect information about the method and provide accurate assessment"*.

Participants required open-ended space in each template to add general notes and recommendations to the design method's subsequent activities. They argued that adding extra space in each template would enhance the communication between different design methods. During the sessions, participants wanted to record observations and point them out to stakeholders in the next activity. For instance, the AI expert insisted on informing the design

team about the importance of helping users interpret LIME explanations. Participants mentioned, *“Designer must know that these explanations are recommendation-specific explanations”*.

9.9.2 INCREASED FOCUS ON THE DESIGN PROBLEM

C-XAI method is meant to increase the design team's focus on potential risks during the interaction between human decision-makers and the XAI interface. It is meant to help the design team and different stakeholders in having a shared understanding of the potential users' errors. It is also to make the XAI interface effective in mitigating these errors. The observer indicated that the C-XAI method helped participants understand the problem from actual context rather than their assumptions. C-XAI method has well-designed templates and recommendations and can assist stakeholders in the design process as a reference point. C-XAI method was created to focus the attention on XAI interface design when trust calibration is the main goal, i.e., supporting users' in applying cognitive thinking and forming correct interpretations of a given explanation. Most responses in the GAQ expressed that C-XAI helped increase the focus on trust calibration problem and removed the focus from dealing with the explanation as informational content only, e.g., *“the method reflects the reality of trust calibration problem and have a link to many explainable AI literature and human behaviour as well ... it helped me to focus my attention on combining these three areas together”*.

9.9.3 INCREASED EMPATHY TOWARDS USERS

C-XAI is meant to help the design and development team to build understanding towards the users and recognise their errors, needs and context. The term empathy means understanding and predicting people behaviours and psychological states. The observer indicates that using C-XAI templates allowed the design and different stakeholders to identify and feel empathy for the end-users and understand their errors. With the C-XAI method in mind, creative design ideas were proposed with the aim of supporting users in calibrating their trust. Most responses in the GAQ expressed that the C-XAI was useful and effective to react to and empathising with users' errors, e.g., the method templates enabled the design team to closely understand the user errors and behaviours that triggered the designers to set solutions. P1 stated, *“I found method shorten the distance between designers and real users”*.

9.9.4 BETTER SOFTWARE PRODUCT DESIGN

C-XAI is meant to help produce better designs with a more valuable set of design features and make design decisions more informed. Understanding the potential behaviour and users' errors can facilitate new ideas and features, informing better design. During the evaluation, the observer indicates that the designers focused on specific trust calibration risks and tried to

mitigate those risks. This helped the designers make better decisions by focusing on a specific problem and referring them to support their design ideas. Most responses in the GAQ expressed that C-XAI was valuable and practical for better software design, e.g., *“using this method designers would set the right features for the right problem and risk”*.

9.9.5 RISKS OF USING C-XAI METHOD IN THE DESIGN PROCESS

- The method did not provide sufficient information about the nature of trust calibration risks which made participants struggle and rely on the psychology expert during the trust calibration risk assessment activity, P6 struggled to contextualise rush understanding risk and indicated, *“I was not really able to reflect how people would rush understand an explanation ... I expected more elaboration”*. A supporting document was added to the method to elaborate on each of the potential trust calibration risks.
- Five participants mentioned that the amount of effort to complete the templates could increase exponentially when the number of tasks and XAI method increase. This led them to recommend future recommendations to the method, e.g., ensure that an adequate number of stakeholders should be recruited based on these parameters. Therefore, C-XAI v3 included a recommendation to the system analyst to consider this point during the participants' recruitment process.
- Trust calibration risks include users' errors and potential cognitive biases during Human-AI collaborative decision-making tasks. Also, the method included templates that reveal several limitations of the XAI method. This information facilitated the stakeholder to make assumptions about the end-users and XAI methods and they may generalise these assumptions over other XAI methods or tasks.
- C-XAI booklet was confusing to participants. Many participants had to ask questions to the system analyst to guide them through the booklet. Many responses in GAQ asked to split the documents into several documents, where each document is only relevant for the current activity. Other participants were also asked to specify each stage's input and output at the beginning of each activity. P5, a UX expert, mentioned, *“the booklet contains so much relevant information for each activity ... I would recommend splitting it into multiple files”*.

9.10 EVALUATION VALIDITY

This section discusses the main threats to the validity of the evaluation study.

- The activities and the amount of time assigned to them were based on the research team experience and estimations. This factor might affect the quality of the information

provided by participants. However, the researcher asked the system analyst before starting each activity that additional time can be given to participants to complete the templates.

- Participants might misinterpret that the supporting documents and the assessment templates. Considering this potential limitation, the researcher made it clear that the participants and system analysts were not restricted to the template's elements; therefore, they were given options to add additional elements on the templates that they deem relevant to explainability and trust calibration.
- While many participants may have benefited from the evaluation study to help ensure the elicitation of diverse viewpoints, the online environment may have limited participants engagement in the activities. However, the system analyst and the researcher ensured that all participants gave each stage feedback by asking the participants questions.
- The limited-time and difficulties to approach participants during the COVID-19 circumstances limited the number of recruited participants to 15. However, all participants involved in the evaluation study were experts in their domains, which meant that various requirements, constraints and assessments were still considered. As the intention here is a proof of concept, therefore the number of participants would not affect the design method's results.

9.11 CHAPTER SUMMARY

This chapter has evaluated the proposed C-XAI method for designing XAI interfaces to enhance the trust calibration process. The chapter discusses the process employed to evaluate the method. The chapter's outcome showed that the C-XAI method helped the analyst and design team identify and understand the design problem. A case study approach was employed to assess the potential of the C-XAI method to aid the design process. The thesis conclusion and future work will be discussed in the next chapter.

10. CHAPTER 10: DISCUSSION, FUTURE WORK AND CONCLUSION

A fundamental success for AI-based assisted decision-making tools is to support users in forming a correct mental model of the system (Bansal et al., 2019). That is, human decision-makers need to know when to trust or distrust the AI-based recommendations. When users fail to calibrate their trust, collaborative decision-making would be affected, and dramatic failures could happen in high-stakes application domains. The research community discussed the challenges for humans to understand and develop an accurate mental model of an AI since opaque ML black-box models are increasingly used (Springer, 2019). For instance, users may fail to follow up with the AI-based recommendations because of their dynamic and uncertain nature (Zhang et al., 2020a). In such cases, users might follow an incorrect recommendation (over-trust) or reject a correct recommendation (under-trust). A design goal that aims to attain and manage trust refers to calibrated trust. This thesis emphasised that this goal is distinct from enhancing trust in AI. For example, enhancing users' trust could be done by providing indicators and metrics for AI abilities and performance (Yin et al., 2019). Enhancing trust also does not require AI users to understand and develop accurate mental models of the AI. On the other hand, calibrating users' trust may require extra effort from the users' and shall be done on the recommendation level where decision-makers can identify situations when to rely on the AI and use their judgment (Yu et al., 2019).

Humans' decision-makers require a user interface to reflect the current state or logic of the recommendation to attain trust calibration. Studies such as (Muir, 1994, Zhang et al., 2020b, Yang et al., 1994) have emphasised the role of XAI in calibrating users' trust. Such communication is a major facilitator of trust calibration within a Human-AI collaborative decision-making task. Explanations help calibrated trust by showing the rationale and reasoning behind single recommendations and their overall logic (Cai et al. 2019). However, such studies often assumed that users would engage cognitively with explanations and form correct interpretations from them. Recent studies showed that even though explanations are communicated to people, trust calibration is not improved (Zhang et al., 2020a, Bussone et al., 2015). Such failure of XAI systems in enhancing trust calibration has been linked to factors such as humans' cognitive biases, e.g., people are selective of what they read and rely on (Naiseh et al., 2020b). Also, others showed that XAI failed to improve calibrated trust because of undesired human behaviour with AI-based explanations, e.g., human laziness to engage in what they perceived as effortful behaviour (Wagner and Robinette, 2021). Overall, users of XAI systems fail, on average, to calibrate their trust, i.e., human decision-makers working collaboratively with an AI can still be notably following incorrect recommendations or rejecting correct ones.

Recently a surge of advances in explainability led to increasing interest in model-agnostic explanation methods which interpret any machine learning model by focusing primarily on the input and the output of the machine learning model (Hohman et al. 2019). Model-agnostic explainable models generate different classes of explanations to answer different user questions (Carvalho et al. 2019). This approach is motivated by preserving the confidentiality of the model and also increasing the cost-efficiency of generating the explanation and increasing its usability (Feng et al. 2019, Zhang et al. 2020, Wachter et al. 2017). Different model-agnostic methods may generate explanations with distinct explanation output, but they may vary in their performance, fidelity and completeness of the underlying AI model (Arrieta et al., 2020). Hence, model-agnostic explanation methods have a wide range of properties and features that may not be appropriate for a particular Human-AI task or require an intervention on the design level to operationalise these explanations. For example, an XAI method that has a high probability to generate novel explanations may need the design team to make users' engagement with the explanation in every interaction essential.

As a result, this research proposes the need for a systematic approach to support XAI designers in developing XAI interfaces that engage users with AI explanations based on the properties of the underlying XAI model. This thesis has conducted several explanatory studies and a literature review to develop an understanding of the research problem. The findings from these studies were used to develop a design C-XAI design method. C-XAI design method aid the design of XAI interfaces to make the interface engaging and support users in interpreting and contextualising the explanation. Achieving the aim of the design is always paired with the underlying properties of the XAI model. C-XAI can have a useful implementation under the term of responsible AI design in practice. There are several benefits to the C-XAI method in the industry:

- Capturing trust calibration risks from real context, e.g., by proposing trust calibration risk assessment.
- Involve wide and diverse members in the design phase to the liability and responsibility of the proposed design.
- Inform the decisions of the design, e.g., by introducing empirical-based design recommendations for XAI interface.
- Parts of the method, e.g., assessment or design phase, can be used as an evaluation tool to current user interfaces in industry.

To conclude this thesis. **Section 10.2** describes the thesis contributions to knowledge, followed by thesis limitations in **Section 10.3**, as well as potential future work in **Section 10.4**.

10.1 RESEARCH QUESTIONS AND OBJECTIVES REVISITED

As outlined in the introductory chapter of this thesis, the purpose of this research was to investigate.

Objective 1: To conduct a literature review on calibrated trust design, XAI and related areas.

The author reviewed the literature from different areas to get more insights and understanding. Domains such as XAI design, XAI algorithms and models, trust and trust calibration were reviewed. Moreover, the review served to find different requirements and patterns of designing Human-Automation interfaces. This includes a way of personalising the XAI interface, modalities of interacting with explanations and delivery methods. The literature review concluded with a comprehensive understanding to inform the design of further explanatory studies.

Objective 2: To provide a taxonomy of XAI methods classes.

A systematic literature review was conducted to provide a taxonomy for different explanation classes in the literature of XAI. The research then mapped several users' questions that each explanation class could answer. Finally, the research conducted an expert evaluation to evaluate the completeness, coherence and understandability of the generated taxonomy. This objective aimed to provide a reference point for the stakeholders of the C-XAI method about what can be explained to users given a black-box model. Furthermore, the results from this stage were used in further explanatory studies.

Objective 3: To explore the lived experience of XAI users in a Human-AI collaborative decision-making task.

To achieve Objective 3, a multi-stage mixed approach was followed. This included three studies (within-subject experiment, semi-structured interviews and think-aloud protocol). This approach aimed to explore the lived experience of participants during a Human-AI collaborative decision-making task and their usage and interaction style with XAI interfaces. This was meant to identify potential users' errors and difficulties while interacting with AI-based explanations. Furthermore, this approach aimed to explore the role of explanation class in calibrating users' trust and what makes the explanation effective in calibrating users' trust. This study used a think-aloud protocol during a decision-making task study, semi-structured interviews, and within-subject experiments in terms of data collection. This allowed the research to explore different trust calibration risks and enable participants to discuss their opinions and concerns

regarding AI-based explanations in their everyday Human-AI collaborative decision-making task.

Objective 4: To propose XAI interface design principles that help trust calibration.

The research explored design principles that can enhance the role of explainability in calibrating users' trust. Based on the findings from Objective 3, the research identified various trust calibration risks and requirements for the XAI interface to calibrate users' trust. Results from Objective 4 illustrated how to implement AI-based explanations in an XAI interface to enhance explanation role to calibrate users' trust. A co-design approach was followed to gather how the solution would look from the users' perspective. In this stage, the researcher discussed and negotiated with representative users' ways of utilising AI-based explanations to serve their needs, enhance trust calibration, and mitigate potential trust calibration risks. This had been achieved by giving the participants initial prototypes of the design problem to help them visualise the idea and then provoke brainstorming related to the design problem. The deriving framework to discover protentional design solutions was digital nudging (Caraban et al.) and de-biasing principles (Soll et al., 2014).

Objective 5: To create, evaluate and refine a method for helping trust calibration in the XAI interface.

The researcher developed a method to assist designers and system analysts in developing XAI interfaces to help trust calibration. The method adopts a participatory design approach to ensure that all relevant stakeholders are involved in the early and later stages of XAI interface design. The method includes seven activities supported with templates and supporting documents from previous objectives. The method was evaluated using a case study approach to assess the effectiveness, completeness and clarity of the methods. The evaluation also focused on the engagement and communication between stakeholders.

10.2 CONTRIBUTION TO KNOWLEDGE

This thesis has contributed to the knowledge of XAI interface design focusing particularly on mapping the design with the underlying XAI model properties to support trust calibration. In the next sub-sections, the main contributions of this thesis will be provided.

Conceptualising trust calibration problem in relation to XAI interface design

The first contribution of this thesis is the conceptualisation of XAI interface design and trust calibration. The discussions and results presented earlier in this thesis would help future research to understand why the current utilisation of explanations for Human-AI tasks is not effective in supporting trust calibration. This has been achieved through a multi-stage mixed approach (within-subject experiment, think-aloud protocol and semi-structured interviews).

Eliciting users' errors that limit XAI interface in supporting trust calibration

The second contribution of this thesis is exploring the real-world experience of XAI users who used AI-based tools before and understanding potential users' errors that limit the role of XAI interface to support trust calibration. The approach followed a think-aloud protocol for data collection, where 16 medical experts used screening prescription AI-based tools in their settings. In the past, most of the empirical literature concerning trust calibration and XAI interface design has relied on retrospective data collection methods and non-expert users for Human-AI tasks, i.e., they usually approach participants with no experience in the Human-AI task. This presents limitations related to the potential for recall bias and questions regarding ecological validity.

Proposing XAI design principles to support trust calibration

The thesis proposed several design implications that point towards supporting the interface design with techniques and principles to increase users' interaction with the XAI interface to help trust calibration. For example, the thesis results suggest that presenting explanations for trust calibration should mainly be designed to avoid undesired behaviour such as skipping explanation and habits formation. This contribution was based on a qualitative approach that provides a detailed look at explainability and trust calibration. The approach consisted of two qualitative phases: (a) exploration phase, which aims to provide a contextual understanding of the problem, (b) design phase to reveal main concepts and design techniques that improve the role of explanation in trust calibration.

A systematic approach to design XAI interfaces with trust calibration goal in mind

The fourth contribution of the thesis is a method for designing XAI interfaces to support trust calibration C-XAI. The proposed method includes various templates for the different stakeholders to express their concerns and potential trust calibration risks regarding the underlying XAI model (see Chapter 8). The proposed method will enable various stakeholders such as system analysts, psychologists, designers, and AI experts to participate in the design stage. The templates were developed primarily based on the thesis findings.

10.3 THESIS LIMITATIONS

The research used convenience sampling, and all participants were volunteers, which may have biased the sample. It would also provide higher validity if the sample size were larger. Further, as the sample was recruited from the mailing list and included participants from the UK, it was not fully representative.

The studies covered a range of trust calibration scenarios using a case study from the medical domain. Also, these studies were conducted online using hypothetical scenarios and focused on

screening prescription as an ML classification problem. These settings were chosen due to participants availability and time constraints during the COVID-19 situations. This shed the light on another bias that indicates that there are still missing additional requirements and trust calibration risks due to the adopted case study and participants' background. There is a need for different research method designs to increase the validity of results from this thesis in other case studies.

In Chapter 5, the trust questionnaire was used to measure participants' trust. This instrument is simple that does not cater for the complexity of trust. However, other behavioural indicator measures were used to mitigate the effect of self-reported trust measurement.

This thesis has mainly targeted high stakes application domains, where participants have good knowledge about the studied task. Approaching these participants may limit the findings to those individuals. Also, the choice of using screening prescription case studies may have a potential influence on the results, i.e., analysing other applications might lead to discovering different types of errors and additional concepts related to trust calibration risks.

Finally, the initial version of C-XAI was itself based on substantial qualitative and quantitative studies. In addition, C-XAI is supported by previous studies, so their creation was both empirical and literature-based. However, the C-XAI output is not meant to guarantee that the generated designs would improve trust calibration. Instead, the author advocates that the method is a starting point for a variety of processes, such as the requirements of elicitation, tailoring, and self-diagnosis, both for the XAI model properties and trust calibration risks. The method by itself is effective in understanding the design problem and providing aid to multiple stakeholders during the XAI interface development stage.

10.4 RECOMMENDATIONS FOR FUTURE RESEARCH

The findings of this thesis have vital implications. However, more research is still needed to draw more compatible conclusions about linking specific XAI model properties and XAI classes to trust calibration risks. For instance, the perceived complexity of an explanation could be correlated with only a particular explanation class.

Future research needs to consider using other methods and data sources based on the data collection methods, perhaps based on objective measures and more in real-time. For example, a diary study could be used to collect longitudinal and temporal information, report events and experiences in context and the moment. This may result in a better understanding of the user experience and behaviour. Additionally, in this thesis, the C-XAI design method was evaluated using a case study evaluation approach. The author recommends similar studies for future work conducted with different case studies and larger representative samples to enable generalisation and the validity of the generated designs from the C-XAI design method. Moreover, future work

may investigate the role of user knowledge and experience in the Human-AI tasks in trust calibration and how this would improve the C-XAI design method.

11. REFERENCES

- Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y. and Kankanhalli, M., 2018, April. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-18).
- Abraas, C., Maloney-krichmar, D. & Preece, J. 2004. *User-Centred Design*, Encyclopedia of Human-Computer Interaction, 2014. Sage Publications, Thousand Oaks.
- Adadi, A. & Berrada, M. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160.
- Aigner, W. & Miksch, S. Supporting protocol-based care in medicine via multiple coordinated views. 2004. *IEEE*, 118-129.
- Aleven, V., Stahl, E., Schworm, S., Fischer, F. & Wallace, R. 2003. Help-seeking and help design in interactive learning environments. *Review of educational research*, 73, 277-320.
- Amadiou, F., Mariné, C. and Laimay, C., 2011. The attention-guiding effect and cognitive load in the comprehension of animations. *Computers in Human Behavior*, 27(1), pp.36-40.
- Andreou, A., Venkatadri, G., Goga, O., Gummadi, K., Loiseau, P. and Mislove, A., 2018, February. Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In *NDSS 2018-Network and Distributed System Security Symposium* (pp. 1-15).
- Annett, J., 2003. Hierarchical task analysis. *Handbook of cognitive task design*, 2, pp.17-35.
- Apley, D.W. and Zhu, J., 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4), pp.1059-1086.
- Apter, M., 1997. Reversal theory: what is it?. *PSYCHOLOGIST-LEICESTER-*, 10, pp.217-220.
- Armstrong, J.S. ed., 2001. *Principles of forecasting: a handbook for researchers and practitioners* (Vol. 30). Boston, MA: Kluwer Academic.
- Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, pp.82-115.
- Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A. and Mourad, S., 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.
- Ashby, W.R., 1961. *An introduction to cybernetics*. Chapman & Hall Ltd.
- IEEE Standards Association, 2018. The IEEE global initiative on ethics of autonomous and intelligent systems. https://standards.ieee.org/developlindconn/ec/autonomous_systems.html.
- Banerjee, A.V., 1992. A simple model of herd behavior. *The quarterly journal of economics*, 107(3), pp.797-817.
- Bansal, G., Nushi, B., Kamar, E., Weld, D.S., Lasecki, W.S. and Horvitz, E., 2019, July. Updates in human-ai teams: Understanding and addressing the performance/compatibility

- tradeoff. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 2429-2437).
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T. and Weld, D., 2021, May. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-16).
- Barocas, S., Selbst, A.D. and Raghavan, M., 2020, January. The hidden assumptions behind counterfactual explanations and principal reasons. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 80-89).
- Barria-Pineda, J., Akhuseyinoglu, K. and Brusilovsky, P., 2019, June. Explaining need-based educational recommendations using interactive open learner models. In Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (pp. 273-277).
- Bass, E.J., Zenyuh, J.P., Small, R.L. and Fortin, S.T., 1996, August. A context-based approach to training situation awareness. In Proceedings Third Annual Symposium on Human Interaction with Complex Systems. HICS'96 (pp. 89-95). IEEE.
- Bastani, O., Kim, C. and Bastani, H., 2017. Interpretability via model extraction. arXiv preprint arXiv:1706.09773.
- Baxter, P. & Jack, S. 2008. Qualitative case study methodology: Study design and implementation for novice researchers. The qualitative report, 13, 544-559.
- Bennetot, A., Laurent, J.L., Chatila, R. and Díaz-Rodríguez, N., 2019. Towards explainable neural-symbolic visual reasoning. arXiv preprint arXiv:1909.09065.
- Berg, B.L., 2004. Methods for the social sciences. Qualitative Research Methods for the Social Sciences. Boston: Pearson Education, p.191.
- Bergus, G.R., Chapman, G.B., Gjerde, C. and Elstein, A.S., 1995. Clinical reasoning about new symptoms despite preexisting disease: sources of error and order effects. Family medicine, 27(5), pp.314-320.
- Berkovsky, S., Taib, R. and Conway, D., 2017, March. How to recommend? User trust factors in movie recommender systems. In Proceedings of the 22nd International Conference on Intelligent User Interfaces (pp. 287-300).
- Berner, E.S., 2009. Clinical Decision Support Systems: State of the Art Agency for Healthcare Research and Quality. Rockville, Maryland.
- Bhaskar, R., 2013. A realist theory of science. Routledge.
- Bien, J. and Tibshirani, R., 2011. Prototype selection for interpretable classification. The Annals of Applied Statistics, 5(4), pp.2403-2424.
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J. and Shadbolt, N., 2018, April. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 Chi conference on human factors in computing systems (pp. 1-14).
- Biran, O. and Cotton, C., 2017, August. Explanation and justification in machine learning: A survey. In IJCAI-17 workshop on explainable AI (XAI) (Vol. 8, No. 1, pp. 8-13).
- Blundell, C., Cornebise, J., Kavukcuoglu, K. and Wierstra, D., 2015, June. Weight uncertainty in neural network. In International Conference on Machine Learning (pp. 1613-1622). PMLR.

- Boehm, B., 1989, September. Software risk management. In European Software Engineering Conference (pp. 1-19). Springer, Berlin, Heidelberg.
- Bofeng, Z., Na, W., Gengfeng, W. and Sheng, L., 2004, June. Research on a personalized expert system explanation method based on fuzzy user model. In Fifth World Congress on Intelligent Control and Automation (IEEE Cat. No. 04EX788) (Vol. 5, pp. 3996-4000). IEEE.
- Bofeng, Z. and Yue, L., 2005, November. Customized explanation in expert system for earthquake prediction. In 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05) (pp. 5-pp). IEEE.
- Bostandjiev, S., O'Donovan, J. and Höllerer, T., 2012, September. TasteWeights: a visual interactive hybrid recommender system. In Proceedings of the sixth ACM conference on Recommender systems (pp. 35-42).
- Bradley, S., 2013. Design fundamentals: Elements, attributes, & principles. Colorado: Vansco design.
- Brank, J., Grobelnik, M. and Mladenic, D., 2005, October. A survey of ontology evaluation techniques. In Proceedings of the conference on data mining and data warehouses (SiKDD 2005) (pp. 166-170). Citeseer Ljubljana, Slovenia.
- Braun, V. and Clarke, V., 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), pp.77-101.
- Braunhofer, M., Elahi, M. and Ricci, F., 2014, September. Usability assessment of a context-aware and personality-based mobile recommender system. In International conference on electronic commerce and web technologies (pp. 77-88). Springer, Cham.
- Brehm, S.S. and Brehm, J.W., 2013. Psychological reactance: A theory of freedom and control. Academic Press.
- Buçinca, Z., Malaya, M.B. and Gajos, K.Z., 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), pp.1-21.
- Bunt, A., Lount, M. and Lauzon, C., 2012, February. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (pp. 169-178).
- Burrell, G. and Morgan, G., 2017. Sociological paradigms and organisational analysis: Elements of the sociology of corporate life. Routledge.
- Buskermolen, D.O. and Terken, J., 2012, August. Co-constructing stories: a participatory design technique to elicit in-depth user feedback and suggestions about design concepts. In Proceedings of the 12th Participatory Design Conference: Exploratory Papers, Workshop Descriptions, Industry Cases-Volume 2 (pp. 33-36).
- Bussone, A., Stumpf, S. and O'Sullivan, D., 2015, October. The role of explanations on trust and reliance in clinical decision support systems. In 2015 international conference on healthcare informatics (pp. 160-169). IEEE.
- Cai, C.J., Jongejan, J. and Holbrook, J., 2019, March. The effects of example-based explanations in a machine learning interface. In Proceedings of the 24th international conference on intelligent user interfaces (pp. 258-262).

- Cai, C.J., Winter, S., Steiner, D., Wilcox, L. and Terry, M., 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW), pp.1-24.
- Campbell, C., Mattison Thompson, F., Grimm, P.E. and Robson, K., 2017. Understanding why consumers don't skip pre-roll video ads. *Journal of Advertising*, 46(3), pp.411-423.
- Caraban, A., Karapanos, E., Gonçalves, D. and Campos, P., 2019, May. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
- Carvalho, D.V., Pereira, E.M. and Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), p.832.
- Castelvecchi, D., 2016. Can we open the black box of AI?. *Nature News*, 538(7623), p.20.
- Chaffee, B.W. and Zimmerman, C.R., 2010. Developing and implementing clinical decision support for use in a computerized prescriber-order-entry system. *American Journal of Health-System Pharmacy*, 67(5), pp.391-400.
- Char, D.S., Shah, N.H. and Magnus, D., 2018. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11), p.981.
- Chavaillaz, A., Wastell, D. and Sauer, J., 2016. System reliability, performance and trust in adaptable automation. *Applied Ergonomics*, 52, pp.333-342.
- Chen, D., Fraiberger, S.P., Moakler, R. and Provost, F., 2017. Enhancing transparency and control when drawing data-driven inferences about individuals. *Big data*, 5(3), pp.197-212.
- Chen, L., 2009, October. Adaptive tradeoff explanations in conversational recommenders. In *Proceedings of the third ACM conference on Recommender systems* (pp. 225-228).
- Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z. and Zha, H., 2019, July. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 765-774).
- Cheshire, C., 2011. Online trust, trustworthiness, or assurance?. *Daedalus*, 140(4), pp.49-58.
- Chouldechova, A., 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2), pp.153-163.
- Chromik, M., Eiband, M., Buchner, F., Krüger, A. and Butz, A., 2021, April. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces* (pp. 307-317).
- Chromik, M., Eiband, M., Völkel, S.T. and Buschek, D., 2019, March. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. In *IUI workshops* (Vol. 2327).
- Chun Tie, Y., Birks, M. and Francis, K., 2019. Grounded theory research: A design framework for novice researchers. *SAGE open medicine*, 7, p.2050312118822927.
- Clement, A., McPhail, B., Smith, K.L. and Ferenbok, J., 2012, August. Probing, mocking and prototyping: participatory approaches to identity infrastructuring. In *Proceedings of the 12th Participatory Design Conference: Research Papers-Volume 1* (pp. 21-30).

- Cooper, R., 2002. Order and disorder in everyday action: the roles of contention scheduling and supervisory attention. *Neurocase*, 8(1), pp.61-79.
- Corbin, J. and Strauss, A., 2014. Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage publications.
- Cowan, D.D. and Lucena, C.J.P.D., 1995. Abstract data views: An interface specification concept to enhance design for reuse. *IEEE Transactions on software engineering*, 21(3), pp.229-243.
- Crandall, B., Klein, G.A. and Hoffman, R.R., 2006. Working minds: A practitioner's guide to cognitive task analysis. Mit Press.
- Creswell, J.W. and Creswell, J.D., 2017. Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications.
- Creswell, J.W. and Poth, C.N., 2016. Qualitative inquiry and research design: Choosing among five approaches. Sage publications.
- Curley, S.P., Young, M.J., Kingry, M.J. and Yates, J.F., 1988. Primacy effects in clinical judgments of contingency. *Medical Decision Making*, 8(3), pp.216-222.
- Dabkowski, P. and Gal, Y., 2017. Real time image saliency for black box classifiers. arXiv preprint arXiv:1705.07857.
- Dash, S., Günlük, O. and Wei, D., 2018. Boolean decision rules via column generation. arXiv preprint arXiv:1805.09901.
- Datta, A., Sen, S. and Zick, Y., 2016, May. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In 2016 IEEE symposium on security and privacy (SP) (pp. 598-617). IEEE.
- De Carolis, B., de Rosis, F., Grasso, F., Rossiello, A., Berry, D.C. and Gillie, T., 1996. Generating recipient-centered explanations about drug prescription. *Artificial Intelligence in Medicine*, 8(2), pp.123-145.
- De Koning, B.B., Tabbers, H.K., Rikers, R.M. and Paas, F., 2009. Towards a framework for attention cueing in instructional animations: Guidelines for research and design. *Educational Psychology Review*, 21(2), pp.113-140.
- de Visser, E.J., Cohen, M., Freedy, A. and Parasuraman, R., 2014, June. A design methodology for trust cue calibration in cognitive agents. In International conference on virtual, augmented and mixed reality (pp. 251-262). Springer, Cham.
- De Visser, E.J., Monfort, S.S., Goodyear, K., Lu, L., O'Hara, M., Lee, M.R., Parasuraman, R. and Krueger, F., 2017. A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. *Human factors*, 59(1), pp.116-133.
- Denzin, N.K. and Lincoln, Y.S., 2008. Introduction: The discipline and practice of qualitative research.
- Dhurandhar, A., Chen, P.Y., Luss, R., Tu, C.C., Ting, P., Shanmugam, K. and Das, P., 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. arXiv preprint arXiv:1802.07623.
- Díaz-Agudo, B., Recio-Garcia, J.A. and Jimenez-Díaz, G., 2018. Data explanation with CBR. *ICCBR 2018*, p.64.

- Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K. and Dugan, C., 2019, March. Explaining models: an empirical study of how explanations impact fairness judgment. In Proceedings of the 24th international conference on intelligent user interfaces (pp. 275-285).
- Doherty, K. and Doherty, G., 2018. Engagement in HCI: conception, theory and measurement. *ACM Computing Surveys (CSUR)*, 51(5), pp.1-39.
- Dong, Y., Su, H., Zhu, J. and Bao, F., 2017. Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv preprint arXiv:1708.05493*.
- Duff, B.R. and Faber, R.J., 2011. Missing the mark. *Journal of Advertising*, 40(2), pp.51-62.
- Ehrlich, K., Kirk, S.E., Patterson, J., Rasmussen, J.C., Ross, S.I. and Gruen, D.M., 2011, February. Taking advice from intelligent systems: the double-edged sword of explanations. In Proceedings of the 16th international conference on Intelligent user interfaces (pp. 125-134).
- Eiband, M., Buschek, D., Kremer, A. and Hussmann, H., 2019, May. The impact of placebic explanations on trust in intelligent systems. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-6).
- Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M. and Hussmann, H., 2018, March. Bringing transparency design into practice. In 23rd international conference on intelligent user interfaces (pp. 211-223).
- Eiband, M., Schneider, H. and Buschek, D., 2018. Normative vs. Pragmatic: Two Perspectives on the Design of Explanations in Intelligent Systems. In *IUI Workshops*.
- Eiband, M., Völkel, S.T., Buschek, D., Cook, S. and Hussmann, H., 2019, March. When people and algorithms meet: User-reported problems in intelligent everyday applications. In Proceedings of the 24th international conference on intelligent user interfaces (pp. 96-106).
- Ekman, F., Johansson, M. and Sochor, J., 2017. Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems*, 48(1), pp.95-101.
- Elahi, Mehdi, Mouzhi Ge, Francesco Ricci, Ignacio Fernández-Tobías, and Schlomo Berkovsky. "Interaction design in a mobile food recommender system." In *IntRS 2015 Interfaces and Human Decision Making for Recommender Systems: Proceedings of the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2015)*, vol. 1438, pp. 49-52. CEUR-WS, 2015.
- Elo, S. and Kyngäs, H., 2008. The qualitative content analysis process. *Journal of advanced nursing*, 62(1), pp.107-115.
- Eslami, M., Krishna Kumaran, S.R., Sandvig, C. and Karahalios, K., 2018, April. Communicating algorithmic process in online behavioral advertising. In Proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1-13).
- Evans, J.S.B., 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59, pp.255-278.
- Faulkner, S.L. and Trotter, S.P., 2017. Theoretical saturation. *The International Encyclopedia of Communication Research Methods*, pp.1-2.
- Fazio, R.H., Ledbetter, J.E. and Towles-Schwen, T., 2000. On the costs of accessible attitudes: Detecting that the attitude object has changed. *Journal of personality and social psychology*, 78(2), p.197.

- Feng, S. and Boyd-Graber, J., 2019, March. What can ai do for me? evaluating machine learning interpretations in cooperative play. In Proceedings of the 24th International Conference on Intelligent User Interfaces (pp. 229-239).
- Fernandez, A., Herrera, F., Cordon, O., del Jesus, M.J. and Marcelloni, F., 2019. Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?. *IEEE Computational intelligence magazine*, 14(1), pp.69-81.
- Fischer, P., Schulz-Hardt, S. and Frey, D., 2008. Selective exposure and information quantity: how different information quantities moderate decision makers' preference for consistent and inconsistent information. *Journal of personality and social psychology*, 94(2), p.231.
- Fong, R.C. and Vedaldi, A., 2017. Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE international conference on computer vision (pp. 3429-3437).
- Fonteyn, M.E., Kuipers, B. and Grobe, S.J., 1993. A description of think aloud method and protocol analysis. *Qualitative health research*, 3(4), pp.430-441.
- Ford, K.M., Cañas, A.J. and Coffey, J.W., 1993, April. Participatory explanation. In Proceedings of the sixth Florida artificial intelligence research symposium (pp. 111-115).
- Friedman, B., Kahn Jr, P.H. and Borning, A., 2020. Value sensitive design and information systems. In *The Ethics of Information Technologies* (pp. 289-313). Routledge.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp.1189-1232.
- Gal, Y. and Ghahramani, Z., 2016, June. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (pp. 1050-1059). PMLR.
- Galbraith, J., 1973. *Designing complex organizations*. Reading, Mass.
- Garcia-Lopez, E., de-Marcos, L., Garcia-Cabot, A. and Martinez-Herraiz, J.J., 2015. Comparing zooming methods in mobile devices: effectiveness, efficiency, and user satisfaction in touch and nontouch smartphones. *International Journal of Human-Computer Interaction*, 31(11), pp.777-789.
- Gasparic, M., Janes, A., Ricci, F. and Zanellati, M., 2017, March. GUI design for IDE command recommendations. In Proceedings of the 22nd International Conference on Intelligent User Interfaces (pp. 595-599).
- Gedikli, F., Jannach, D. and Ge, M., 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4), pp.367-382.
- Gill, J. and Johnson, P., 2002. *Research methods for managers*. Sage.
- Gkika, S. and Lekakos, G., 2014, May. The Persuasive Role of Explanations in Recommender Systems. In *BCSS@ PERSUASIVE* (pp. 59-68).
- Glasnapp, J. and Brdiczka, O., 2009, July. A human-centered model for detecting technology engagement. In *International Conference on Human-Computer Interaction* (pp. 621-630). Springer, Berlin, Heidelberg.

- Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E., 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), pp.44-65.
- Gönül, M.S., Önköl, D. and Lawrence, M., 2006. The effects of structural characteristics of explanations on use of a DSS. *Decision support systems*, 42(3), pp.1481-1493.
- Goodfellow, I.J., Shlens, J. and Szegedy, C., 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Goodman, B. and Flaxman, S., 2017. European Union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, 38(3), pp.50-57.
- Graves, A., 2011. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24.
- Gregor, S. and Benbasat, I., 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, pp.497-530.
- Grewal, D. and Monroe, K.B., 1995. Information cues as signals of quality. *ACR European Advances*.
- Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F. and Giannotti, F., 2018. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D., 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), pp.1-42.
- Dietvorst, B.J., Simmons, J.P. and Massey, C., 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), p.114.
- Hadwin, A.F. and Winne, P.H., 2001. CoNoteS2: A software tool for promoting self-regulation. *Educational Research and Evaluation*, 7(2-3), pp.313-334.
- Hall, P. and Gill, N., 2019. *An introduction to machine learning interpretability*. O'Reilly Media, Incorporated.
- Hardiman, P.T., Dufresne, R. and Mestre, J.P., 1989. The relation between problem categorization and problem solving among experts and novices. *Memory & cognition*, 17(5), pp.627-638.
- Harding, R., 1998. *Environmental decision making*. NSW, Australia: The Federation Press.
- Helldin, T., Falkman, G., Riveiro, M. and Davidsson, S., 2013, October. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications* (pp. 210-217).
- Hendrycks, D. and Gimpel, K., 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.
- Henelius, A., Puolamäki, K., Boström, H., Asker, L. and Papapetrou, P., 2014. A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery*, 28(5), pp.1503-1529.

- Henelius, A., Puolamäki, K. and Ukkonen, A., 2017. Interpreting classifiers through attribute interactions in datasets. arXiv preprint arXiv:1707.07576.
- Hergeth, S., Lorenz, L., Vilimek, R. and Krems, J.F., 2016. Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human factors*, 58(3), pp.509-519.
- Herlocker, J.L., Konstan, J.A. and Riedl, J., 2000, December. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241-250).
- Hoffman, R.R., 2017. A taxonomy of emergent trusting in the human-machine relationship. *Cognitive Systems Engineering*, pp.137-164.
- Hoffman, R.R., Coffey, J.W., Ford, K.M. and Carnot, M.J., 2001, October. Storm-lk: A human-centered knowledge model for weather forecasting. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 1, pp. 752-752).
- Hoffman, R.R., Mueller, S.T., Klein, G. and Litman, J., 2018. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
- Holliday, D., Wilson, S. and Stumpf, S., 2013. The effect of explanations on perceived control and behaviors in intelligent systems. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (pp. 181-186).
- Hooker, G., 2004, August. Diagnosing extrapolation: Tree-based density estimation. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 569-574).
- Howells, J., 1996. Tacit knowledge. *Technology analysis & strategic management*, 8(2), pp.91-106.
- Huang, S.H., Bhatia, K., Abbeel, P. and Dragan, A.D., 2018, October. Establishing appropriate trust via critical states. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3929-3936). IEEE.
- Ishii, R. and Nakano, Y.I., 2010, February. An empirical study of eye-gaze behaviors: Towards the estimation of conversational engagement in human-agent communication. In *Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction* (pp. 33-40).
- Israelsen, B.W. and Ahmed, N.R., 2019. "Dave... I can assure you... that it's going to be all right..." A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Computing Surveys (CSUR)*, 51(6), pp.1-37.
- Johansson, U., König, R. and Niklasson, L., 2004. The Truth is In There-Rule Extraction from Opaque Models Using Genetic Programming. In *FLAIRS Conference* (pp. 658-663).
- Johansson, U. and Niklasson, L., 2009, March. Evolving decision trees using oracle guides. In *2009 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 238-244). IEEE.
- John, O.P. and Srivastava, S., 1999. *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives* (Vol. 2, pp. 102-138). Berkeley: University of California.
- Johnson, C.M., Johnson, T.R. and Zhang, J., 2005. A user-centered framework for redesigning health care interfaces. *Journal of biomedical informatics*, 38(1), pp.75-87.
- Josse, J., Prost, N., Scornet, E. and Varoquaux, G., 2019. On the consistency of supervised learning with missing values. arXiv preprint arXiv:1902.06931.

- Jugovac, M., Nunes, I. and Jannach, D., 2018, July. Investigating the decision-making behavior of maximizers and satisficers in the presence of recommendations. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (pp. 279-283).
- Kahneman, D., 2011. *Thinking, fast and slow*. Macmillan.
- Kanehira, A. and Harada, T., 2019. Learning to explain with complemental examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8603-8611).
- Kang, B., Tintarev, N., Höllerer, T. and O'Donovan, J., 2016, November. What am I not seeing? An interactive approach to social content discovery in microblogs. In *International Conference on Social Informatics* (pp. 279-294). Springer, Cham.
- Kang, M., Choo, P. and Watters, C.E., 2015. Design for experiencing: participatory design approach with multidisciplinary perspectives. *Procedia-Social and Behavioral Sciences*, 174, pp.830-833.
- Kaptein, F., Broekens, J., Hindriks, K. and Neerincx, M., 2017, August. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 676-682). IEEE.
- Karga, S. and Satratzemi, M., 2019. Using explanations for recommender systems in learning design settings to enhance teachers' acceptance and perceived experience. *Education and Information Technologies*, 24(5), pp.2953-2974.
- Keil, F.C., 2006. Explanation and understanding. *Annu. Rev. Psychol.*, 57, pp.227-254.
- Kessler, T., Stowers, K., Brill, J.C. and Hancock, P.A., 2017, September. Comparisons of human-human trust with other forms of human-technology trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1303-1307). Sage CA: Los Angeles, CA: SAGE Publications.
- Kim, B., Khanna, R. and Koyejo, O.O., 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.
- Kim, B., Rudin, C. and Shah, J.A., 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in neural information processing systems* (pp. 1952-1960).
- Kitchenham, B., Linkman, S. and Law, D., 1997. DESMET: a methodology for evaluating software engineering methods and tools. *Computing & Control Engineering Journal*, 8(3), pp.120-126.
- Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., El Emam, K. and Rosenberg, J., 2002. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on software engineering*, 28(8), pp.721-734.
- Klein, G. and Borders, J., 2016. The ShadowBox approach to cognitive skills training: An empirical evaluation. *Journal of Cognitive Engineering and Decision Making*, 10(3), pp.268-280.
- Kleinerman, A., Rosenfeld, A. and Kraus, S., 2018, September. Providing explanations for recommendations in reciprocal environments. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 22-30).

- Knijnenburg, B.P. and Kobsa, A., 2013. Making decisions about privacy: information disclosure in context-aware recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiIS)*, 3(3), pp.1-23.
- Koh, P.W. and Liang, P., 2017, July. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning* (pp. 1885-1894). PMLR.
- Konig, R., Johansson, U. and Niklasson, L., 2008, December. G-REX: A versatile framework for evolutionary data mining. In *2008 IEEE International Conference on Data Mining Workshops* (pp. 971-974). IEEE.
- Kool, W. and Botvinick, M., 2018. Mental labour. *Nature human behaviour*, 2(12), pp.899-908.
- Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J. and Getoor, L., 2019, March. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 379-390).
- Krause, J., Perer, A. and Bertini, E., 2018. A user study on the effect of aggregating explanations for interpreting machine learning models. In *ACM KDD Workshop on Interactive Data Exploration and Analytics*.
- Krause, J., Perer, A. and Ng, K., 2016, May. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5686-5697).
- Krebs, L.M., Alvarado Rodriguez, O.L., Dewitte, P., Ausloos, J., Geerts, D., Naudts, L. and Verbert, K., 2019, May. Tell me what you know: GDPR implications on designing transparency and accountability for news recommender systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-6).
- Krippendorff, K., 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Krishnan, R., Sivakumar, G. and Bhattacharya, P., 1999. Extracting decision trees from trained neural networks. *Pattern recognition*, 32(12).
- Krishnan, S. and Wu, E., 2017, May. Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics* (pp. 1-6).
- Krueger, R.A., 2014. *Focus groups: A practical guide for applied research*. Sage publications.
- Kujala, S. and Väänänen-Vainio-Mattila, K., 2009. Value of information systems and products: Understanding the users' perspective and values. *Journal of Information Technology Theory and Application (JITTA)*, 9(4), p.4.
- Kulesza, T., Burnett, M., Wong, W.K. and Stumpf, S., 2015, March. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 126-137).
- Kulesza, T., Stumpf, S., Burnett, M. and Kwan, I., 2012, May. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1-10).
- Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I. and Wong, W.K., 2013, September. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing* (pp. 3-10). IEEE.

- Kunkel, J., Donkers, T., Michael, L., Barbu, C.M. and Ziegler, J., 2019, May. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).
- Lai, V. and Tan, C., 2019, January. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 29-38).
- Lakkaraju, H., Kamar, E., Caruana, R. and Leskovec, J., 2017. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.
- Lamche, B., Adigüzel, U. and Wörndl, W., 2014, September. Interactive explanations in mobile shopping recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems* (Vol. 14).
- Langley, P., Meadows, B., Sridharan, M. and Choi, D., 2017, February. Explainable agency for intelligent autonomous systems. In *Twenty-Ninth IAAI Conference*.
- Laugel, T., Lesot, M.J., Marsala, C., Renard, X. and Detynecki, M., 2017. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*.
- Lazar, J., Feng, J.H. and Hochheiser, H., 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.
- Lazar, J., Feng, J.H. and Hochheiser, H., 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.
- LEE, J. D. & SEE, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46, 50-80.
- Leveson, N.G., 2016. *Engineering a safer world: Systems thinking applied to safety* (p. 560). The MIT Press.
- Lewis, J.D. and Weigert, A., 1985. Trust as a social reality. *Social forces*, 63(4), pp.967-985.
- Liao, Q.V., Gruen, D. and Miller, S., 2020, April. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).
- Lim, B.Y., 2012. *Improving understanding and trust with intelligibility in context-aware applications* (Doctoral dissertation, Carnegie Mellon University).
- Lim, B.Y. and Dey, A.K., 2009, September. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing* (pp. 195-204).
- Lim, B.Y. and Dey, A.K., 2010, September. Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 13-22).
- Lim, B.Y., Dey, A.K. and Avrahami, D., 2009, April. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2119-2128).
- Lim, B.Y., Yang, Q., Abdul, A.M. and Wang, D., 2019. Why these Explanations? Selecting Intelligibility Types for Explanation Goals. In *IUI Workshops*.
- Lords, H.O., 2018. AI in the UK: ready, willing and able?.

- Lou, Y., Caruana, R., Gehrke, J. and Hooker, G., 2013, August. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 623-631).
- Lucassen, G., Dalpiaz, F., van der Werf, J.M.E. and Brinkkemper, S., 2016, March. The use and effectiveness of user stories in practice. In International working conference on requirements engineering: Foundation for software quality (pp. 205-222). Springer, Cham.
- Lundberg, S.M. and Lee, S.I., 2017, December. A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).
- Madhavan, P. and Wiegmann, D.A., 2007. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4), pp.277-301.
- Madsen, M. and Gregor, S., 2000, December. Measuring human-computer trust. In 11th australasian conference on information systems (Vol. 53, pp. 6-8). Brisbane, Australia: Australasian Association for Information Systems.
- Marcus, A. and Wang, W. eds., 2017. Design, User Experience, and Usability: Designing Pleasurable Experiences: 6th International Conference, DUXU 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part II (Vol. 10289). Springer.
- Martens, D. and Provost, F., 2014. Explaining data-driven document classifications. *MIS quarterly*, 38(1), pp.73-100.
- McAllister, D.J., 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal*, 38(1), pp.24-59.
- McCall, J.C. and Trivedi, M.M., 2007. Driver behavior and situation aware brake assistance for intelligent vehicles. *Proceedings of the IEEE*, 95(2), pp.374-387.
- Mejtoft, T., Hale, S. and Söderström, U., 2019, September. Design Friction. In Proceedings of the 31st European Conference on Cognitive Ergonomics (pp. 41-44).
- Millecamp, M., Htun, N.N., Conati, C. and Verbert, K., 2019, March. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In Proceedings of the 24th International Conference on Intelligent User Interfaces (pp. 397-407).
- Miller, C.A. and Parasuraman, R., 2007. Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human factors*, 49(1), pp.57-75.
- Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, pp.1-38.
- Miller, T., Howe, P. and 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. *arXiv preprint arXiv:1712.00547*.
- Milliez, G., Lallement, R., Fiore, M. and Alami, R., 2016, March. Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 43-50). IEEE.
- Mills, S., 2000. The importance of task analysis in Usability Context Analysis-designing for fitness for purpose. *Behaviour & Information Technology*, 19(1), pp.57-68.

- Mishra, S., Sturm, B.L. and Dixon, S., 2017, October. Local Interpretable Model-Agnostic Explanations for Music Content Analysis. In ISMIR (pp. 537-543).
- Montavon, G., Samek, W. and Müller, K.R., 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, pp.1-15.
- Mothilal, R.K., Sharma, A. and Tan, C., 2020, January. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607-617).
- Muhammad, K., Lawlor, A., Rafter, R. and Smyth, B., 2015, September. Great explanations: Opinionated explanations for recommendations. In *International Conference on Case-Based Reasoning* (pp. 244-258). Springer, Cham.
- Muir, B.M., 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11), pp.1905-1922.
- Zhou, L., Dai, L. and Zhang, D., 2007. Online shopping acceptance model-A critical survey of consumer factors in online shopping. *Journal of Electronic commerce research*, 8(1), p.41.
- Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N. and Ali, R., 2021. Explainable recommendations and calibrated trust: two systematic user errors. *Computer*, 54(10), pp.28-37.
- Naiseh, M., Jiang, N., Ma, J. and Ali, R., 2020, April. Personalising explainable recommendations: literature and conceptualisation. In *World Conference on Information Systems and Technologies* (pp. 518-533). Springer, Cham.
- Naiseh, M., Jiang, N., Ma, J. and Ali, R., 2020, September. Explainable recommendations in intelligent systems: Delivery methods, modalities and risks. In *International Conference on Research Challenges in Information Science* (pp. 212-228). Springer, Cham.
- Naiseh, M., Al-Thani, D., Jiang, N. and Ali, R., 2021. Explainable recommendation: when design meets trust calibration. *World Wide Web*, 24(5), pp.1857-1884.
- Naiseh, M., 2020, September. Explainability design patterns in clinical decision support systems. In *International Conference on Research Challenges in Information Science* (pp. 613-620). Springer, Cham.
- Naiseh, M., 2021. Explainable recommendation: When design meets trust calibration—Research protocol.
- Näkki, P. and Antikainen, M., 2008. Online tools for co-design: User involvement through the innovation process. *New Approaches to Requirements Elicitation*, 96.
- Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S. and Doshi-Velez, F., 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*.
- Nathan, L.P., Klasnja, P.V. and Friedman, B., 2007, April. Value scenarios: a technique for envisioning systemic effects of new technologies. In *CHI'07 extended abstracts on Human factors in computing systems* (pp. 2585-2590).
- Naveed, S., Donkers, T. and Ziegler, J., 2018, July. Argumentation-based explanations in recommender systems: conceptual framework and empirical results. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (pp. 293-298).

- Nguyen, A., Yosinski, J. and Clune, J., 2016. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. arXiv preprint arXiv:1602.03616.
- Nielsen, J., 1994. Estimating the number of subjects needed for a thinking aloud test. *International journal of human-computer studies*, 41(3), pp.385-397.
- Nunes, I. and Jannach, D., 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3), pp.393-444.
- O. Nyumba, T., Wilson, K., Derrick, C.J. and Mukherjee, N., 2018. The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and evolution*, 9(1), pp.20-32.
- O'brien, H.L. and Toms, E.G., 2013. Examining the generalizability of the User Engagement Scale (UES) in exploratory search. *Information Processing & Management*, 49(5), pp.1092-1107.
- O'Sullivan, D., Fraccaro, P., Carson, E. and Weller, P., 2014. Decision time for clinical decision support systems. *Clinical medicine*, 14(4), p.338.
- Oinas-Kukkonen, H. and Harjumaa, M., 2009. Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, 24(1), p.28.
- Okamura, K. and Yamada, S., 2018, September. Adaptive trust calibration for supervised autonomous vehicles. In *Adjunct proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications* (pp. 92-97).
- Oppenheimer, D.M., 2004. Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science*, 15(2), pp.100-105.
- Orb, A., Eisenhauer, L. and Wynaden, D., 2001. Ethics in qualitative research. *Journal of nursing scholarship*, 33(1), pp.93-96.
- Ortiz, D. and Greene, J., 2007. Research design: qualitative, quantitative, and mixed methods approaches. *Qualitative Research Journal*, 6(2), pp.205-208.
- Ouellette, J.A. and Wood, W., 1998. Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological bulletin*, 124(1), p.54.
- Parasuraman, R. and Manzey, D.H., 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3), pp.381-410.
- Parasuraman, R. and Riley, V., 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2), pp.230-253.
- Parayitam, S. and Dooley, R.S., 2009. The interplay between cognitive-and affective conflict and cognition-and affect-based trust in influencing decision outcomes. *Journal of Business Research*, 62(8), pp.789-796.
- Paterno, M.D., Maviglia, S.M., Gorman, P.N., Seger, D.L., Yoshida, E., Seger, A.C., Bates, D.W. and Gandhi, T.K., 2009. Tiering drug-drug interaction alerts by severity increases compliance rates. *Journal of the American Medical Informatics Association*, 16(1), pp.40-46.
- Payrovnaziri, S.N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J.H., Liu, X. and He, Z., 2020. Explainable artificial intelligence models using real-world electronic health record

- data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7), pp.1173-1185.
- Peirce, C.S., 1997. *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism*. SUNY Press.
- Petty, R.E. and Cacioppo, J.T., 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion* (pp. 1-24). Springer, New York, NY.
- Petty, R.E., Cacioppo, J.T. and Schumann, D., 1983. Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of consumer research*, 10(2), pp.135-146.
- Plaue, C.M., Miller, T. and Stasko, J.T., 2004. *Is a picture worth a thousand words? An evaluation of information awareness displays*. Georgia Institute of Technology.
- Poole, E.S., Le Dantec, C.A., Eagan, J.R. and Edwards, W.K., 2008, September. Reflecting on the invisible: understanding end-user perceptions of ubiquitous computing. In *Proceedings of the 10th international Conference on Ubiquitous Computing* (pp. 192-201). POPE, C., ZIEBLAND, S. & MAYS, N. 2000. Analysing qualitative data. *Bmj*, 320, 114-116.
- Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Wortman Vaughan, J.W. and Wallach, H., 2021, May. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-52).
- Quinlan, J.R., 1987, August. Generating production rules from decision trees. In *ijcai* (Vol. 87, pp. 304-307).
- Ramachandran, D., Fenty, M., Provine, R., Yeh, P., Jarrold, W., Ratnaparkhi, A. and Douglas, B., 2015, September. A TV program discovery dialog system using recommendations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 435-437).
- Ras, G., van Gerven, M. and Haselager, P., 2018. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and interpretable models in computer vision and machine learning* (pp. 19-36). Springer, Cham.
- Rasmussen, J., 1983. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE transactions on systems, man, and cybernetics*, (3), pp.257-266.
- Renkl, A., 2014. Toward an instructionally oriented theory of example- based learning. *Cognitive science*, 38(1), pp.1-37.
- Renkl, A., Hilbert, T. and Schworm, S., 2009. Example-based learning in heuristic domains: A cognitive load theory account. *Educational Psychology Review*, 21(1), pp.67-78.
- Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. Nothing else matters: model-agnostic explanations by identifying prediction invariance. *arXiv preprint arXiv:1611.05817*.
- Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
- Ribeiro, M.T., Singh, S. and Guestrin, C., 2018, April. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

- Ritchie, J., Lewis, J., Nicholls, C.M. and Ormston, R. eds., 2013. *Qualitative research practice: A guide for social science students and researchers*. sage.
- Robinette, P., Li, W., Allen, R., Howard, A.M. and Wagner, A.R., 2016, March. Overtrust of robots in emergency evacuation scenarios. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 101-108). IEEE.
- Rosenfeld, A. and Richardson, A., 2019. Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6), pp.673-705.
- Russell, S. and Norvig, P., 2002. *Artificial intelligence: a modern approach*.
- Sanders, E.B.N., 2002. From user-centered to participatory design approaches. In *Design and the social sciences* (pp. 18-25). CRC Press.
- Sanders, E.B.N., 2002. From user-centered to participatory design approaches. In *Design and the social sciences* (pp. 18-25). CRC Press.
- Sanneman, L. and Shah, J.A., 2020, May. A Situation Awareness-Based Framework for Design and Evaluation of Explainable AI. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (pp. 94-110). Springer, Cham.
- Sato, M., Ahsan, B., Nagatani, K., Sonoda, T., Zhang, Q. and Ohkuma, T., 2018, March. Explaining recommendations using contexts. In 23rd international conference on intelligent user interfaces (pp. 659-664).
- Sato, M., Ahsan, B., Nagatani, K., Sonoda, T., Zhang, Q. and Ohkuma, T., 2018, March. Explaining recommendations using contexts. In 23rd international conference on intelligent user interfaces (pp. 659-664).
- Saunders, M.N., Lewis, P., Thornhill, A. and Bristow, A., 2015. *Understanding research philosophy and approaches to theory development*.
- Schäfer, H., Hors-Fraile, S., Karumur, R.P., Calero Valdez, A., Said, A., Torkamaan, H., Ulmer, T. and Trattner, C., 2017, July. Towards health (aware) recommender systems. In *Proceedings of the 2017 international conference on digital health* (pp. 157-161).
- Schaffer, J., Giridhar, P., Jones, D., Höllerer, T., Abdelzaher, T. and O'donovan, J., 2015, March. Getting the message? A study of explanation interfaces for microblog data analysis. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 345-356).
- Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A. and Höllerer, T., 2019, March. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 240-251).
- Schaffer, J., O'Donovan, J., Marusich, L., Yu, M., Gonzalez, C. and Höllerer, T., 2018. A study of dynamic information display and decision-making in abstract trust games. *International Journal of Human-Computer Studies*, 113, pp.1-14.
- Scheepers-Hoeks, A.M.J., Grouls, R.J., Neef, C., Ackerman, E.W. and Korsten, E.H., 2013. Physicians' responses to clinical decision support on an intensive care unit—comparison of four different alerting methods. *Artificial intelligence in medicine*, 59(1), pp.33-38.
- Schmell, R.W. and Umanath, N.S., 1988. An experimental evaluation of the impact of data display format on recall performance. *Communications of the ACM*, 31(5), pp.562-570.

- Schrills, T. and Franke, T., 2020, July. Color for Characters-Effects of Visual Explanations of AI on Trust and Observability. In *International Conference on Human-Computer Interaction* (pp. 121-135). Springer, Cham.
- Schulam, P. and Saria, S., 2019, April. Can you trust this prediction? Auditing pointwise reliability after learning. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 1022-1031). PMLR.
- Shilton, K., Koepfler, J.A. and Fleischmann, K.R., 2014, February. How to see values in social computing: methods for studying values dimensions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 426-435).
- Shneiderman, B., Plaisant, C., Cohen, M.S., Jacobs, S., Elmqvist, N. and Diakopoulos, N., 2016. *Designing the user interface: strategies for effective human-computer interaction*. Pearson.
- Short, C., Rebar, A., Plotnikoff, R. and Vandelanotte, C., 2015. Designing engaging online behaviour change interventions: a proposed model of user engagement.
- Simonyan, K., Vedaldi, A. and Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- SOARES, E. & ANGELOV, P. 2019. Fair-by-design explainable models for prediction of recidivism. *arXiv preprint arXiv:1910.02043*.
- Sokol, K. and Flach, P., 2020, January. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 56-67).
- Sokol, K. and Flach, P., 2020. One explanation does not fit all. *KI-Künstliche Intelligenz*, 34(2), pp.235-250.
- Soll, J.B., Milkman, K.L. and Payne, J.W., 2014. A user's guide to debiasing.
- Song, J.H. and Adams, C.R., 1993. Differentiation through customer involvement in production or delivery. *Journal of Consumer Marketing*.
- Spinuzzi, C., 2005. The methodology of participatory design. *Technical communication*, 52(2), pp.163-174.
- Springer, A., 2019. *Accurate, Fair, and Explainable: Building Human-Centered AI*. University of California, Santa Cruz.
- Springer, A. and Whittaker, S., 2019, March. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 107-120).
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), pp.1929-1958.
- Strack, F., 1992. "Order effects" in survey research: Activation and information functions of preceding questions. In *Context effects in social and psychological research* (pp. 23-34). Springer, New York, NY.
- Subbaswamy, A. and Saria, S., 2018. Counterfactual normalization: Proactively addressing dataset shift and improving reliability using causal mechanisms. *arXiv preprint arXiv:1808.03253*.

- Sun, H., 2013. A longitudinal study of herd behavior in the adoption and continued use of technology. *Mis Quarterly*, pp.1013-1041.
- Sutcliffe, A., 2003, September. Scenario-based requirements engineering. In *Proceedings. 11th IEEE International Requirements Engineering Conference*, 2003. (pp. 320-329). IEEE.
- Sutcliffe, A. and Carroll, J., 1998. Generalizing claims and reuse of HCI knowledge. In *People and computers XIII* (pp. 159-176). Springer, London.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R., 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- Tan, S., Caruana, R., Hooker, G. and Lou, Y., 2018, December. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 303-310).
- Te'eni, D. and Sani-Kuperberg, Z., 2005. Levels of abstraction in designs of human–computer interaction: The case of e-mail. *Computers in human behavior*, 21(5), pp.817-830.
- Teixeira, L., Ferreira, C. and Santos, B.S., 2012. User-centered requirements engineering in health information systems: A study in the hemophilia field. *Computer methods and programs in biomedicine*, 106(3), pp.160-174.
- Ter Hoeve, M., Heruer, M., Odijk, D., Schuth, A. and de Rijke, M., 2017, August. Do news consumers want explanations for personalized news rankings. In *FATREC Workshop on Responsible Recommendation Proceedings*.
- Saunders, M., Lewis, P. and Thornhill, A., 2009. *Research methods for business students*. Pearson education.
- Tintarev, N. and Masthoff, J., 2011. Designing and evaluating explanations for recommender systems. In *Recommender systems handbook* (pp. 479-510). Springer, Boston, MA.
- Tolomei, G., Silvestri, F., Haines, A. and Lalmas, M., 2017, August. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 465-474).
- Tomsett, R., Braines, D., Harborne, D., Preece, A. and Chakraborty, S., 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*.
- Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G. and Kaplan, L., 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, 1(4), p.100049.
- Torning, K. and Oinas-Kukkonen, H., 2009, April. Persuasive system design: state of the art and future directions. In *Proceedings of the 4th international conference on persuasive technology* (pp. 1-8).
- Torrey, C., Powers, A., Marge, M., Fussell, S.R. and Kiesler, S., 2006, March. Effects of adaptive robot dialogue on information exchange and social relations. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (pp. 126-133).
- Tsai, C.H. and Brusilovsky, P., 2017, July. Providing control and transparency in a social recommender system for academic conferences. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 313-317).

- Tsai, C.H. and Brusilovsky, P., 2019, March. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 391-396).
- Tubbs, R.M., Gaeth, G.J., Levin, I.P. and Van Osdol, L.A., 1993. Order effects in belief updating with consistent and inconsistent evidence. *Journal of Behavioral Decision Making*, 6(4), pp.257-269.
- Tucker, C. and Zhang, J., 2011. How does popularity information affect choices? A field experiment. *Management Science*, 57(5), pp.828-842.
- Tulabandhula, T. and Rudin, C., 2014. On combining machine learning with decision making. *Machine learning*, 97(1-2), pp.33-64.
- Twycross, A., 2004. Research design: Qualitative, quantitative and mixed methods approaches. *Nurse Researcher* (through 2013), 12(1), p.82.
- Jacko, J.A. ed., 2012. *Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications*. CRC press.
- Van Doorn, J., Lemon, K.N., Mittal, V., Nass, S., Pick, D., Pirner, P. and Verhoef, P.C., 2010. Customer engagement behavior: Theoretical foundations and research directions. *Journal of service research*, 13(3), pp.253-266.
- Verbert, K., Parra, D., Brusilovsky, P. and Duval, E., 2013, March. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 351-362).
- Vermeeren, A.P., Law, E.L.C., Roto, V., Obrist, M., Hoonhout, J. and Väänänen-Vainio-Mattila, K., 2010, October. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries* (pp. 521-530).
- Verplanken, B. and Wood, W., 2006. Interventions to break and create consumer habits. *Journal of Public Policy & Marketing*, 25(1), pp.90-103.
- Veryzer, R.W. and Borja de Mozota, B., 2005. The impact of user- oriented design on new product development: An examination of fundamental relationships. *Journal of product innovation management*, 22(2), pp.128-143.
- Wachter, S., Mittelstadt, B. and Russell, C., 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31, p.841.
- Wagner, A.R. and Robinette, P., 2021. An explanation is not an excuse: Trust calibration in an age of transparent robots. In *Trust in Human-Robot Interaction* (pp. 197-208). Academic Press.
- Wagner, A.R. and Robinette, P., 2021. An explanation is not an excuse: Trust calibration in an age of transparent robots. In *Trust in Human-Robot Interaction* (pp. 197-208). Academic Press.
- Wang, B., Ester, M., Bu, J. and Cai, D., 2014, June. Who also likes it? generating the most persuasive social explanations in recommender systems. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.
- Wang, D., Yang, Q., Abdul, A. and Lim, B.Y., 2019, May. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-15).

- Wang, N., Pynadath, D.V. and Hill, S.G., 2016, March. Trust calibration within a human-robot team: Comparing automatically generated explanations. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 109-116). IEEE.
- Wei, D., Dash, S., Gao, T. and Gunluk, O., 2019, May. Generalized linear rule models. In International Conference on Machine Learning (pp. 6687-6696). PMLR.
- Weld, D.S. and Bansal, G., 2019. The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6), pp.70-79.
- West, D.M., 2018. The future of work: Robots, AI, and automation. Brookings Institution Press.
- Westfall, L., 2009. Sampling methods. *The Certified Quality Engineer Handbook*.
- Wever, R., Van Kuijk, J. and Boks, C., 2008. User- centred design for sustainable behaviour. *International journal of sustainable engineering*, 1(1), pp.9-20.
- Whitefield, A., Wilson, F. and Dowell, J., 1991. A framework for human factors evaluation. *Behaviour & Information Technology*, 10(1), pp.65-79.
- Wickens, C.D., 1995. Designing for situation awareness and trust in automation. *IFAC Proceedings Volumes*, 28(23), pp.365-370.
- Wiebe, M., Geiskkovitch, D.Y. and Bunt, A., 2016, March. Exploring user attitudes towards different approaches to command recommendation in feature-rich software. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (pp. 43-47).
- Wilson, J., 2014. *Essentials of business research: A guide to doing your research project*. Sage.
- Wolcott, M.D. and Lobczowski, N.G., 2021. Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning*, 13(2), pp.181-188.
- Wood, W. and Rünger, D., 2016. Psychology of habit. *Annual review of psychology*, 67, pp.289-314.
- Wood, W., Tam, L. and Witt, M.G., 2005. Changing circumstances, disrupting habits. *Journal of personality and social psychology*, 88(6), p.918.
- Woods, D.D. and Roth, E.M., 1988. Cognitive systems engineering. In *Handbook of human-computer interaction* (pp. 3-43). North-Holland.
- Wrobel, G.M., Grotevant, H.D., Samek, D.R. and Korff, L.V., 2013. Adoptees' curiosity and information-seeking about birth parents in emerging adulthood: Context, motivation, and behavior. *International journal of behavioral development*, 37(5), pp.441-450.
- Wu, W., Su, Y., Chen, X., Zhao, S., King, I., Lyu, M.R. and Tai, Y.W., 2020. Towards global explanations of convolutional neural networks with concept attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8652-8661).
- Yang, F., Huang, Z., Scholtz, J. and Arendt, D.L., 2020, March. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 189-201).
- Yang, X.J., Unhelkar, V.V., Li, K. and Shah, J.A., 2017, March. Evaluating effects of user experience and system transparency on trust in automation. In 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 408-416). IEEE.

- Yang, X.J., Unhelkar, V.V., Li, K. and Shah, J.A., 2017, March. Evaluating effects of user experience and system transparency on trust in automation. In 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 408-416). IEEE.
- Yin, M., Wortman Vaughan, J. and Wallach, H., 2019, May. Understanding the effect of accuracy on trust in machine learning models. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-12).
- Yu, K., Berkovsky, S., Taib, R., Zhou, J. and Chen, F., 2019, March. Do I trust my machine teammate? An investigation from perception to decision. In Proceedings of the 24th International Conference on Intelligent User Interfaces (pp. 460-468).
- Yuan, X., He, P., Zhu, Q. and Li, X., 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9), pp.2805-2824.
- Zanker, M. and Schoberegger, M., 2014, September. An empirical study on the persuasiveness of fact-based explanations for recommender systems. In Joint Workshop on Interfaces and Human Decision Making in Recommender Systems (Vol. 1253, pp. 33-36).
- Zhang, X., Prybutok, V.R., Ryan, S. and Pavur, R., 2009. A model of the relationship among consumer trust, web design and user attributes. *Journal of Organizational and End User Computing (JOEUC)*, 21(2), pp.44-66.
- Zhang, X., Solar-Lezama, A. and Singh, R., 2018. Interpreting neural network judgments via minimal, stable, and symbolic corrections. *arXiv preprint arXiv:1802.07384*.
- Zhang, Y., Liao, Q.V. and Bellamy, R.K., 2020, January. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 295-305).
- Zhang, Y., Liao, Q.V. and Bellamy, R.K., 2020, January. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 295-305).
- Zhao, G., Fu, H., Song, R., Sakai, T., Chen, Z., Xie, X. and Qian, X., 2019. Personalized reason generation for explainable song recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(4), pp.1-21.
- Zhou, Y. and Hooker, G., 2016. Interpreting models via single tree approximation. *arXiv preprint arXiv:1610.09036*.
- Zhou, Z.H., Jiang, Y. and Chen, S.F., 2003. Extracting symbolic rules from trained neural network ensembles. *Ai Communications*, 16(1), pp.3-15.
- Zucker, L.G., 1986. Production of trust: Institutional sources of economic structure, 1840–1920. *Research in organizational behavior*.

12. APPENDIX

12.1 APPENDIX 1 - EXPLAINABLE MODELS TAXONOMY VALIDATION

Appendix 1.1 The invitation email:

Hi all,

My name is Mohammad Naiseh, a PhD candidate in the faculty of Science and Technology, Bournemouth University, Bournemouth, UK. I am conducting a card-sorting focus group to validate a taxonomy developed for explainable machine learning.

I hope you can join our session and provide your valuable insights and comments.

Date: Friday the 24th June Time: 10:00 AM to 11:AM

Location: PG13, Poole House, Talbot Campus.

Notes: all documents mentioned in table 2 will be provided in the session. We recommend having the following tasks before the study to ensure better discussions and fruitful collaboration.

- Read the participant sheet.
- Fill the demographic survey and resend it through email.

Best wishes,

Mohammad Naiseh

Appendix 1.2 Participant Information sheet:

Introduction

Given the multidisciplinary nature of explaining machine learning models to end-users, we build a taxonomy to conceptualise related concepts and factors. We have collected five main categorisations for explaining ML models and linked them to users' questions from the literature. The goal of this study is to validate and revise the link between the explainable model and the users' questions.

Procedure.

Participants will evaluate and modify predetermined concepts and strategies, along with their created ones, then refine them offline through a post-exercise questionnaire. The study will be conducted on Friday the 24th June, Time: 10:00 AM to 11:AM, Location: PG13, Poole House, Talbot Campus. Participants will be divided into two groups. Each one will have a leader while the researcher acts as a facilitator. The exercise set will be as shown in the table below.

Table 1: Study structure

Stage No.	Name	Description	Note	Est time.
1	Prepare	The researcher will brief the participants about the study goals and structure.	-----	10 mins
2	Evaluate	The participants will be provided with a copy of the taxonomy to individually evaluate and make notes in the notes form.	Notes might include missing users' questions or concepts, structuring issues and probably refining suggestions	15 mins
3	Map	Each group will be provided with the same set of taxonomy concepts cards to start mapping the available users' questions to the explainable model with an option to add questions. This is meant to be a group activity.	Participants will be informed to delete/add questions and re-map them as they think appropriate. Disagreements were expected to arise but resolved during the session.	15 mins
4	Discussion	Each group discussed the other groups' mapping findings and highlighted the agreements and recommendations for further corrections.	Each group was given 10 mins	20 mins

Table 2: The provided documents during the mapping session

Document No.	Documents	Description
1	Taxonomy structure	The first draft of the taxonomy that appeared from our literature review was provided in the email invitation.
2	Notes form	Each participant used this form to amend the taxonomy and make notes about the taxonomy structure.
3	Catalogue	The catalogue defines the relative concepts of the explainable machine learning models and their usage scenarios.
4	Cards	Taxonomy concepts cards with extra blank ones were provided during the session.

Contact for further information

If you have any queries about this research, please contact my PhD supervisor, Dr Nan Jiang, by email at mnaiseh@bournemouth.ac.uk or by post to:

Dr Mohammad Naiseh
Department of Computing and Informatics
Faculty of Science and Technology
Bournemouth University
BH12 5BB

Complaints

If you have any complaints about this project please contact Professor Tiantian Zhang, Deputy Dean for Research and Professional Practice of the Faculty of Science and Technology at Bournemouth University at the following address:

Professor Tiantian Zhang
Talbot Campus, Fern Barrow, Poole, BH12 5BB
E-mail: researchgovernance@bournemouth.ac.uk
Tel: 01202 965721

Thank you for taking the time to read this information sheet, and please do not hesitate to contact me if you have any queries.

Appendix 1.3 Consent Form

We are asking for your kind help in validating the taxonomy through a focus group session. You have volunteered to take part in this study to improve the organisation of the taxonomy.

To have a complete record of participants' comments, the discussion in phase 4 will be audio recorded. While your privacy of identification will be safeguarded, no sensitive information will be collected. We will use the recordings to develop a final version of the taxonomy. Your participation is voluntary, and your answers will remain confidential.

I have volunteered to participate in this focus group, and I permit for the collected data to be used for the purposes stated above.

Participants' Name: -----

Participants' Signature: -----

Date: -----

Appendix 1.4 Background knowledge questionnaire:

Gender:

- ☐ Male
- ☐ Female

Age Group:

- ☐ 20 – 30
- ☐ 30-40
- ☐ 40-50
- ☐ 50-60
- ☐ Above 60

Your background knowledge: -----

Your title/position: -----

Years of experience: -----

A number of publications in your field:

- ☐ 1-5
- ☐ 6-10
- ☐ 11-15
- ☐ Above 15

Please rate your familiarity with the provided concepts:

	Very poor	Poor	Average	Good	Very good
Machine Learning					
User experience					
Explainable AI					
Human-AI Interaction					
Transparency					
Decision making					
Human-Centred computing					
Other: -----					

Name:

Email:

Appendix 1.5 Taxonomy structure before the focus group evaluation:

Global explanations	Global feature importance	Ranking the data features.	Why does the system think so? What are the main reasons for that?
		Dependencies between data features	What is the relation between features A and B?
		Influence Function	What is the effect of feature A?
	Decision tree	How the system reaches to that conclusion? What if?	

	approximation	Why not? Why?	
	Rule extraction	AND-OR rules	How the system reaches to that conclusion? What if? Why not? Why?
		If-then rules	How the system reaches to that conclusion? What if? Why not? Why?
	Explain prediction	Local feature importance	Why the system generates this prediction? What is the effect of specific feature on local prediction?
Local rules and trees		Why the system generates this prediction? How the system reached to this prediction?	
Example-based	Prototype	With what example could this prediction have happened? What cases could result the same prediction? What else similar to this prediction?	
	Counterfactual example	What examples could a change A to this prediction result? What else (small changes)?	
	Influential example	What examples make this prediction to happen? What anomalies examples could affect the prediction? What are the abstract examples of such a prediction?	
Counterfactual	Feature Influence	When the prediction could change? What if feature A has the value B?	
	Counterfactual features	When the prediction could change? How to change the prediction? Why not?	
Confidence	How certain is the system of this prediction? How accurate the system is?		

Appendix 1.6 Notes form:

Please use this form so you can make notes regards the taxonomy.

Concept/category name	Note
General Notes: please use the below space to mention any other issues or concerns regards the taxonomy structure.	

Appendix 2.1. Rating exercise.

We aim to provide explanatory information that help the medical practitioners to calibrate their trust in Collaborative Human-AI decision-making tools. We consulted with two AI experts and one medical expert, presenting them the explainable interface and asked them for their expert opinion regarding the relevance of the explanations. We used these opinions as well as the results from our pilot study to refine the interface design. We presented ten individual patient scenarios to every participant. They have been initialised with fictional names and profiles to make it more realistic to our practitioners. Each scenario was accompanied with one different explanation class and was meant to be either correct recommendation or incorrect recommendation. We asked our participants to self-report their cognition-based trust components in each explanation class using 5 Likert Scale question. Examples of the mock-up interfaces are shown in Figure 1.

Q1: I will use the explanation because I can understand how the system behaves.

- 1: Completely disagree, would not even consider looking at the explanation as part of my decision-making process
- 2: Disagree, unlikely to consider the explanation
- 3: Neutral, may use the explanation but unlikely to rely on it in any way
- 4: Agree, likely to use explanation and consider as a part of my decision-making process
- 5: Completely agree, almost certain to use the explanation together with its details when considering further medical decisions

Q2: I can rely on the explanation in the task properly.

- 1: Completely disagree, would not even consider looking at the explanation as part of my decision-making process
- 2: Disagree, unlikely to consider the explanation
- 3: Neutral, may use the explanation but unlikely to rely on it in any way
- 4: Agree, likely to use explanation and consider as a part of my decision-making process
- 5: Completely agree, almost certain to use the explanation together with its details when considering further medical decisions

Q3: The explanation has sound knowledge and accurate about this recommendation.

- 1: Completely disagree, would not even consider looking at the explanation as part of my decision-making process
- 2: Disagree, unlikely to consider the explanation
- 3: Neutral, may use the explanation but unlikely to rely on it in any way
- 4: Agree, likely to use explanation and consider as a part of my decision-making process

5: Completely agree, almost certain to use the explanation together with its details when considering further medical decisions

Patient: Jack (Mr) Gender: Male, Born: 12-June-1989 (40Y), SAP Number: marke0007

Palliative Gemcitabine and Paclitaxel Frequency: 21 days

Prescription: Prescribed - Approved by AI tool. Certainty: 78%

Time	Drug	Dose	Administration	Frequency	Route	Duration
T+30mins	Chlorpheniramine	10 mg			IV Bolus	1 minutes
T+1hr	Metoclopramide	10 mg		Three times a day	Oral	3 days
T+0	Sodium Chloride 0.9%	330mg	volume dependent upon requirement		IV Flush	

(a)

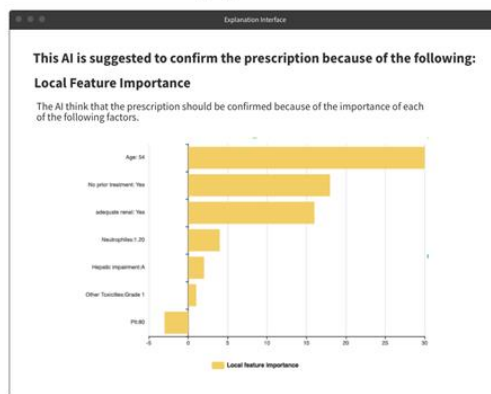
This AI is suggested to confirm the prescription because of the following:

Example-based

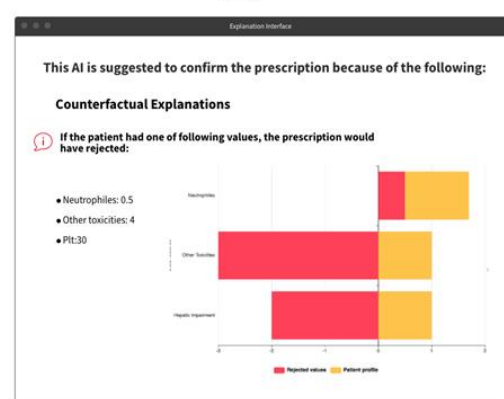
Top 5 similar patients to Jack who have previously had the same prescription in the past.

Age	No Prior tr.	ER status	Adequate renal	Neutrophils	Hepatic imp.	Other Tox.	PLT
55	YES	Positive	YES	1.22	A	Grade 1	80
54	YES	Positive	YES	1.2	B	Grade 1	75
54	YES	Positive	YES	1.22	A	Grade 2	80
56	YES	Positive	YES	1.24	A	Grade 1	76
58	YES	Positive	YES	1.18	A	Grade 2	78

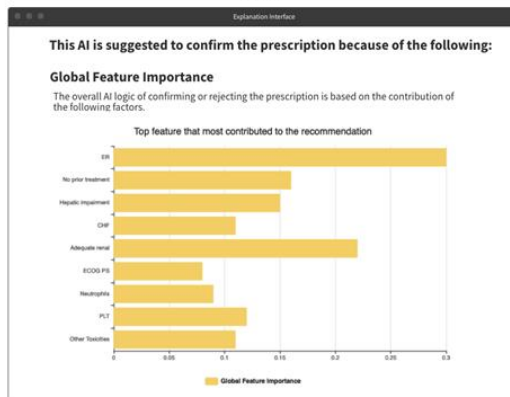
(b)



(c)



(d)



(e)

Fig 1. Four explanation classes mock-up interface presented to our participants. (a) No explanation. (b) Example-based explanation (c) Local feature importance (d) Counterfactual explanation (e) Global feature importance.

Appendix 2.2. follow-up interview questions.

1. How would you summarise why the AI-supported decision tool made the recommendations?
2. What do you think about this explanation and how do you evaluate it in helping you to understand the AI recommendation?
3. How do you assess it in helping you to rely on the AI recommendation?
4. How do you assess it in helping you to identify the correctness the AI recommendation?
5. Why might you be agreeing or disagreeing on [understandability, reliability and technical competence] of the recommendation and the explanation presented in this way?
6. What information led you to agree or disagree with the [understandability, reliability and technical competence] of the recommendation and its explanation?
7. What information is missing that might help you to assess the [understandability, reliability and technical competence] confidently or effectively?
8. What would you like to change about this explanation to help the assessment of the [understandability, reliability and technical competence]?

Appendix 2.3. follow-up interview questions.

Scenario Number	Explanation class	Type of recommendation
SC1	Confidence	Correct
SC2	Confidence	Incorrect
SC3	Counterfactual	Correct
SC4	Counterfactual	Incorrect
SC5	Global	Correct
SC6	Global	Incorrect
SC7	Local	Correct
SC8	Local	Incorrect
SC9	Example-based	Correct
SC10	Example-based	Incorrect

Table 35 Scenarios characteristics. Scenarios numbers do not represent the order of presentation.

Appendix 2.4. Sample of scenarios' characteristics.

	SC1	SC2	SC3	SC4
	Male:54 CHF	Male:47 CHF	Female:56 CHF	Male: 44 not CHF
ER	Positive	Positive	Positive	Negative
No prior treatment with CDK 4/6	Yes	Yes	Yes	Yes
Adequate renal and hepatic function	Yes	Yes	Yes	Yes
ECOG PS	2	0	1	2
Neutrophils	1.20	0.9	1.00	0.7
Plt	80	74	33	84
Hepatic impairment	A	B	A	C
Other Toxicities	Grade 1	Grade 2	Grade 1	Grade 4

Table 36 Four examples of four patients' profiles presented in the scenarios.

Appendix 2.5. Sample of a full scenarios characteristics.

74-year-old female

ER positive, HER2 negative invasive ductal carcinoma with lung and bone metastases

PMH/ DHX – nil and NKDA

Initiated on first line Palbociclib + letrozole and denosumab

Cycle 1 – Palbociclib 125mg od for days 1-21 of a 28-day cycle

Bloods taken at cycle 2 day 1 show Neut 0.6, therefore to defer 1 week

Bloods taken 1 week later show Neuts 1.0 therefore to proceed with treatment

Cycle 2 initiated at 125mg od for days 1-21 of 28-day cycle

Bloods taken at cycle 3 day 1 – neutropenic again, neut=0.6, therefore patient deferred again for 1 week or until blood count recovery NB this is second deferral for neutropenia so would also warrant a DR to 100mg od

Bloods taken 1 week later show Neuts 0.9 therefore defer 1 further week

Cycle 3 initiated at 100mg od for days 1-21 of 28-day cycle

CT scan post cycle 3 shows disease response with shrinking of metastases

Bloods taken at cycle 4 day 1 neut 1.1 – proceed with treatment at same dose (100mg od).

Cycle 4, 5 & 6 proceed with neut >1.0 at day 1 of each cycle.

CT scan post cycle 6 shows stable disease.

Cycles 7-9 prescribed based on blood counts from cycle 7 day 1 as patient has been stable for >3 months & now only required 3 monthly blood monitoring. Cycles 8 & 9 dispensed when due based still on bloods from cycle 7 day 1.

CT scan post cycle 6 shows stable disease.

Cycle 10 patient has bloods which are all adequate to proceed with treatment but is reporting significant fatigue and therefore decision is made to dose reduce to 75mg od.

Cycle 10 prescribed 75mg od days 1-21 of 28-day cycle

Patient reviewed at cycle 11 day 1 (bloods all adequate) – patient feel 75mg od much more bearable, energy levels have improved and will continue at this dose.

Cycles 11-12 prescribed. Cycle 12 dispensed when due based still on bloods f

Appendix 2.6 Screening survey

1. Please provide your age category.
 - ☐ 20-30
 - ☐ 30-40
 - ☐ 40-50
 - ☐ 50-60
2. Please provide your gender.
 - ☐ Male
 - ☐ Female
-
3. Approximately how long have you been practicing clinically?
 -
 - -----
 -
4. Please check all statements that apply regarding your level of experience screening chemotherapy prescriptions.
 - ☐ I know what screening prescription is.
 - ☐ I have used a screening prescription software in practice.

5. Please indicate your level of agreement with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Agree Strongly
Artificial Intelligence will play an important role in the future of medicine					
There are too many complexities and barriers in medicine for AI to help in clinical settings.					
I have reservations about using AI in clinical settings.					

12.3 APPENDIX 3. EXPLAINABLE RECOMMENDATION AND CALIBRATED TRUST: TWO SYSTEMATIC USERS' ERRORS

Calibrated trust has become an important design goal when designing Human-AI collaborative decision-making tools. It refers to a successful understandability, reliability and predictability to the AI-based tool behaviour and recommendations. Explainable AI is an emerging field where explanations accompany AI-based recommendations to help the human-decision maker understand, rely on, and predict AI behaviour. Such an approach is supposed to improve humans' trust calibration while working collaboratively with an AI. However, evidence from the literature suggests that explanations have not contributed to improved trust calibration and even introduced other errors. Designers of such explainable systems often assumed that humans would engage cognitively with AI-based explanations and use them in their Human-AI collaborative decision-making task. This research explores users' behaviour and interaction style with AI-based explanations during a Human-AI collaborative decision-making task. Such an investigation will help further studies address design solutions for AI explanations to enhance trust calibration and operationalize explainability during a Human-AI decision-making task. To achieve this goal, we conduct a multi-stage qualitative study. It includes think-aloud protocol, follow-up interviews and observations. The results of these studies will guide the research to develop an understanding of the main research question in the literature: "Why explanations do not improve trust calibration?". It will also help our future research to devise a design method for the XAI interface to enhance trust calibration. In the following subsection, we explained the procedures and provided the supplementary materials used in each study. The study workflow is summarised in Figure 1.

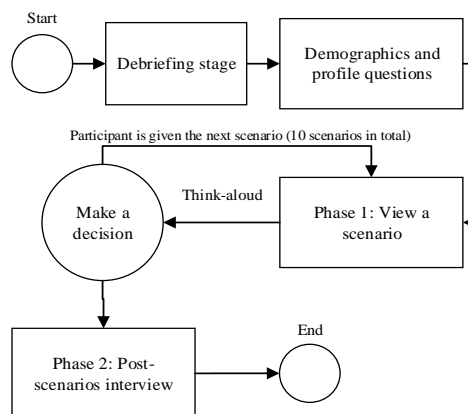


Figure 1 Study workflow for each participant

Phase 1: think-aloud protocol – first stage.

We aim to provide explanatory information that supports the medical practitioners in their trust calibration during Human-AI collaborative decision-making task. Our participant's inclusion criteria were based on their experience of using clinical decision support systems in their settings and experience in screening chemotherapy prescription. We designed ten recommendations accompanied by ten different explanations. The adopted recommendations were generated to be non-trivial, which was based on a literature review on related work and medical expert judgment. We tested the 2 material and activities with two participants and refined them to optimise their fulfilment of these criteria (See Appendix B). Also, we validated the material with a medical oncologist with a focus on the border cases that need an investigation from the participants in the actual study. This ultimately helped put our participants, who were medical experts, in a realistic setting: exposing them to an imperfect AI-based recommendation and its explanations where trust calibration is needed and where errors in that process are possible. We consulted with one AI expert and one medical expert,

presenting them with ten explainable interfaces, and asked them for their expert opinion regarding the relevance of the explanations and the validity of the recommendation. We used these opinions, as well as the results from our pilot study, to refine the interface design. Each scenario considered a hypothetical patient profile and AI-recommendations that suggests either rejecting or accepting a chemotherapy prescription for the patient. Patients have been initialised with fictional names and profiles to make it more realistic to our practitioners. Each scenario was accompanied by one different explanation class and was meant to be either correct recommendation or incorrect recommendation. We used our five main explanation classes revealed from our previous literature review. We encouraged them to think aloud during their decision-making process. Then, they were asked to think freely and encouraged to make optimal decisions. Examples of explainable interfaces used in our study settings are shown in Figure 2.

Patient: Jack (Mr)
Gender: Male, Born: 11-June-1988 (49Y), SAP Number: mskmrk4007

Palliative Gemtuzabine and Plicicavil Frequency: 21 days

Height: 185 Weight: 75 Surface Area: 1.81 Performance Status: 0

☒ Add drug(s) ☒ Re-schedule ☒ Cancel appointment ☒ Add cycle(s) ☒ Add course ☐ Remember Cycle ☐ View Treatment Summary

Prescribed - Approved by AI tool. Certainty: 78%

Prescription Pathology Results Toxicities Notes (1) Documents (1) Allergies Diagnosis History

Time	Drug	Dose	Administration	Frequency	Route	Duration
Box 1 (ASO) 2005 Chemotherapy Day 1						
T+30mins	Chlorpheniramine	10 mg			IV Bolus	1 minutes
T+1hr	Metoclopramide	10 mg		Three times a day	Oral	3 days
T+0	Sodium Chloride 0.9%	250mg	volume dependent upon requirement	IV Flush		

Confirm Prescription Cancel authorisation (return to prescription) View PDF prescription Day 1 Day 2

(a)

Explanation Interface

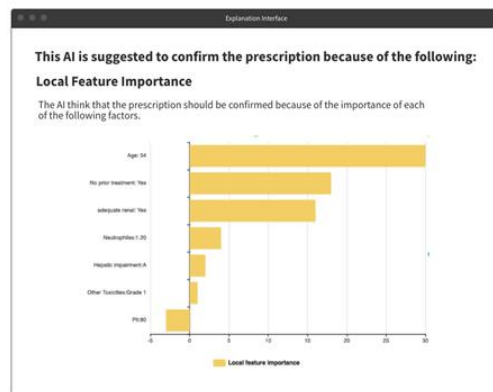
This AI is suggested to confirm the prescription because of the following:

Example-based

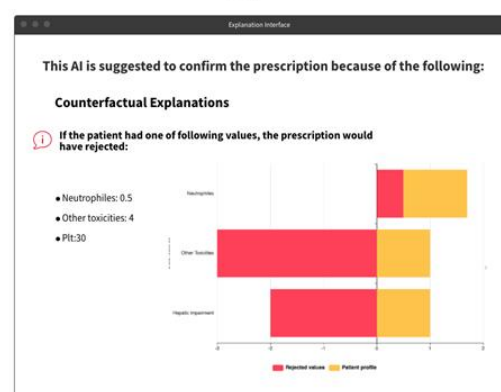
Top 5 similar patients to Jack who have previously had the same prescription in the past.

Age	No Prior tr...	ER status	adequate renal	Neutrophils	Hepatic imp...	Other Tox...	PLT
55	YES	Positive	YES	1.22	A	Grade 1	80
54	YES	Positive	YES	1.2	B	Grade 1	75
54	YES	Positive	YES	1.22	A	Grade 2	80
56	YES	Positive	YES	1.24	A	Grade 1	76
58	YES	Positive	YES	1.18	A	Grade 2	78

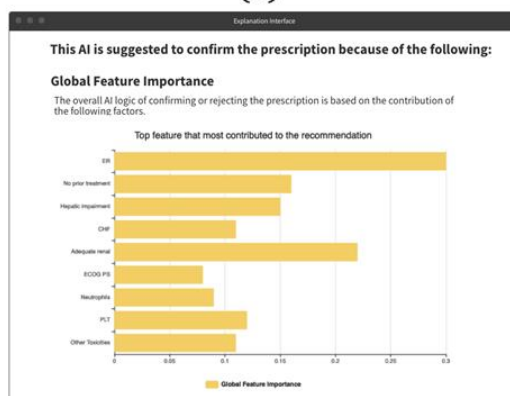
(b)



(c)



(d)



(e)

Fig 2. Five explanation classes mock-up interface presented to our participants. (a) confidence explanation. (b) Example-based explanation (c) Local feature importance (d) Counterfactual explanation (e) Global feature importance.

Phase 2. Post-interview questions.

At this stage, follow-up interviews were used to clarify the collected observations and participants think-aloud data and gather insights from the participants about their lived experience with AI explanations. This helped us to understand the nature of the users' errors and confirm our observations. The following questions summarises the questions asked to the participants.

General questions.

1. How would you summarise why the AI-supported decision tool made the recommendations?
2. What did you think of this explanation?
3. Can you explain the results of the AI recommendation in your own words?
4. How do you think the explanation could help you in your everyday decision-making activity?

Questions regarding a specific action during the think-aloud protocol.

1. Can you tell us why did you do that?
2. What did you think about that scenario?
3. What would you do in that scenario if you were in your clinic?

Code

Quotes from qualitative data

Qualitative data codes and examples

<u>Lack of curiosity</u>	<p>P5 mentioned, “... to be honest with you, I was not really interested in reading the explanation ... I mean I did not feel that could add something new to me”.</p> <p>P12 mentioned, “ ... I am not curious to find why the AI thinks so ... I just want to know the results”.</p>
<u>Perceived goal impediment.</u>	<p>P12 mentioned, “... that [explainability] experience was good in general ... but I doubt that it could work in real-world ... doctors and pharmacists are too busy to validate each decision with an explanation”</p> <p>P6 added, “... I cannot see how these explanations will work in everyday prescriptions screening”.</p> <p>P9 asked to customise the number of data features in Local explanations, “The average pharmacist does not need to see all these factors that the AI is considering, some of them are just simple rules”.</p>
<u>Redundant information</u>	<p>P9 stated, “The average pharmacist does not need to see all these factors that the AI is considering, some of them are just simple rules”.</p> <p>P6 criticized Counterfactual explanation and stated, “... mentioning the AI could change its decision if age was 29 does not consider as a useful explanation in our setting ... I mean we all know that ... explanations should be smart enough”.</p> <p>P12 mentioned, “Does this mean that I have to look at all these factors each time I make a decision?”.</p>
<u>Perceived complexity.</u>	<p>P11 ignored a Global explanation but read and engaged with Counterfactual explanation, and mentioned: “It could be useful, but I won’t bother digging what does that mean”.</p> <p>P12 stated during Global explanation scenario, “I would usually look for the first three or four values”.</p> <p>P6 stated, “.... sometimes we are so busy; I won’t have that time to validate the AI through its explanation; in my opinion, a simple explanation targeting main patient issues would be enough with an option to investigate more when needed”.</p>
<u>Lack of context.</u>	<p>P8 stated, “I find this irrational, the explanation is saying the prescription would have been prescribed if the patient age is 50 ... I mean patient age is not something we can change ... I expected something like a blood test or any other variable that we can do something about it”.</p> <p>P9, who skipped a Global explanation mentioned, “I would like also to see correlations between patient information to judge whether this is valid information in this case”.</p> <p>P12 commented on the Local explanation presented during the study: “I feel this could be biased in some way, so that means the majority of the decisions will be made based on the tumour size”.</p>
<u>Misinterpretation.</u>	<p>P2 stated: “... so shall we screen all prescriptions only on blood results?”.</p> <p>P9 had a false interpretation of a Confidence explanation and stated that</p>

	<p><i>“44% certainty in a diagnosis is a good value”.</i></p> <p>P9 mentioned, <i>“I think it is unfair for AI to explain this way because it just does whatever it was designed to explain for, so it does not give us to see the big picture ... I would like to know what it means to have a patient age with 35% influence on the AI decision? and how this could be interpreted for this patient?”.</i></p> <p>P8 commented on the Global explanation encountered during the study, <i>“I saw that blood test is the influential factor, and I was wondering we should screen prescriptions on that factor only?”.</i></p>
<u>Mistrust.</u>	<p>P8 noted, <i>“I am wondering if an experienced pharmacist has looked at this before”.</i></p> <p>P5 wondered if Local explanation considered data coming from different hospitals, <i>“we have got to know which hospital this explanation covers, this could completely change my opinion about this explanation”.</i></p> <p>. P7 stated, <i>“As far as that is concerned, I cannot tell whether this explanation is right or wrong without knowing it is up to date”</i></p> <p>P13 added, <i>“from time to time we get emails to tell us the treatment x got recognised for diagnosing breast cancer. We need to ensure that the system knows this information”.</i></p> <p>P1 mentioned, <i>“I was wondering if this explanation could be generalised for another patient”</i></p>
<u>Confirmatory search</u>	<p>P4 who is a pharmacist stated, <i>“Well, I would look for the examples that I’ve already experienced in the past”.</i></p> <p>P1 commented, <i>“I think this is crucial when I am sitting in the clinic and I need to make a decision, examples allow me to ask a whole range of questions even it is one that what will your prognosis be what will the outcome be what how should I treat the patient how can I tell what events would be”.</i></p> <p>P2 commented, <i>“There is a lot of correlation between the treatment cycle and the patient history when you are aware that the system considering this correlation, I will be able to tell if the explanation is accurate”.</i></p>
<u>Rush understanding.</u>	<p>P4 stated, <i>“Well in many cases I could predict how the AI work after reading the explanations in the first two cases”.</i></p> <p>P7 mentioned, <i>“... I would say that I have a confidence to tell how it worked”.</i></p> <p>P8 described <i>“I think it was easier to read and recognise when this explanation [Local] groups patient history information in one value”.</i></p>
<u>Habits formation.</u>	<p>P4 who showed similar behaviour mentioned, <i>“I think this is similar to the previous explanation”.</i></p>

12.4 APPENDIX 4. CALIBRATED TRUST DESIGN PRINCIPLES STUDY MATERIAL

Calibrated trust has become an important design goal when designing Human-AI collaborative decision-making tools. It refers to a successful understandability, reliability and predictability to the AI-based tool behaviour and recommendations. eXplainable AI (XAI) is an emerging field where explanations accompany AI-based recommendations to help the human-decision maker

understand, rely on, and predict AI behaviour. Such an approach is supposed to improve humans’ trust calibration while working collaboratively with an AI. However, evidence from the literature suggests that explanations have not contributed to improved trust calibration and even introduced other errors. Designers of such explainable systems often assumed that humans would engage cognitively with AI-based explanations and use them in their Human-AI collaborative decision-making task. In this paper, we devise XAI design techniques and principles for XAI interfaces to enhance the role of explanations in calibrating users’ trust. We focus on model-agnostic explanations in high stake applications. We used screening prescription as a Human-AI collaborative decision-making task where the human medical practitioner uses the AI to check whether the prescription can be approved to a given patient. Such a task reflects an everyday Human-AI collaborative decision-making task where trust calibration errors are possible. We follow a multi-stage qualitative research method, including think-aloud protocol and co-design sessions with medical practitioners. Our results shed light on the nuances of the lived experiences of users of XAI and how the design can help their trust calibration.

First, we conducted a systematic literature review to identify what can be explained to end-users given a black-box AI model. Second, we conducted a think-aloud session to observe and understand how human decision-makers interact with AI-based explanations during a Human-AI collaborative decision-making task, i.e., what kind of errors could happen in real-time interaction. Finally, we conducted a co-design study with end-users to identify techniques and principles to guide the XAI interface to help trust calibration and mitigate errors. Figure 1 summarises the research method. The following sections describe each phase and its used material.

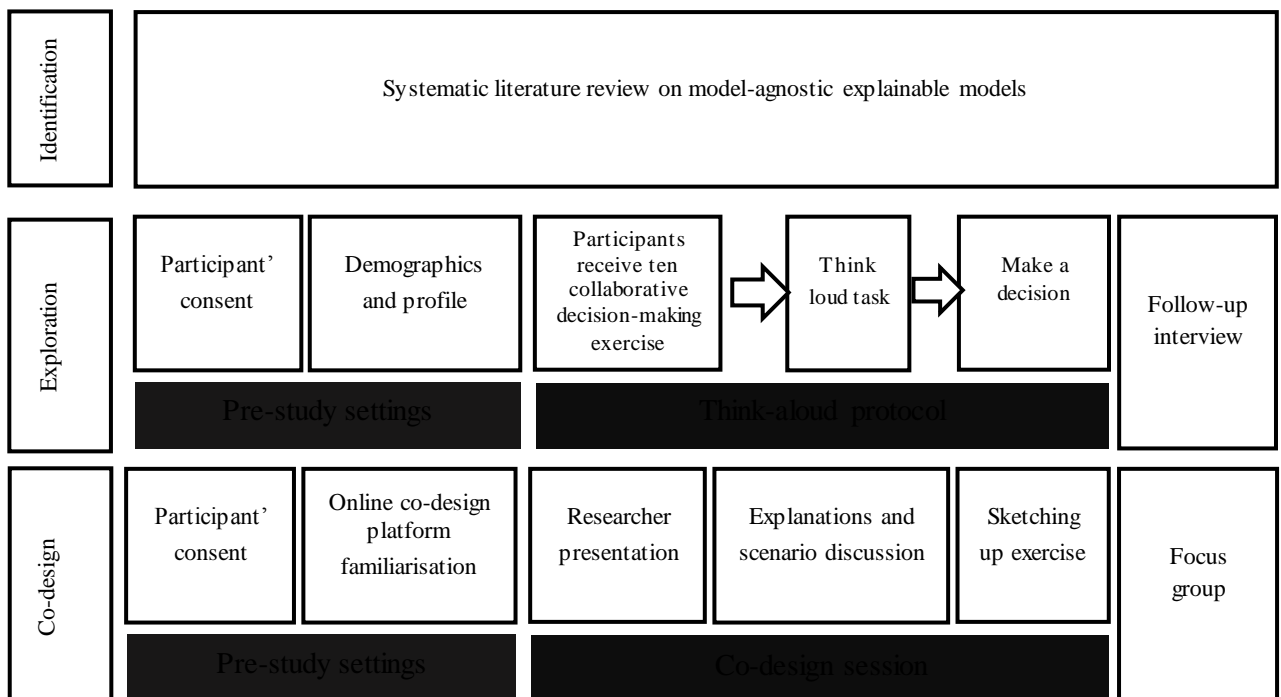


FIGURE 52 MULTI-STAGE QUALITATIVE STUDY WORKFLOW

Co-Design.

We conducted two co-design sessions with eight participants, i.e., four participants in each session. The main aim of this stage was to explore how the design can play an effective role in enhancing users’ trust calibration during a Human-AI collaborative decision-making task. We

used the same inclusion criteria employed in the exploration stage, i.e., expert users in the studied task. We chose to recruit different participants to avoid the learning effect (Lazar et al., 2017) and increase the credibility of our findings as existing users already learned the objective of the study and were part of the underpinnings for this next study. Co-design method enables users who might be potential users in future AI-supported decision-making tools to reflect their experience in the design process, and this is supposed to increase the acceptance of the proposed solutions (Poole et al., 2008). Co-design can lead to a better understanding of the end-user needs, which enhances the possibility of the designs' acceptance (Song and Adams, 1993). In this phase, we discussed and negotiate how to embed AI explanations to serve users' needs, task workflow and trust calibration. Together with the participants, we conceptualised and sketched design features to support users in utilising AI explanation and reduce trust calibration errors revealed from the exploration phased. This was achieved by giving the participants initial prototypes or mock-ups (Clement et al., 2012) of the problem to help them visualise the idea and then provoke brainstorming related to the research problem. All these dynamics were hard to capture during the exploration phase. Therefore, co-design method helped us to come up with innovative designs of how the solution should look from a user perspective.

Participants were divided into two design sessions based on their availability. Due to the COVID-19 situation, we chose to conduct the study online using FreeHand tool from Invision³. Also, it has been shown that online tools for co-design can make the process easier, cheaper and flexible for participants (Näkki and Antikainen, 2008). To mitigate any potential issues that could arise from using online platforms, e.g., readability of the instructions and the tool usability issues, we conducted a pilot study with two post-graduate researchers and one academic in an interdisciplinary research group residing in the departments of Computing and Psychology in Bournemouth University. This also helped us in the preparation of the training and induction stage for the participants in the real study. All participants attended a training session to familiarise themselves with the tools' functionalities and how they can communicate online. The training session lasted for 15-20 minutes. Then participants were invited to try the tool till they felt all capable of using it. They had the ability to ask questions and one of the authors answered them.

We adopted four techniques during the co-design sessions in order to reach the goal of our study (See Figure 2); researcher presentation, participants discussion, sketching-up exercise and focus groups. This also helped to enhance the credibility of the study and to ensure that data bias was eliminated. Each of the sessions lasted for around 2 hours. Both sessions, including the four main steps, were audio-recorded and transcribed. Audio recording for the design session helped the authors analysing main design needs and issues revealed from participants discussions. The following sections describe each technique that we used in our design sessions.


1) Researcher presentation (10 mins). The researcher gave a 10-minute presentation on AI-based decision-making tools and an overview regarding the first phase findings, particularly those about different types of errors that emerged during the exploration study. This helped to immerse the participants in the research problem, and it involved a warming-up activity in getting the participants involved in the design sessions.

2) Explanation and scenario discussion (25 mins): In this stage, participants started by introducing themselves. We then asked each participant to talk about how AI-based tools could help their everyday decision-making process. Then, we provided a definition for explainability methods introduced in previous interpretable machine learning surveys (Adadi and Berrada, 2018). We provided different e-cards describing different explanation types in simplified examples. This was meant to illustrate explainability definition and potential uses of these explanations. To answer our research question, the participants needed first to immerse in a fictional problem as recommended in (Buskermolen and Terken, 2012). In our study, the fictional problem was collaborative decision-making between the medical expert and the AI. Specifically, a screening prescription using and AI-based tool. The researcher invited participants to discuss the designed scenario of an AI-based collaborative decision-making tool

³ <https://freehand.invisionapp.com/freehand/new>.

of a screening prescription and its generated explanations. We used a random forest classifier as an ML algorithm to train our model. We then generated explanations from current state-of-art model-agnostic explanations to examine how users would like to receive these explanations and develop prototypes for effective utilisation for such explanations in real-world scenarios. This stage was meant to scope the discussion and facilitate focused conversations using the provided scenario. This was also meant to immerse the participants with the research problem and facilitate their understanding of the researcher presentation. Our participants discussed a wide range of trust calibration scenarios using the explanation interfaces through the provided material in this stage. This stage provided a sense of realism to the problem and encouraged careful consideration of solutions to cater to different contexts and usage styles. The following scenario and explanations were presented to our participants and discussed in this stage. We asked our participants to use the output from five explainable models and sketch up designs that help them to have appropriate trust in the AI recommendation and help them in their everyday Human-AI collaborative decision-making task. Below we describe the provided scenario and how we generate the explanations.

John is a doctor using AI-supported decision-making tool that recommends if a prescription shall be confirmed or rejected. While John was trying to understand why the AI is recommending that, he wanted to make informed decision using the below explanations. This might trigger two circumstances: either to reject correct AI recommendation or follow incorrect AI recommendation.



Patient **Emily** is 27 years old, does not smoke, and she is getting a treatment for cervical cancer.

Our AI suggests that the provided prescription shall be rejected with a confidence score 72.6%

The AI explains its' recommendation using the following explanations:

(See Figure 7 that describes patient scenario).

FIGURE 53 PROVIDED PATIENTS' PROFILE IN THE DESIGN SESSIONS

Using the provided explanations, please answer the following questions:

1. How do you think each explanation should be designed to help you understand the AI recommendation?
2. How would you design the explanation to help you assess of the reliability of AI explanations?
3. How would you design the explanation to help you in judging the accuracy of the AI recommendation and its explanation?

Global feature importance. We used eli5⁴ library in python to generate the global feature importance explanation. Below we see the importance features in the overall model recommendation.

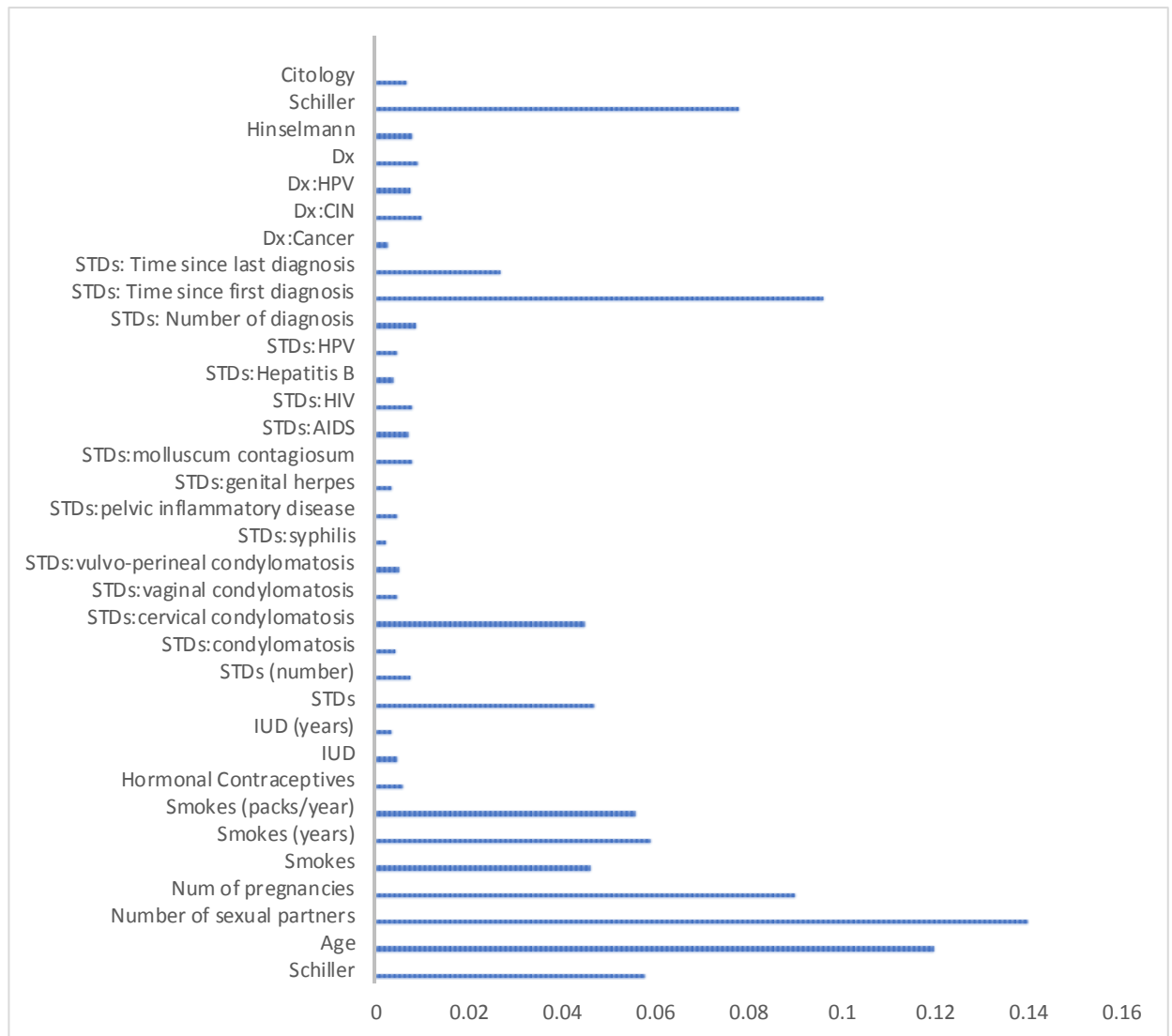


Figure 54 Global feature importance

Local Feature importance. We used LIME⁵ to generate local feature importance given a patient record. Our model recommended that the patient does not have a cancer with 72.6% confidence. Figure 9 shows the generated local feature importance that shows why our system provide this recommendation (Minus values contributed to patient has a cancer, whereas positive values contributed to patient does not have cancer).

⁴ <https://eli5.readthedocs.io/en/latest/overview.html>

⁵ <https://github.com/marcotcr/lime>

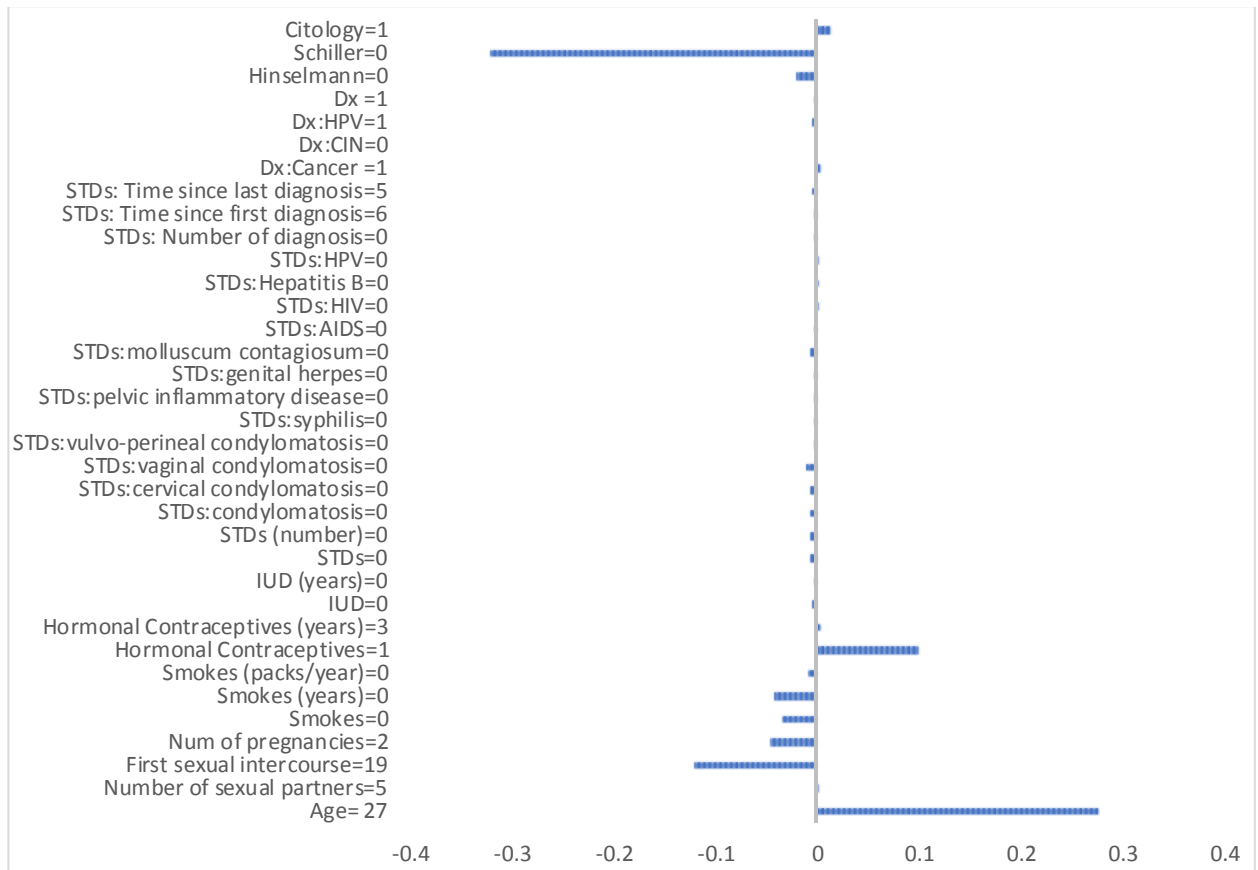


Figure 55 Local feature importance explanation

Counterfactual explanations. We used Alibi⁶ library to generate counterfactuals given the same patient record. Below the generated counterfactual explanation that shows why our system provide this recommendation. Figure 10 shows the generated counterfactual explanation.

The patient would have a cancer with 67% confidence, if the First sexual intercourse=29 and Hormonal Contraceptives (years) = 13.

Figure 56 Counterfactual explanation

Example-based explanation. We used K-nearest neighbour algorithm to retrieve the k neighbours for the same patient record.

⁶ <https://pypi.org/project/alibi/>

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	IUD (years)	STDs	STDs (number)	STDs:condylomatosis
0	18	4.0	15.0000	1.000000	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0
1	15	1.0	14.0000	1.000000	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0
2	34	1.0	16.9953	1.000000	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0
3	52	5.0	16.0000	4.000000	1.0	37.000000	37.000000	1.0	3.00	0.000000	0.000000	0.0	0.0	0.0
4	46	3.0	21.0000	4.000000	0.0	0.000000	0.000000	1.0	15.00	0.000000	0.000000	0.0	0.0	0.0
5	42	3.0	23.0000	2.000000	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	0.0	0.0	0.0
7	26	1.0	26.0000	3.000000	0.0	0.000000	0.000000	1.0	2.00	1.000000	7.000000	0.0	0.0	0.0

Figure 57 Example-based explanations using KNN.

Confidence score. For confidence score, we used the function `predict_proba` implemented in the Random Forest library. The algorithm confidence score for the patient record was **72%**.

3) Sketching-up exercise (40 mins): Participants were then encouraged to start sketching-up their designs using FreeHand tool from InVision. We gave each participant a blank e-page to sketch up designs considering five explanation types (Local, Global, Example-based, Counterfactual and Confidence explanations). The online platform provided several creation tools (e.g. coloured pens, shapes and sticky notes). The participants were also asked to not limit themselves to the given explanation classes and consider any extra features they would like to see in XAI interfaces to help them in utilising the explanation during a collaborative decision-making task. We deliberately asked our participants to work individually, think outside of the box, and consider different kinds of potential solutions. In this stage, our participants designed their explanations and provided multiple usage scenarios for them. They created a wide variety of usage scenarios covering different purposes and task requirements, e.g., grouping data features in Local explanations to reduce the explanation complexity. Below we provide a screenshot of the design space provided to our participants (Figure 12).

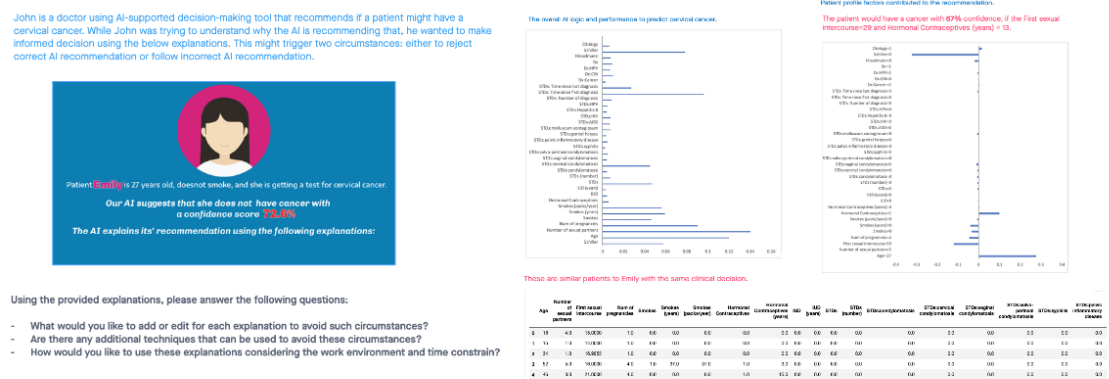


FIGURE 58 DESIGN SPACE PROVIDED TO OUR PARTICIPANTS

4) Focus group (45 mins). After each participant completed the sketching activity, each participant presented their ideas to the group. This was meant to critically analyse and evaluate the ideas by the participants in order to formulate robust solutions. This activity allowed our participants to explore and discuss various ways of using AI explanations in their work environment, considering trust calibration as the primary goal. Figure 13 shows a sample of the designs generated during the focus-group stage.

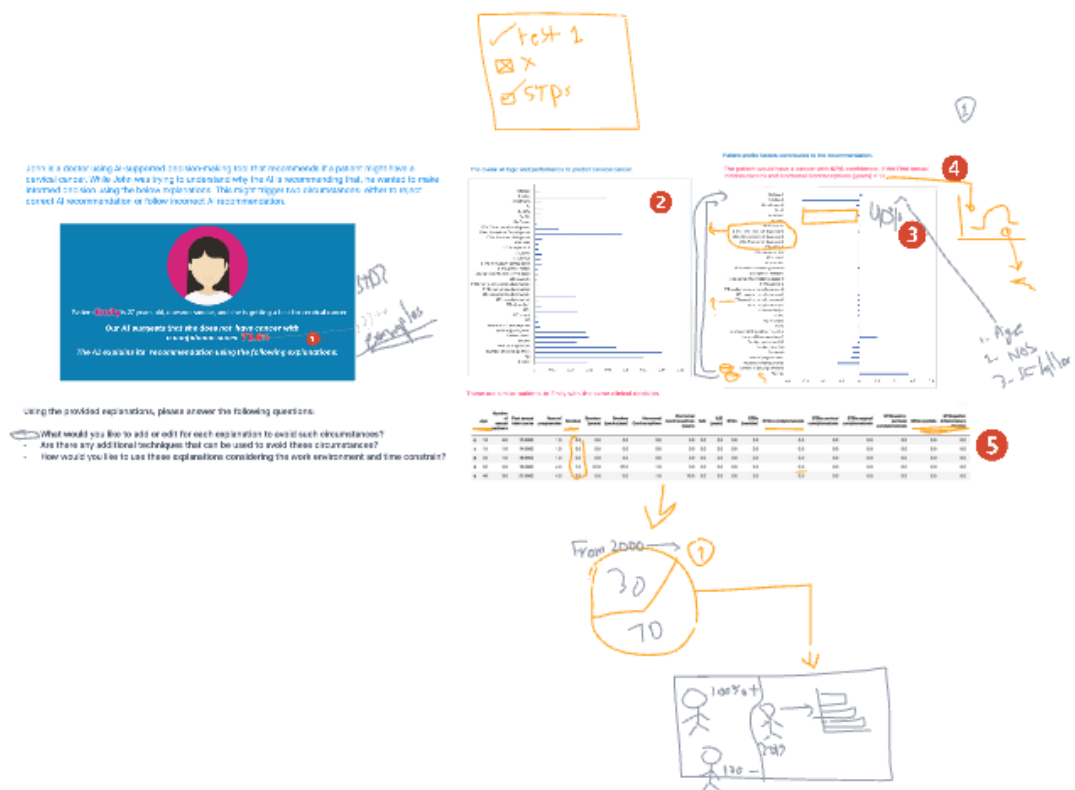
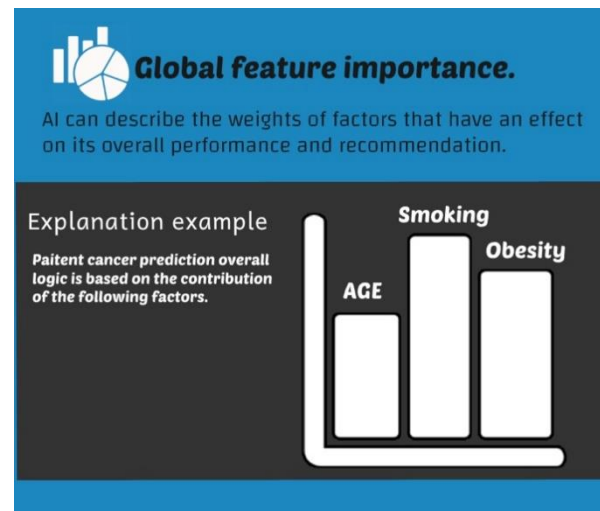
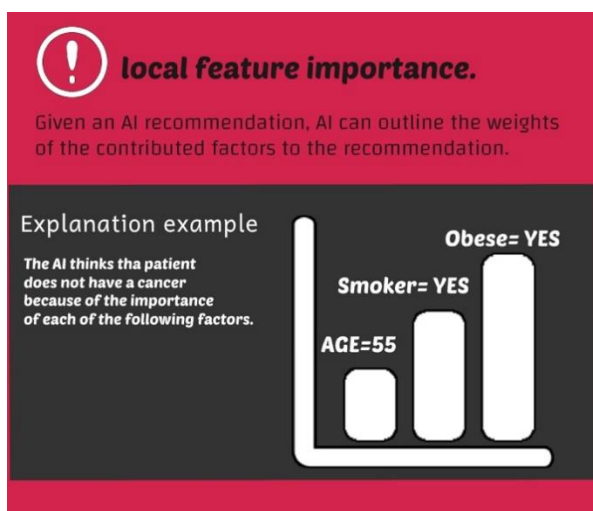


FIGURE 59 SAMPLE OF THE COLLECTED DATA.

Cards provided to the participants to facilitate the discussion.





Counterfactuals

Given an AI recommendation, AI can explain what circumstances could change the recommendation.

Explanation example



The AI suggests to reject the loan.

AI denied the loan because of annual income was £30,000. If the income had been £45,000, the loan would have been offered (Wachter et al.2017).

How confident?

Given an AI recommendation, AI can say how much confidence have on the provided recommendation?

95%



Explanation example

The AI has 78% confidence that the patient has a cancer



Examples

Given an AI recommendation, AI can explain its recommendation using previous similar examples.

Explanation example

The AI suggests that the patient does not have cancer because he is similar to the following patients:

