Evaluating Explanations of Artificial Intelligence Decisions: the Explanation Quality Rubric and Survey

Charlotte Young

Bachelor of Information Technology (Professional Practice)

Bachelor of Science (Honours)

Submitted in total fulfilment of the requirements for the degree of

Doctor of Philosophy

Federation University Australia

April 2022

School of Science, Engineering, Information Technology

PO Box 663

University Drive, Mount Helen

Ballarat Victoria 3353

Australia

ABSTRACT

The use of Artificial Intelligence (AI) algorithms is growing rapidly (Vilone & Longo, 2020). With this comes an increasing demand for reliable, robust explanations of AI decisions. There is a pressing need for a way to evaluate their quality.

This thesis examines these research questions:

> *What would a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations look like?*
>
> *How can a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations be created?*
>
> *Can a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations be used to improve explanations?*

Current Explainable Artificial Intelligence (XAI) research lacks an accepted, widely employed method for evaluating AI explanations. This thesis offers a method for creating a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations. It uses this to create an evaluation methodology, the XQ Rubric and XQ Survey. The XQ Rubric and Survey are then employed to improve explanations of AI decisions.

The thesis asks what constitutes a good explanation in the context of XAI. It provides:

1. a model of good explanation for use in XAI research
2. a method of gathering non-expert evaluations of XAI explanations
3. an evaluation scheme for non-experts to employ in assessing XAI explanations (XQ Rubric and XQ Survey).

The thesis begins with a literature review, primarily an exploration of previous attempts to evaluate XAI explanations formally. This is followed by an account of the development and

iterative refinement of a solution to the problem, the eXplanation Quality Rubric (XQ Rubric). A Design Science methodology was used to guide the XQ Rubric and XQ Survey development.

The thesis limits itself to XAI explanations appropriate for non-experts. It proposes and tests an evaluation rubric and survey method that is both stable and robust: that is, readily usable and consistently reliable in a variety of XAI-explanation tasks.

STATEMENT OF AUTHORSHIP AND ORIGINALITY

Except where explicit reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis by which I have qualified for or been awarded another degree or diploma. No other person's work has been relied upon or used without due acknowledgment in the main text and bibliography of the thesis. No editorial assistance has been received in the production of the thesis without due acknowledgement. Except where duly referred to, the thesis does not include material with copyright provisions or requiring copyright approvals.

Charlotte Young

Federation University Australia

15 June 2022

ACKNOWLEDGEMENTS

My research was supported by an Australian Government Research Training Program (RTP) Stipend and an RTP Fee-Offset Scholarship through Federation University Australia.

I wish to thank Dr Peter Vamplew, Dr Cameron Foale of Federation University, and Dr Richard Dazeley of Deakin University for their contributions and guidance. I appreciate the help of Dr Andrew Stranieri and Dr Sally Firmin.

I am grateful for the assistance of the focus group participants and Amazon Mechanical Turk participants.

CONTENTS

FIGURES

TABLES

APPENDICES

STATEMENT OF ETHICS APPROVAL

The Ethics Approvals are in Appendix B – HREC Approvals. The associated applications are provided in Appendix B5 – Human Research Ethics Committee Applications.

They consist of:

Appendix B1 – Focus Group

Appendix B2 – Initial Validation

Appendix B3 – Constructive Feedback

Appendix B4 – Independent Validation

ABBREVIATIONS

Abbreviation	Meaning
AI	Artificial Intelligence
DSRM	Design Science Research Methodology by Peffers, Tuunanen, Rothenberger, and Chatterjee (2007)
GDPR	General Data Protection Regulation
IT	Information Technology
ML	Machine Learning
MTurk	Amazon Mechanical Turk
NLG	Natural Language Generation
NN	Neural Network
XAI	eXplainable Artificial Intelligence
XQ Rubric	eXplanation Quality Rubric
XQ Survey	eXplanation Quality Survey

1 INTRODUCTION

Artificial Intelligence (AI) is the field of science and engineering whose object is to investigate, design, and construct machines capable of acting in intelligent ways, following schemes of reasoning analogous to those of humans (Russell & Norvig, 2016).

AI has found application in education, government, business, industry, communications, and warfare: in every aspect, it seems, of modern life (Vilone & Longo, 2021b). It has been used to set bail terms and prison sentences (Brennan, Dieterich, & Ehret, 2009), to decide who receives a job offer (Dastin, 2018), and AI tools are being developed that can determine whether a person's loan request should be accepted or rejected (IBM Research, 2019).

A bad decision by an AI is increasingly likely to be harmful, so it is important that AI decisions should be able to be evaluated (Wang, Zhang, & Lim, 2021). Explainable Artificial Intelligence (XAI) is a response to this concern (Longo, Goebel, Lecue, Kieseberg, & Holzinger, 2020; Zhang & Lim, 2022). XAI deals with the explanation of an AI's actions, both explanations proffered by the AI itself and those tendered on its behalf by its human designers and users (Arrieta et al., 2020).

XAI has a pivotal role in the development of AI (Kim et al., 2021). Clearer, better explanations lead to deeper understanding of an AI's reasoning. This means improved AI design and increased accountability for an AI's actions.

My review of XAI literature found a lack of consistency about what constitutes good XAI explanation. In particular, there has been little research on possible schemes for gathering non-expert assessments of XAI explanations. In response, this thesis develops and proposes an assessment method, a rubric— the XQ Rubric — to guide human evaluation of XAI explanation. Associated with this is a survey methodology founded on user-centred design, that is, user-driven, iterative development of the explanation project.

A rubric (from the Latin 'rubeus': 'red') originally referred to a red-coloured heading that designated the aspect under which related matters in a document were considered (Santa Cabrera, Castillo, & Jimenez, 2017). The term ‘rubric’ now commonly refers to tables used to mark students' assignments (McNeill & Krajcik, 2006) which set out detailed requirements for each level of points allocated, separating the marking into aspects for evaluation (Figure 1-1).

Component	Levels		
	0	1 & 2	3
Claim – An assertion or conclusion that answers the original question.	Does not make a claim, or makes an inaccurate claim.	Makes an accurate but incomplete claim.	Makes an accurate and complete claim.
Evidence – Scientific data that supports the claim. The data needs to be appropriate and sufficient to support the claim.	Does not provide evidence, or only provides inappropriate evidence (Evidence that does not support claim).	Provides appropriate, but insufficient evidence to support claim. May include some inappropriate evidence.	Provides appropriate and sufficient evidence to support claim.
Reasoning – A justification that links the claim and evidence and shows why the data counts as evidence to support the claim by using the appropriate and sufficient scientific principles.	Does not provide reasoning, or only provides reasoning that does not link evidence to claim.	Provides reasoning that links the claim and evidence. Repeats the evidence and/or includes some scientific principles, but not sufficient.	Provides reasoning that links evidence to claim. Includes appropriate and sufficient scientific principles.

Figure 1-1 ‘Base Rubric’ reproduced from McNeill and Krajcik (2006)

Rubrics, a general assessment tool, offer many advantages over other evaluation methods in education and related fields. The rubric method is a good fit for this research project, for it can provide usable data and feedback (Moskal & Leydens, 2000), and it is adaptable to different purposes (Lizotte, Harris, McNeill, Marx, & Krajcik, 2003).

Rubric creation and use are usually envisaged in an educational setting and must be broadly interpreted for use in the creation of rubrics appropriate for XAI. The evaluation researchers Boston (2002), McNeill and Krajcik (2006), and Moskal and Leydens (2000) all advocate the use of rubrics. They stress the need for a rubric’s designer to examine the requirements of the assessment task and tailor the rubric to it.

McNeill and Krajcik (2006) suggest a method for constructing a more specific rubric from a base-explanation rubric (a base explanation rubric is a “*general rubric for scoring an inquiry practice across different content and learning tasks*” (McNeill & Krajcik, 2006, p. 12)). They demonstrate

that more specific rubrics encourage better scientific explanations than those proposed by researchers who did not use rubrics.

Moskal and Leydens (2000) stressed the need for a reliable and valid rubric. A reliable rubric generates the same score no matter who the marker is, and a valid rubric reflects the quality of the explanation assessed in its scoring. Very little research has been done on the validation of rubrics.

1.1 MOTIVATION

There is abundant evidence that AI researchers and practitioners have, until recently, given little attention to the issue of explaining how and why an AI made its decisions (Arrieta et al., 2020). Too often, problems that arise from inadequate explanation have gone unrecognised, and the usefulness of good explanation has been undervalued (Zhang & Lim, 2022).

In 2018, Amazon.com admitted that when its recruitment division used a Machine Learning (ML) algorithm to select candidates for employment, the program disproportionately rejected women candidates (Dastin, 2018). Moreover, even after the recruitment team removed obvious gender markers (such as names and pronouns) from applications, the selection algorithm could still detect that the application had been submitted by a woman from other information in their résumé, such as membership of a sorority or a women's sports team (Tambe, Cappelli, & Yakubovich, 2019).

What was needed was an explanation that unambiguously revealed why the AI had accepted or rejected a candidate. It is highly significant that Amazon’s problems in its recruitment method were discovered only after efforts were made to explain the AI's decisions.

The problem of poorly communicated and misleading explanations can have large repercussions. In the case of the AI used by Amazon's hiring processes, a plausible but inadequate explanation of why the AI recommended hiring or not hiring could easily have given the AI users misplaced

confidence in its functional capability, generating the false belief that the users of the AI had selected the best person for the job.

It is clear from the Amazon case that a well-formed and complete explanation of an AI's decisions is an important part of its operational design and history.

The consequences of bad explanations suggest that a method of evaluating explanations is urgently needed, a method that would establish whether an explanation is robust, complete, and true. It is also worth stressing that a good explanation can also be immensely useful to the operators of the AI and its designers and developers (Arrieta et al., 2020).

The importance of methods for evaluating XAI explanations has, until recently, not received the attention it deserves in XAI research literature (Vilone & Longo, 2021b). The current methods for evaluating XAI are discussed in Chapter 3.

1.2 Research Questions

What would a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations look like?

How can a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations be created?

Can a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations be used to improve explanations?

This thesis responds to these questions by developing a process to create an evaluation scheme, the XQ Rubric and Survey. It demonstrates that the XQ Rubric and Survey is a practical solution to the question of what such an evaluation scheme should look like. The thesis then it verifies that the XQ Rubric and Survey can be used to improve explanations iteratively.

The research question emerged from the literature review, from which it became clear that no rigorous method existed that met evaluators' needs in the assessment of AI decisions. Using a design science and user-centred approach, an evaluation scheme was developed in response.

Giving priority to the needs of the user (in this case, the interests and requirements of the explanation-evaluator) is referred to as 'human-centred' design (Cooley, 2000). Human-centred design rejects the notion that there is a single best way to do something. It instead emphasises the needs of the users in the production of an artefact that matches their requirements (Zoltowski, Oakes, & Cardella, 2012).

To complement the human-centred approach to design, this thesis takes a constructivist approach to the research and creation of its evaluation scheme for XAI explanation. The constructivist paradigm regards knowledge as a human construct, developed through experience and reflection. (Adom, Yeboah, & Ankrah, 2016).

1.3 CONTRIBUTIONS

This thesis proposes ways of correcting three shortcomings of apparent in the literature of explainable AI (Chapter 3):

1. there is no clear account of what constitutes a good explanation of the decisions of an AI
2. non-experts have no ready way to evaluate candidate explanations formally
3. there is no straightforward and reliable method to prepare evaluations of XAI explanations for use by non-experts

In response to these shortcomings, this thesis provides:

1. a model of a good explanation for use in XAI research (Section 2.5.4)
2. a methodology for acquiring non-expert evaluations of XAI explanations
3. an evaluation scheme for non-experts to use in evaluating XAI explanations (XQ Rubric and XQ Survey).

These three outcomes are the primary contributions of this thesis. The second and third are listed separately because they can be used separately: the methodology is not tied to the evaluation scheme, and vice versa. The outcomes offer solutions to gaps in the literature (Chapter 3) and provide comprehensive answers to the research questions.

1.4 Objectives and Approach

The explanation-evaluation methodology described in this thesis was created in four stages, following the research methodology of Peffers et al. (2007) (Chapter 4).

First, an extensive review of current literature on XAI was conducted to ascertain the characteristics of good explanations and evaluation methods (Chapter 3). From this was drafted (Chapter 5) a prototype evaluation rubric, the XQ Rubric.

Second, the prototype evaluation rubric was assessed using an iterative design methodology (Chapter 4). Employing a Delphi approach (Chapter 5), the XQ Rubric was presented to a group of experts in various fields (to eliminate bias, none of them XAI). After three rounds of discussion and improvement, the new rubric was given to a group of users on Amazon Mechanical Turk (MTurk) (Chapter 6). The feedback from the first MTurk experiment was used as the basis for further improvements to the XQ Rubric.

Third, after the XQ Rubric had been created, and a case study was edited using results from the Initial Validation Experiment, MTurk workers evaluated the edited case study to ascertain

whether the explanation had improved (Chapter 7). This experiment was designed to show that the XQ Rubric and Survey could be used as a feedback mechanism to improve explanations.

In order to demonstrate that the XQ Rubric and its methodology constituted an effective approach for the improvement of explanations, I recruited an independent XAI researcher, Dr Francisco Cruz, of Deakin University, to apply the evaluation methodology in his own research. Dr Cruz presented a new set of XAI explanations to MTurk workers to evaluate. This experiment demonstrated that the tools developed could be usefully incorporated into a typical XAI research project (Chapter 8). I was able to establish that my rubric and methodology enabled Dr Cruz to improve his explanation of his AI's decisions.

Scope

The scope of this thesis was limited to:

1. Evaluators who were not XAI experts
2. Explanations designed for AI or XAI

The people who participated in the experiments in this thesis were not XAI experts. The XQ Rubric and Survey created by this thesis was intended for use by non-experts, so non-experts were consulted in their creation. The inclusion of non-experts was inspired by interviews of stakeholders in studies conducted by Bhatt, Andrus, Weller, and Xiang (2020). Stakeholders requested more involvement by non-academics and non-experts in XAI.

The type of explanation used in the development of the XQ Rubric and Survey was limited to that appropriate for XAI. The thesis considers no other form of explanation.

1.5 Thesis Layout

This thesis begins with an overview of important concepts in XAI explanation and evaluation. This is followed by a review of previous on the subject. The literature review creates a new two-dimensional scheme for viewing XAI evaluation methods. This showed that there were a considerable number of explanations that relied on a supposedly 'self-evident' view of what constitutes a good explanation, that it is self-evidently a good one. The literature included no significant attempt to incorporate the views of non-experts into explanatory designs.

The methodology of this thesis is then discussed, with special attention to Design Science and User-Centred Design, its main methodological tools.

A series of experiments were conducted. The first, The Focus Group Using the Delphi Methodology, developed and established an evaluation method for XAI (the XQ Rubric). The second, Initial Validation of XQ Rubric and XQ Survey as Evaluation Tools, developed and established a methodology (the XQ Survey) for using the evaluation method.

The third experiment, Validation of XQ Surveys as a Tool for Constructive Feedback, demonstrated the use of the XQ Rubric and XQ Survey to revise an XAI explanation successfully.

The fourth and final experiment, Independent Validation of the XQ Rubric and the XQ Survey, demonstrated that someone who was not familiar with the creation of the XQ Rubric and XQ Survey could nevertheless successfully revise an XAI explanation using the methodology and tools proposed.

The thesis concludes with a discussion of the results of these experiments and offers recommendations for future research.

2 BACKGROUND

This chapter gives examples of cases in which a good XAI explanation might have prevented the failure of an AI to achieve its objective, then offers a grounding in ideas and theories related to XAI and the evaluation of XAI explanations.

2.1 EXAMPLES OF FAILURES OF EXPLANATION IN ARTIFICIAL INTELLIGENCE

To provide more context for later discussion, this section has two examples of explanation failures. By clarifying the reasons for which the ML system made its decision and so allowing developers and users to notice errors in its logic, better assessment methodology would clearly have been apparent (Arrieta et al., 2020).

2.1.1 The Importance of Background Features in Training Neural Networks

A research team trained a Neural Network (NN) to recognise tanks in photographs by showing it two sets of photographs, one set with and one set without tanks (Whitby, 2009). By this method, the NN learned to differentiate between the two sets and was able to decide whether a given photograph included a tank. The NN achieved great accuracy, even 'detecting' a tank out of sight behind a dune (Whitby, 2009). This, of course, was impossible. The high accuracy rate claimed for the NN was not actually achieved.

The cause was investigated. It emerged that the photographs of tanks had been taken in the morning. Those without tanks had been taken in the afternoon (Whitby, 2009). The NN had identified the position of shadows as the best predictor of whether there was a tank in the photograph (Whitby, 2009). While this was an accurate way of differentiating the photographs, it did not achieve the creators' goal of reliably recognizing tanks in photographs.

2.1.2 Systematic Discrimination Repeated by AI

Courts in some U.S. jurisdictions use an advisory risk assessment algorithm called “Correctional Offender Management Profiling for Alternative Sanctions” (COMPAS) for bail setting and sentencing (Brennan et al., 2009). It appears that offenders whose outcomes were influenced by references to COMPAS were given bail and prison terms less lenient than those whose terms were decided only by officials of the court (Zhang & Han, 2022).

The reason seems to be that COMPAS was trained partly on historical sentencing data, inheriting the biases of human sentencing (Zhang & Han, 2022). (Unfortunately, the algorithm is proprietary, and the developers have made very little public about its design and operations (Angwin, Larson, Mattu, & Kirchner, 2016).)

Setting aside the social and political issues of 'predictive policing' and legal questions of access and fair play, from an AI point of view, the COMPAS algorithm, if it is indeed based too heavily on historical data, must be judged inadequate to its purpose, too liable to recycle historical prejudice.

2.1.3 Conclusion

ML algorithms are trained on historical patterns. If the historical patterns are biased, the ML algorithms will be biased (Zhang & Han, 2022). However, a robust XAI system should be able to explain why its decision was made, and subsequent human analysis should be able to determine whether the decision was fair (Wang et al., 2021).

2.2 Important Concepts in Artificial Intelligence

2.2.1 Black Boxes

The concept of 'black boxes' is important in computer science and XAI (Holzinger, 2017). In the world of computing, a 'black box' is a program whose internal operations cannot easily be understood (du Boulay, O'Shea, & Monk, 1981). AI methods such as deep learning and NN are often regarded as black-box algorithms.

The intricacies of many ML algorithms, such as NN, are difficult to explain to non-experts, and insight into the decisions of black-box operations is often limited and difficult to interpret (Holzinger, 2017). Given enough time, it may be possible for an expert to understand at least some of the decisions made by the ML algorithm, but time and expertise is not always available.

The issue of 'black boxes' is not directly examined by the XQ Rubric developed here. However, many of the algorithms referred to in this thesis, including the case studies discussed in the experiment chapters (Section 4.6), are effectively black boxes, in that their inner workings cannot be readily examined and explained. All the case studies chosen for this thesis were generated after the algorithm had finished learning. This is known as the post hoc explanation method (Moradi & Samwald, 2021).

2.2.2 Data

AI algorithms are trained and tested with data. For many important AI applications, the data is obtained from real-world datasets. Compared to artificial or simulated datasets, real-world data often encodes significant biases, either by un-noticed selection or, intentionally or unintentionally, by encoding human prejudice (Akter et al., 2021).

The misuse of data can cause an algorithm to become biased and inaccurate (Ntoutsi et al., 2020). If AI developers and users fail to discover that the data has been misused, the algorithm

may make unwarranted claims and draw inaccurate conclusions (Akter et al., 2021). See Section 2.1.2 for an example of historical data leading to bias in an algorithm's prediction.

A good explanation will expose an algorithm's bias and inaccuracies, including those caused by insufficient and inadequate data. A good explanation will help mitigate the effects of bad data.

The XQ Rubric attempts to ameliorate the misuse of data by its Verifiability Category (Section 5.5.3). Though this category does not directly consider data, it evaluates the explanation on verifiability, with the aim of helping evaluators by providing more information about how the AI made its decision. This information can be scrutinised for symptoms of data misuse.

2.3 Important Concepts in XAI Literature

2.3.1 Authority and Trust

AI may be given spurious authority by users, represented as having information and competencies sufficient to make well-informed and appropriate decisions (Robinette, Li, Allen, Howard, & Wagner, 2016). These decisions can lead to unthinking compliance and obedience on occasions when people feel directionless (Cialdini, 2007, p. 176). Ribeiro, Singh, and Guestrin (2016) link trust to action; if someone trusts an ML model, they will act on its advice. Importantly, people are more likely to trust a model whose conclusions they understand (Polonski, 2018; Zhang & Lim, 2022).

To demonstrate why care must be taken to make sure a trusted AI is genuinely trustworthy, Robinette et al. (2016) conducted an experiment in which a robot, designated a 'Fire-Safety' robot, attempted to lead subjects the wrong way in a fire evacuation drill. In this experiment, most subjects followed the robot even when they had reason to believe that they were going in the wrong direction (for example, the fire exit sign pointed the opposite way to the direction they were being led). This experiment demonstrates why great care must be taken, for people may blindly follow an AI's bad advice, even against their better judgement.

Simpson (2012) points out that trust is essential for human social life and that humans should, in most circumstances, trust AIs to an appropriate degree. A trusted AI is less likely to be questioned about its decisions, and it will be more likely to be given a place in the human world. However, the blind, unquestioning trust demonstrated in the Robinette et al. (2016) experiment is dangerous. An AI must be able to explain its decisions well (Zhang & Lim, 2022). Only then will the AI deserve the trust placed in it (Wang, Yang, Abdul, & Lim, 2019). Should the AI's explanation suggest to people that they should not trust it, they will be able to make an informed choice on the matter (Zhang & Lim, 2022).

An AI should be able to explain how it came to its conclusions and justify its actions. Without this, people may put their trust in an AI that has come to the wrong conclusion. The XQ Rubric evaluates the explanation of an AI's decisions to make sure that the explanation seems a reliable guide to its trustworthiness.

2.3.2 Law and Social Justice

The increasing use of AI and XAI has profound implications for the law and for justice, including social justice.

Clearly, without knowing the reasons why an AI made its judgements, it is impossible to assess the justice of its decisions. Doshi-Velez et al. (2017) believe that XAI, combined with an understanding of the data employed by it, will be required for legal systems to judge whether an AI system has complied with the law. The European Union General Data Protection Regulation (EU GDPR) has moved to create laws to regulate how data is gathered and processed by ML algorithms (European Union, 2017).

The GDPR has been enforced since May 2018 (Layton, 2022). It updates the Data Protection Directive and explicitly mentions ML algorithms (Wu, 2017). Significantly for XAI, the EU GDPR concluded that people have a legal right to an explanation of an AI algorithm's decision (Doshi-Velez et al., 2017).

Since the EU GDPR has been in effect for only a short while, it is difficult to predict how its regulations will change the design and use of ML algorithms. However, it has been suggested that the code of ML algorithms will become more complex, making it harder to explain their results (B. Goodman & Flaxman, 2017). All this suggests that there will be ever-increasing regulatory complexity. Already a Canadian Marketing Association (2022) report found that the GDPR created a "*staggering regulatory burden*" (p. 7) because of its "*overly complex, prescriptive or otherwise disproportionate provisions*" (p. 5). Despite this, many commentators hope that the European Union limits on the use of AI will provide legal structures for consumers to demand explanations of an AI's decision (Doshi-Velez et al., 2017).

2.3.3 Truth

In the matter of what constitutes the truth of XAI explanations, this thesis employs the common-sense, 'correspondence' theory, which considers truth to be an alignment between the truth value of a proposition and the reciprocally related fact in the world (David, 2020). Philosophical debate on the matter is outside the scope of the thesis.

2.4 Overview of Assessment and Evaluation in Adjacent Fields to XAI

2.4.1 Formative Evaluation

As discussed in the Methodology chapter (Section 4.2), this thesis follows an iterative design philosophy, that is, a process of continuous evaluation and improvement. Formative Evaluation, the evaluation theory used for the XQ Rubric, also follows a method of improvement based on iterative evaluation (Nieveen & Folmer, 2013). Formative Evaluation is an evaluation methodology that emphasises improvement as a use of evaluation (Nieveen & Folmer, 2013).

In accordance with the Formative Evaluation view of evaluation, rather than simply creating an assessment scheme to rank explanations by criteria judged to be suitable, the XQ Rubric and XQ Survey aim to improve the explanation itself. The do this by gathering feedback in a structured

way. This feedback can then provide a point of reference to help users construct an explanation and improve their existing explanations.

Muller (2019) identifies the ability to game metrics as a significant flaw in their use. The term 'to game a metric' expresses the idea that people who know that "*they will be evaluated by some numeric score, […] will be encouraged to perform in ways that will produce better scores*" (Best, 2018). The XQ Rubric and Survey is not a competitive ranking scheme, so its users have no reason to game it, and this flaw can be avoided. Indeed, since users of the rubric system know how their explanations will be evaluated, they are free, and even encouraged, to use the XQ Rubric to improve their explanation.

2.4.2 Utilization-Focused Evaluation

Useful Evaluation or Utilization-Focused Evaluation is a technique that employs 'functional evaluation', in which a useful evaluation is deemed one that is created by consulting the opinions of its intended users.

In his theory of Useful Evaluation, Patton (2008) gives four standards for evaluation: utility, feasibility, propriety, and accuracy. The feasibility and accuracy standards mean just that: practicality and alignment with the truth. The utility standard concerns the evaluation's relevance and its intended use, and the propriety standard concerns the ethical and legal correctness of the evaluation. The XQ Rubric and XQ Survey were designed to comply with these evaluation standards.

The XQ Rubric and Survey meet the accuracy standard by accurately representing the judgement of a non-expert on the XAI explanation. They meet the utility standard by serving as a relevant evaluation tool appropriate for its intended use. The XQ Rubric and Survey meet all ethical and legal concerns about evaluation as defined by Patton (2008).

2.5 Broad Overview Of Explanation

2.5.1 Explanation

XAI literature has not yet developed a comprehensive theoretical understanding of at least two of its important elements, explanation and transparency (Doshi-Velez & Kim, 2017). Doshi-Velez and Kim (2017, p. 1) criticised current approaches to XAI, suggesting that most approaches rely on the evaluator's recognising a good explanation upon encountering one.

It is to be hoped that as XAI matures, its literature will include broader and deeper discussion of these and other theoretical matters peculiar to its field. Below is a brief overview of important concepts relating to explanation in the field of XAI.

2.5.2 Explainer

In assessing an explanation, it is important to determine who is offering it. In XAI, the explainer is often assumed to be the machine accounting for its decision or the designer explaining the design (the ML algorithm). As noted in Section 2.3.1, humans are capable of inappropriately ceding authority to an AI and trusting the AI's explanations too readily.

The background and preconceptions of the explainer can influence the effectiveness of the explanation and the trust people place in it (Brown, 2006), including what they think the message is, who the audience is, and what they believe the explanation already shows. If the explainer gets the audience wrong, for example by offering a too-simplistic explanation to experts, the explainer risks losing their audience (Brown, 2006). When, where and how the explainer offers an explanation influences the audience's response (Brown, 2006).

2.5.3 Audience

Depending on who the audience is, what is already known, and what preconceptions have been formed (Brown, 2006), different levels and types of information may be required. The Institute of Electrical and Electronics Engineers (IEEE) identifies five audience types for XAI explanations:

"*users, the general public and bystanders, safety certification agencies, incident/accident investigators and lawyers/expert witnesses*" (Winfield et al., 2021, p. 1). These audience types have different requirements for a good explanation.

Sometimes the people receiving the explanation are not the intended audience, and for that reason they may find it less convincing (Brown, 2006). For example, an explanation targeted at an expert audience may be incomprehensible to non-experts. Since the reaction of the audience will vary depending on the context of the explanation, it is important to consider how, when, and where the audience is being offered the explanation (Brown, 2006). For this reason, an explanation should not be evaluated without considering its intended audience, for the composition and character of the audience will affect how well the explanation is understood and regarded.

2.5.4 Definition of XAI Explanation

What constitutes an explanation, what is a good explanation, and what is a good explanation in XAI? These are vexed questions, and answers vary, with no single definitional solution that can usefully be regarded as final (Rosenfeld & Richardson, 2019, p. 2).

This thesis accepts, within limits, the ordinary view that, at least to begin with, you know a good explanation when you see it: some proposed explanations are better than others and, for reasons that when necessary can be made explicit, this or that explanation is to be preferred over competing explanations.

Without some such preliminary understanding of what makes a good explanation, any attempt to close in on a better one is doomed to failure, for it is impossible to articulate why one definition is better than another without some prior acquaintance with examples of what is at issue. The usually accepted opinions expressed on the matter should be considered and provisionally adopted while the question is analysed and refined.

Model of a Good XAI Explanation

Though there may be no general solution to some of these issues, a model of a good XAI explanation can nevertheless be created by referencing authors in the XAI field and related fields (such as Lester and Porter (1997); Miller (2017); Miller, Howe, and Sonenberg (2017); Sevian and Gonsalves (2008)). It is crucial to create such a model because this thesis, and several 'Evaluation Methodologies' (Section 3.4.2) and 'Evaluation Concepts' (Section 3.4.3), base their evaluations on a such a model of good XAI explanation.

With these general considerations in mind, a good XAI explanation may be defined as a statement or series of statements, sometimes illustrated with graphs or images, that seeks to illuminate, justify, or clarify an action. It is:

a) understandable, logical, clear, and concise (similar to the ideas of 'Writing Style' and Coherence in Lester and Porter (1997) and similar to the view of good explanation put forward by Sevian and Gonsalves (2008))
b) truthful, believable, and given in good faith (similar to the idea of Correctness in Lester and Porter (1997))
c) created by someone with insight into AI generally, or the AI under analysis

This list of features is not intended to be definitive or exhaustive. However, it includes features generally considered part of a good explanation. While the emphasis on a truthful explanation given in good faith is unusual (though with precedent in the work of Lester and Porter (1997)), it is included in the list because without assuming the explanation is true, evaluating its worth to an audience is not useful. An explanation can be used to inform the audience and guide them to make informed decisions only if it is true. Moreover, a deceitful explanation will, if discovered, destroy trust people have placed in it (Hutson, 2021).

An XAI researcher uninterested in truth might concentrate on providing explanations that aim to maximize some desired impact on the audience's behaviour regardless of whether the explanation is true or complete (Hutson, 2021). Consider, for example, an explanation of why an AI recommended a user buy a particular product that fails to acknowledge that the AI's creator receives a commission for sales of the product and is therefore likely to recommend the product over other, possibly better, products. The explanation may benefit the AI's creator, but it will not benefit the audience (Hutson, 2021).

2.6 Conclusion

This chapter discussed the background to a number of important concepts in XAI, such as black boxes, illustrative examples of failures in AI, and an overview of concepts relating to explanation.

An understanding of explanation is vital to understanding XAI. Each part of an explanation is essential: the explainer, the audience, and the explanation itself. The concept of a good explanation is vital to the creation of an effective XAI. A true explanation is one created in good faith which adequately conforms to the facts.

3 LITERATURE REVIEW

3.1 INTRODUCTION

This literature review discusses commonly-used methodologies of explanation-evaluation in the current literature about XAI. Different taxonomies of XAI evaluation are analysed. A new two-dimensional model of XAI evaluation is presented and used to group common approaches to evaluation.

The literature review used Google Scholar to find relevant papers. The search strings were "explainable artificial intelligence" and "evaluate". The keywords were combined in various ways.

The papers discovered by these searches were filtered by the citation count of the article, the publisher (publications of no scholarly standing were discarded), and their relevance to XAI evaluation and explanation. Of particular interest were papers that used an evaluation method or clearly described why none had been employed.

3.2 PREVIOUS ATTEMPTS TO EVALUATE EXPLANATIONS

Little has been published about the evaluation of explanations in the field of XAI. Existing work "*tends to suffer from a lack of usability, practical interpretability and efficacy on real uses*" (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018, p. 1). Moreover, in their survey of XAI papers, Miller et al. (2017) found that researchers in the field of XAI were not building on knowledge from other fields.

Most evaluation of AI systems is experimental; that is, an experiment will show that one AI system performs better than another in tests (Japkowicz & Shah, 2011). For example, a system may produce better classification accuracy. However, this evaluation style does not give priority to explanations and is not immediately useful in the evaluation of XAI explanation generally.

Biran and Cotton (2017) identified two trends of current research in the XAI field: "*Interpretable Models*" (p. 4) and "*Prediction Interpretation and Justification*" (p. 3). "*Interpretable Model*" research describes models that create interpretable AI. For example, Zhang, Wu, and Zhu (2018) use a type of "*Interpretable Model*" that aims to modify convolutional neural networks to render them interpretable. In contrast, "*Prediction Interpretation and Justification*" research attempts to understand and justify the predictions of minimally interpretable AI. The research by Doshi-Velez and Kim (2017) is an example.

Sanneman and Shah (2020) suggest a taxonomy of XAI, ordered by the purpose and situation of an XAI explanation:

> *Level 1: XAI for Perception - explanations of what an AI system did or is doing, and the decisions made by the system*
>
> *Level 2: XAI for Comprehension - explanations of why an AI system acted in a certain way or made a particular decision and what this means in terms of the system's goals*
>
> *Level 3: XAI for Projection - explanations of what an AI system will do next, what it would do in a similar scenario, or what would be required for an alternate outcome*
>
> (Sanneman & Shah, 2020, p. 98)

The framework of Sanneman and Shah (2020) allows researchers to assess which situation the explanation aims to meet and then to assess it appropriately.

3.3 Competing Taxonomies of XAI Evaluation

Towards A Rigorous Science of Interpretable Machine Learning (Doshi-Velez & Kim, 2017) draws attention to concerns about the safety of AI algorithms, arguing that interpretable ML will help to make AI algorithms safer and allow for their optimisation. The paper suggests a methodology

for producing evaluations of interpretable AI and provides a taxonomy of approaches. It outlines evaluation methodologies rather than discussing specific methods in detail.

The taxonomy of evaluation approaches suggested by Doshi-Velez and Kim (2017, pp. 4-5) divides approaches into three types: '*Application-Grounded*', '*Human-Grounded*', and '*Functionality-Grounded*'. The '*Application-Grounded*' approach evaluates AI's interpretability with human experiments and real tasks. The '*Human-Grounded*' approach uses humans to evaluate explanations (this thesis employs a '*Human-Grounded*' approach). The '*Functionality-Grounded*' approach does not use humans in its evaluations. It employs instead a "*formal definition of interpretability as a proxy for explanation quality*" (Doshi-Velez & Kim, 2017, p. 5). That is, in order to label new explanations, it uses ML algorithms that have been trained on explanations that were labelled, presumably by humans, as good or bad. A similar category is defined in the two-dimensional model of XAI evaluation presented by this thesis (Section 3.4.2.2).

The taxonomy of Doshi-Velez and Kim (2017) is broad but cumbersome. The two-dimensional taxonomy suggested below (Section 3.4) gives greater weight to how the evaluation methods could be used in the real world.

Evaluating Arguments Made with Natural Language Generation

Natural Language Generation (NLG) is a sub-field of AI and computational linguistics that aims to create systems able to generate meaningful text from non-linguistic information (Reiter & Dale, 2000). While it is not directly concerned with XAI, there has been some work in the field of NGL on evaluating explanations generated by NLG algorithms.

In their paper on evaluating NLG explanations, Carenini and Moore (2006) discuss three methods of evaluation:

1. Using a panel of human judges to score outputs from an NLG algorithm
2. Using "*corpus-based evaluation*" (Carenini & Moore, 2006, p. 938) to evaluate the input and output of an NLG algorithm,
3. Using the method of "*task efficacy*" (Carenini & Moore, 2006, p. 938) to evaluate human responses to the explanation.

The first method of evaluation, employing a panel of human judges, was used by Lester and Porter (1997). This experiment produced one of the first papers that attempted to evaluate an NLG explanation systematically.

Lester and Porter (1997, pp. 68-69) created an algorithm that produced explanations (the KNIGHT algorithm), following this with an experiment to test if the algorithm's explanation was as good as an explanation created by a panel of experts. A computer then created candidate explanations. Two panels were formed from experts in the field of biology. One panel generated explanations; the other evaluated them using a grading scheme from A (the highest grade) to F (fail).

The KNIGHT algorithm scored within half a grade of the explanations created by the panel of experts. Only the writing style and coherence of the human and machine explanations were given statistically significant different grades. KNIGHT's performance even exceeded that of one of the biologists (Lester & Porter, 1997, p. 90).

The second method of evaluation, "*corpus-based evaluation",* used to evaluate the input and output of an NLG algorithm (Carenini & Moore, 2006), describes a complex algorithmic process of evaluation, best used with a simple algorithm where inputs and outputs can be compared directly.

The third method of evaluation, "*task efficacy*" (Carenini & Moore, 2006), was used to evaluate human responses to the explanation. Unlike the earlier two methods, this pays attention to the

users of the evaluations. Di Eugenio, Glass, and Trolio (2002) used the task efficiency method to evaluate their NLG tutoring tool. They examined how users interacted with it. This was followed by a survey of the users.

3.4 Two-Dimensional Model of XAI Evaluation Employed by this Thesis

This section presents a two-dimensional model of methods to evaluate explanations of XAI decisions. While there are many ways of arranging a taxonomy of explanation, the most suitable group similar methodologies and theories. This two-dimensional model is based on findings in the literature about the evaluation of XAI explanations. Several quite different methodologies and theories were identified. (A small number of papers did not evaluate their explanations at all.)

The taxonomy employed by this thesis has two dimensions. One is Evaluation Methodology; the second is Evaluation Concept. The methodology governs the evaluation of the explanation. The concept of evaluation drives the questions that the evaluation seeks to answer. 'Evaluation Concept' refers to what is being investigated by the evaluation. 'Evaluation Methodology' refers to how the explanation is evaluated. Researchers typically choose similar methodologies but pair them with different concepts of evaluating explanation and vice versa. By separating the evaluation schemes into two: Evaluation Concept and Evaluation Methodology, the themes in XAI evaluation schemes can be interrogated more effectively.

3.4.1 Self-Evident Argument

The category of self-evident argument covers explanations that are not adequately evaluated by the researchers who created them. Some authors assume or assert without evidence that a particular style of explanation will be appropriate. Frequently this method goes no further (Doshi-Velez & Kim, 2017). The self-evident argument approach is common in XAI literature (Doshi-Velez & Kim, 2017), and researchers frequently use this method to explain XAI decisions

(Holzinger, 2017; Lane, Core, Lent, Solomon, & Gomboc, 2006; Malle, 1999; Tangermann, 2018) without justifying their methods or explanations. This prevents comparison between systems.

The related concept of "*you will know it when you see it*" is discussed by Doshi-Velez and Kim (2017, p. 1) in their taxonomy paper. They appear unconcerned about the lack of rigour of this approach, noting that "*the notions of interpretability […] appear reasonable because they are reasonable*" (Doshi-Velez & Kim, 2017, p. 1). Nevertheless, Doshi-Velez and Kim (2017) also argue for a more rigorous scheme for evidence-based evaluation in Interpretable Machine Learning. They also believe that XAI needs more formal approaches to evaluation, not merely "*you will know it when you see it*".

In 2018 IBM produced a research paper which outlines a new proposal for explainable AI and trustworthy AI (Hind et al., 2018). Elizalde, Sucar, Noguez, and Reyes (2009) compared their explanations to a human expert in the same field. They did not effectively validate their comparisons, however, relying instead on their own judgement as to whether the explanation was reasonable.

The method of self-evident evaluation is hard to replicate, and since it is not supported by any system of standard assessment, it is vulnerable to the evaluator's biases. However, the 'self-evident' methodology is cheap and quick, and at least to the evaluator, the quality of the explanation seems obvious.

3.4.2 Evaluation Methodology

'Evaluation Methodology' refers to the techniques used by a researcher to evaluate an XAI explanation. It is used without reference to the underpinning theoretical understanding of explanation (which comes under the heading of Evaluation Concept). The term Evaluation

Methodology is used to group similar methods of evaluating XAI, making it possible for trends in the literature to be identified and discussed.

3.4.2.1 Surveying People (Questionnaires, Focus Groups, Surveys)

Some researchers have used surveys to evaluate explanations of AI decisions. Thellman, Silvervarg, and Ziemke (2017), for example, conducted surveys to gauge how people explained the behaviour of both robots and humans.

User studies are another way of surveying people. Abdul, von der Weth, Kankanhalli, and Lim (2020) employed a user study to assess the cognitive load of ML modelling explanations. Wang et al. (2021) employed a user study to assess user-perception of explanations.

For the taxonomy presented here, since they constitute a method of collecting responses about how good an explanation is, surveys are regarded as an Evaluation Methodology. A survey by itself is not an Evaluation Concept because it does not specify how to evaluate an explanation nor suggest what questions should be asked.

Surveying non-experts is cheaper than surveying experts, for expertise is expensive (Sternberg & Frensch, 1992). Surveys can be created and conducted more quickly and cheaply than rubrics or ML systems (Krosnick, 1999). A survey methodology should explain who is being surveyed, how many people are being surveyed, and the software used (Kotrlik & Higgins, 2001). Surveying people does not require an expert to manage the survey, the focus group, and the questionnaire.

While surveys and questionnaires can be conducted relatively quickly, they take time to set up, conduct and interpret. Moreover, although the evaluators do not need to be experts to conduct a survey, they need to be familiar with the survey process and need to know how to ask pertinent questions (Fowler Jr, 2013). A survey that uses leading or poorly-worded questions risks having illegitimate or weak conclusions being drawn from its work (Kotrlik & Higgins, 2001).

3.4.2.2 Model of Good Explanation

Models of Good Explanation form a subcategory of the 'Evaluation Methodology' category. This subcategory groups together methods of evaluating explanations that start with an already-established idea of a good explanation.

While surveys may assume, or arbitrarily define, what is to be counted as a good explanation, rubrics and ML evaluation methods rely on pre-established and carefully articulated notions of good explanation. The success of the evaluation depends upon the suitability of the model; a flawed or incomplete model leads to a weak or incomplete evaluation.

Methodologies included in this category have the advantage of having clear goals defined before starting, with clear definitions on which to base the evaluation. A good example of a scheme for the evaluation of XAI that has clearly-defined goals is that suggested by Kuwajima and Ishikawa (2019). It is modelled on ISO standard ISO/IEC 25000, also known as SQuaRE (System and Software Quality Requirements and Evaluation). Their evaluation scheme uses the ISO standard as a model of a good explanation.

Rubric

A rubric "*lists the criteria for a piece of work, or 'what counts' (for example, purpose, organisation, details, voice, and mechanics are often what count in a piece of writing); it also articulates gradations of quality for each criterion, from excellent to poor*" (Andrade, 1997).

Rubrics offer many advantages here, including an established methodology. Rubrics can provide usable data and feedback and are readily adaptable to new purposes (Lizotte et al., 2003). They are used in many kinds of evaluations, from scientific abstracts to student assignments. Rubrics can be easily modified to fit the needs of a researcher evaluating XAI explanations. Of course, a rubric cannot necessarily be adapted to cover every explanation and may neglect some of a explanation's unique characteristics (Moskal & Leydens, 2000).

Rubric schemes are not without flaws. They sometimes rely too heavily on the opinion of the marker, and marking is difficult to standardise across a group of people too different in their assessments. A rubric is usually more reliable as an assessment method when it is coupled with another, different system of criteria, for there is a danger that a rubric by itself will fail to include relevant aspects of the matter being investigated.

ML Evaluation of XAI Explanations

ML approaches to evaluation often simply amount to the use of ML algorithms to evaluate human actions. Typically an algorithm is used to set a benchmark performance against which the human is compared, or the results of one ML algorithm are compared with those of another (Japkowicz & Shah, 2011). Currently there is no standard way to evaluate explanations by this method.

An ML model used to evaluate an XAI explanation would first require a dataset of explanations that have a score or a label attached, such as ‘good explanation’ or ‘bad explanation’. The ML model could then be trained on this data, and the model used to evaluate subsequent explanations.

The third category of evaluation proposed by Doshi-Velez and Kim (2017) is very close to this category of ML Evaluation of XAI Explanations. Although Doshi-Velez and Kim (2017) support the use of ML to evaluate explanations, they do not clearly set out how this should be done. Doshi-Velez and Kim (2017) themselves do not offer any examples of ML algorithms being used to evaluate explanations. However, ML-based evaluation approaches may be at least potentially quicker and cheaper than other methods of evaluation (excluding the 'Self-Evident Argument' approach (Section 3.4.1)).

Evaluating XAI explanations using ML technology can be quick, repeatable, and objective. It is also cheaper if equipment for the task is already available. On the other hand, it is costly and time-consuming to set up an ML evaluation, for there is no established ready-to-use

methodology, and an ML specialist will be needed, both to perform the initial setup and, later, to conduct the evaluation.

3.4.3 Evaluation Concept

Evaluation Concept refers to the evaluation concept being measured, that is, the concept that underlies the methodology of the evaluation. While Evaluation Methodology refers to the practical considerations of evaluating an XAI explanation, 'Evaluation Concept' groups similar conceptual views of it.

Evaluation Concept shapes the methodology employed and directs the overall assessment of a target explanation. Naturally, a researcher should choose a type of evaluation appropriate to the research being undertaken. The researcher should not only consider what is to be explained but also what features a good explanation should have. A concept that fits well with the evaluation methodology will allow the researcher to recognise and understand the features of a good explanation.

3.4.3.1 Audience Behaviour

To some degree, the efficacy of an explanation depends on the interests and expectations of its audience. Generally speaking, an explanation should be tailored to suit the people at whom it is directed. In this regard, a good explanation is an explanation judged so by its audience.

Trust

Ribeiro et al. (2016) link explanation to trust, arguing that an unexplained, or poorly explained, AI will not be trusted as much as a well-explained AI. A sense of how good an explanation is can therefore be inferred from how trusted an AI is and vice versa. Since trustworthiness is dispositional, people's trust in an explanation can be deduced by aspects of their behaviour or calculated directly by asking whether they trust the explanation and the system that produced it (Lipton, 2018).

A significant advantage of linking trust to good explanation is that it is possible to determine the effectiveness of an explanation experimentally by conducting human trials.

Ribeiro et al. (2016) used trust to evaluate whether their explanation system for ML worked effectively. They asked users to pick the better algorithm, using explanations they supplied. Users were selected according to their knowledge of the algorithm's subject, not their knowledge of ML.

Understanding

One reason for creating an XAI system is to increase understanding of the reasons for an AI's decisions. This increases our ability to predict from the explanation what the algorithm will choose to do next. A good explanation gives the audience insight into the machine's actions, making its behaviour easier to predict (van der Waa, Nieuwburg, Cremers, & Neerincx, 2021).

van der Waa et al. (2021) compared two explanation styles and tested how they affected the audience's understanding of the explanation. They rated the explanations on their *"system understanding (Experiment I), persuasive power and task performance (Experiment II)"* (van der Waa et al., 2021, p. 4). The results from the van der Waa et al. (2021) experiment align with the two-dimensional model of Concept and Methodology of this thesis.

Conclusion

It is sometimes held that the reactions of an audience to an explanation can be used as a measure of the explanation's quality. The evaluation methodologies that emerge from attention to the audience are typically more expensive in time and resources than other methods, but are more objective and consistent. Attention to audience behaviour helps to make sure the that the intended audience will understand the explanation.

3.4.3.2 Expert Analysis

The decisions of an ML algorithm can be evaluated by a panel of experts. In fact, it is very common to have an expert or panel of experts evaluate the quality of a piece of research presented to it. In academic publishing, for example, the peer-review process uses this method to select articles for inclusion in professional journals.

Expert analysis is commonly paired with the methodology of surveys, ordinary focus groups, and focus groups that use the Delphi methodology. While this is an effective method for evaluating explanations, it is costly and time-consuming (Laidlaw, 2014). Expert panels are helpful when various perspectives are sought or when expert input is needed to make a judgement (Laidlaw, 2014). Expert panels are best used when experts in a field need to discuss a specific topic or problem. The authors of an explanation of an AI's decisions can be claimed to be experts, as can the creators of the AI whose decisions need explaining.

As discussed in Section 3.3, Lester and Porter (1997) created a method to evaluate ML-generated explanations using a panel of experts. Each explanation was evaluated and given a letter grade (from A to F). Two expert panels were used to verify the results. Though expensive in human resources, this is a thorough way of evaluating an explanation (Laidlaw, 2014).

It is costly and time-consuming to use an expert panel, however, and this makes the method unpopular. Another disadvantage of expert panels is that biases can be introduced into their judgements, as Langfeldt (2004) explains:

> *"A situation with no clear bases or rules for peer judgements means there is a wide scope of 'acceptable' outcomes of evaluation, and also various kinds of bias"* (Langfeldt, 2004, p. 57)

As with the method of evaluation using audience behaviour, evaluation using expert panels is expensive and slow (Laidlaw, 2014). Expert panels can be helpful; they may identify errors in the

reasoning of the explanations, and should the audience be composed of experts, they will be able to assess whether the explanation is appropriate. However, an expert panel may be less well able to judge whether the explanation is suited to a lay audience.

Expert analysis allows people who are more likely to know what they expect from an explanation to have a say, and this gives the evaluation credibility. Experts can also judge on the basis of their experience the truthfulness of the proffered explanation.

3.4.4 Desirable Properties of Evaluations

Two comparison tables, Evaluation Methodologies (Table 3-1) and one Evaluation Theories (Table 3-2) are provided below. These set out the factors which may affect the choice of a method to evaluate an explanation. '*Self-evident argument*' is included in both tables as it is both an Evaluation Methodology and Evaluation Concept.

To compare the Evaluation Methodologies and Evaluation Concepts, four categories were chosen: '*Low Cost*' and *Quick'*, '*Reproducible'* and '*Standard Across Evaluators*'. These comparison categories were divided into two groups: '*Ease of Use*' and '*Allows for Comparison*'.

Ease of Use: A method that is easy to use will take less time and cost less, making it more likely to be adopted.

> **Low cost**: The proposed method does not require additional funds or equipment that would not usually be available to researchers. This is a desirable quality because it means that a broad range of people can use the method.
>
> **Quick**: The proposed method is relatively quick. This is a desirable quality because it means that the results from the method can be returned quickly, making it easier for multiple explanations to be evaluated.

Allows for Comparison: This is a group of categories that allow comparison. They permit researchers to compare the results of evaluations of different explanations.

> **Reproducible**: The proposed method can be repeated with the same or similar results. Reproducibility is important because it allows comparison between results (Moskal & Leydens, 2000).
>
> **Standard Across Evaluators**: The proposed method will return similar opinions. That is, evaluators will not use the evaluation method differently (Moskal & Leydens, 2000). As with reproducibility, this is important because it allows comparison between results. It also reduces a potential cause of bias, the particular and special concerns and interests of individual evaluators.

3.4.4.1 Evaluation Methodology Comparison Table

Table 3-1 compares different Evaluation Methodologies. While methodologies have their own strengths and weaknesses, the advantages of the self-evident argument (discussed below in Section 3.4.1) does not outweigh its many disadvantages.

Table 3-1 Evaluation Methodology' Comparison Table

	Self-Evident Argument	**Surveys**	**Evaluation using a definition of a good explanation**	
			Rubrics	ML-based evaluation
Easy to Use	Yes	Yes*	Yes	No
Low Cost	Yes	No	Yes*	Yes*
Quick	Yes	No	Yes*	No
Allows for Comparison	No	Yes	Yes	Yes
Reproducible	No	Yes	Yes	Yes
Standard Across Evaluators	No	Yes	Yes*	Yes

*Dependent on the creator

The self-evident argument (Section 3.4.1) scheme of explanation-evaluation is the least robust. Although it is low cost, quick, and easy to use, it is not reproducible and does not allow for

comparison (Doshi-Velez & Kim, 2017). The self-evident argument does not account for a variety of opinions, it does not match to what a different method would find, and different evaluators will not necessarily return similar results.

By contrast, the survey methodology (Section 3.4.2.1) is reproducible, and it allows for comparison (Fowler Jr, 2013). Surveys are often easy to use, though this, of course, depends on the skill of the survey's designer (Krosnick, 1999). Surveys allow the evaluators of an explanation to express a variety of opinions. Survey results are dependent heavily on how the survey is conducted and its results assessed (Kotrlik & Higgins, 2001).

The broader category of '*Evaluation using a model of good explanation*' (Section 3.4.2.2) describes evaluation methods that rely on a good explanation model. This category is further divided into two: Rubrics and ML-based evaluation.

Rubrics are the best 'Evaluation Methodology' overall as scored by Table 3-1 (though this naturally depends on the rubric's design). Rubrics are quick, easy to use, and cost relatively little (Turley & Gallagher, 2008). They allow for comparison, and are reproducible (Moskal & Leydens, 2000). They permit a variety of opinions to be collected about the explanation (Lizotte et al., 2003).

In contrast, an ML-based evaluation scheme is not quick and easy to use, and it does not allow a variety of opinions to be collected. If the creators of an ML evaluation have the necessary equipment, it will be comparatively cheap. On the other hand, an ML-based evaluation scheme is reproducible, allows for the comparison of different explanations, and is standard across evaluators (Doshi-Velez & Kim, 2017).

*3.4.4.2 'Evaluation **Concept**' Comparison Table*

Table 3-2 does not indicate how long a concept may take to implement. How much the concept might cost to establish is also not considered because this depends on the methods ('Evaluation Methodology') used. While different theories have different strengths and weaknesses, the advantages of the Self-Evident Argument evaluation concept do not outweigh its many disadvantages (Doshi-Velez & Kim, 2017) (discussed in Section 3.4.1).

Table 3-2 'Evaluation concept' Comparison Table

	Self-Evident Argument	**Audience Behaviour**	**Expert Panels**
Easy to use	Yes	No	Yes*
Reproducible	No	Yes	Yes*
Standard Across Evaluators	No	Yes	Yes*
Allows for comparison	No	Yes	Yes*

*Dependent on the creator

The Audience Behaviour category describes an evaluation concept, common in XAI research, that ties the worth of an XAI explanation to its effect on an audience (see, for example, the work of Lipton (2018)). The Audience Behaviour category and the Expert Panels category are conceptually superior to the Self-Evident Argument category, for they can be reproduced and compared. They are also not dependent on the creators of the evaluation, and they allow for a variety of opinions.

The Expert Panels evaluation concept uses experts to evaluate the XAI explanation. However, finding suitable experts for a panel can be time-consuming and costly (Sternberg & Frensch, 1992). Moreover, since experts may offer what could turn out to be simply their own opinions, it is possibly less objective than other evaluation concepts (Langfeldt, 2004). How the evaluation of the XAI explanation is conducted determines whether the results will be uniform across evaluators and whether the evaluations are reproducible. This level of agreement cannot be assumed.

3.4.5 Summary

The two-dimensional model of methods to evaluate XAI explanation discussed above does not rate Evaluation Methodologies and Evaluation Concepts. Different methodologies and concepts suit different purposes and styles of explanation (Sandelowski, 2000). The model simply groups together similar methods and concepts of evaluation.

One of the most common views of XAI evaluation is the 'Self-Evident Argument', or the "you'll know it when you see it" approach. More systematic methods of XAI evaluation are beginning to emerge, however, especially methods of surveying people and of assessing the response of the explanation's audience to gauge the explanation's usefulness.

3.5 CONCLUSION

To discuss trends in the current literature about evaluating XAI explanations, this section introduced a two-dimensional model of XAI evaluation. Each method of evaluating XAI in this chapter was composed of at least one Evaluation Methodology and an Evaluation Concept. Evaluation Methodology refers to the tools and techniques used to evaluate the XAI explanation. Evaluation Concept refers to the underpinning theory of explanation used in the corresponding methodology.

The literature on XAI explanation has two major gaps:

1. there is no formal method for eliciting non-expert evaluations of XAI explanations
2. there is no evaluation scheme for non-experts to use to evaluate XAI explanations

The first gap is evident by the lack of interest in the literature of XAI in non-expert evaluation. The second gap is evident from the lack of methods and schemes designed to cater to non-expert audiences.

4 METHODOLOGY OF THE RESEARCH

This chapter gives an overview of the thesis' research methodology. It also briefly discusses the experimental tools used, evaluation approaches, and case studies.

Design Science and User-Centred Design (UCD) are well suited to the task of creating a "*rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations*" (See Section 1.2) for these methodologies specialise in the creation of an artefact (in this case, a scheme for evaluation) and tailoring it to suit the user (Johannesson & Perjons, 2014).

4.1 DESIGN SCIENCE

The research methodology used is that of 'Design Science'. Design Science is a multi-disciplinary tool for creating objects and processes to solve unsolved and significant problems (Carstensen & Bernhard, 2019). These artefacts can be found in many different contexts. All have a practical purpose (Johannesson & Perjons, 2014). In this thesis, the artefact is a marking rubric. Its purpose is to help measure the efficacy of an XAI explanation.

The Methodology from Peffers et al. (2007)

Since it was adjusted and modified to meet the requirements of Information Systems Research, the Design Science Research Methodology (DSRM) of Peffers et al. (2007) was adapted for use in this thesis. The DRSM was specifically designed to be used in the creation of an artefact similar to that developed by the thesis. It required relatively little adaptation.

The DSRM is a straightforward methodology (Peffers et al., 2007). For this thesis, it had these steps:

1. define the specific research problem and justify the value of a solution
2. define the objectives for a solution
3. design and develop
4. observe and measure how well the artefact supports a solution to the problem
5. demonstrate.
6. evaluate

Stages 1-3 of the DSRM (Peffers et al., 2007) were performed by the Literature Survey and the focus group experiments. These stages concerned the design and development of the eXplanation Quality Rubric (XQ Rubric).

Stages 4-6 of the DSRM (Peffers et al., 2007) were the first, second, and third MTurk experiments, which evaluated the XQ rubric's utility, quality, and efficacy. In this way, an iterative, user-centred development was achieved which covered all the steps of the DSRM.

4.2 Iterative Design

'Iterative design' is a methodology that makes iterative improvements until the desired result is achieved (Ishii, Kobayashi, & Arita, 1994). An iterative approach employing Design Science methodology is appropriate for applied research, especially research with a pragmatic approach and qualitative measures of success (Nielsen, 1993). Figure 4-1 (below), adapted from the DSRM (Peffers et al., 2007) as described by Hevner and Chatterjee (2010, pp. 28 - 31), illustrates the iterative design theory used in this thesis.

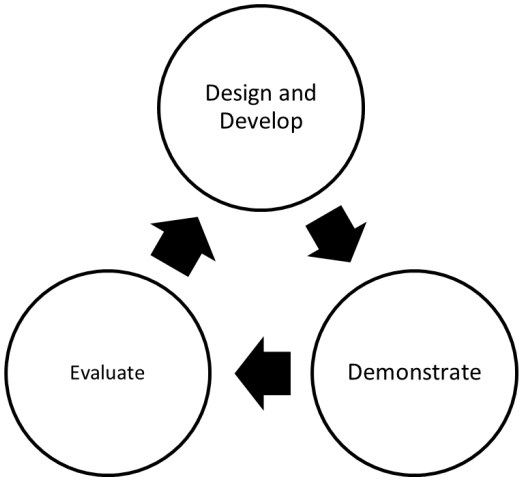

Figure 4-1 Iterative Design Methodology (adapted from Hevner and Chatterjee (2010))

The experiments of this thesis were structured in accordance with design science methodology. Each of the experiments had an iterative cycle of parts, and the thesis used an iterative approach. The iterative approach to design science derives from the same methodology, the DSRM (Peffers et al., 2007).

4.3 User-Centred Design

User-Centred Design (UCD) is a design methodology. It was used in this thesis to augment Design Science methodology. UCD is based on *'International Standard 13407 – Human-Centred Design For Interactive Systems'* (Abras, Maloney-Krichmar, & Preece, 2004). It specially suits the process of the design of products for people who will use them. Like the Design Science methodology, UCD is employed to make the designed object the most appropriate product for the user.

User-Centred Design (UCD) was used throughout this thesis to make sure that users of the XQ Rubric were consulted in its design and development phase. All of the online MTurk experiments had users who were representative of the end-users of the XQ Rubric and Survey. This design methodology suited the research project, for the XQ rubric was designed for use by a user, not for the benefit of its designer.

4.4 Tools Used in This Thesis

This section discusses the online applications used in this thesis to perform its experiments. These tools were selected according to their cost, reliability, and ease of use.

4.4.1 LimeSurvey

LimeSurvey is a popular free and open-source survey software (LimeSurvey, 2021), available to students and staff of Federation University Australia. The survey software had helpful features such as the algorithm function used to create completion codes, an online interface, and the ability to use a single template for multiple surveys.

4.4.2 Google Drive

Google Drive is an online file storage service owned by Google (Google, 2022). Storage up to 15 GB is free for individuals. This service was chosen because it was reliable and cost nothing. It was found to be necessary to use Google Drive rather than the attachment system LimeSurvey provided, which could not reliably display images and documents in XQ Surveys (see Section 6.4.4).

4.4.3 Amazon Mechanical Turk

Amazon Mechanical Turk (MTurk), an online "crowdsourcing marketplace" (Amazon Web Services, 2022), is a popular platform for conducting extensive surveys (Paolacci, Chandler, & Ipeirotis, 2010). MTurk was chosen because there is a significant body of research literature concerning its use in surveys. Moreover, MTurk was able to return a large number of responses in a short time. Ribeiro et al. (2016) used MTurk to evaluate their XAI explanations because it was an easy-to-use recruiting method for gathering responses from people of different cultural backgrounds. Paolacci et al. (2010) draw attention to MTurk's ability to gather respondents from a wide variety of cultural circumstances.

Initially, the survey task could be performed by any MTurk worker. Later, however, because unscreened participants appeared more likely to give nonsensical and irrelevant answers, only MTurk participants rated as “Masters” were given access to the XQ Survey. To be awarded a rating of “Master”, MTurk participants were required to have been highly rated by a broad number of surveyors in a variety of subjects.

MTurk very quickly recruited a suitable number of participants for the surveys. A shortcoming of the MTurk system is that participants have few incentives to answer questions honestly. Some of them, it appears, use automatic answering tools. However, the risk of this was diminished by paying the participants a fair wage, explaining the importance of the research, and checking that all participants read and understood the case studies and questions (Paolacci et al., 2010).

Hendrickson, Navarro, Langsford, Kennedy, and Perfors (2015) argue that the most ethically proper way to treat MTurk workers is as short-term employees and pay at least the equivalent of the US Federal minimum wage (US$7.25/hr when the surveys were conducted (United States Department Of Labor, 2022)).

Completion Code

A completion code, a validation code displayed at the end of the XQ Survey, prevented participants who had not finished the tasks from claiming that they had (G. Paolacci & Chandler, 2014). Participants were required to enter the code on the MTurk website. Using a code such as this is standard practice (Amazon Web Services, 2020).

The LimeSurvey survey platform did not provide an automatic way of giving participants a completion code. It used an equation function instead. The code generated by the function was a number, based partly on the participant’s answers to previous questions. This was used to make sure that the codes were not generalised, although there remained a chance that more than one participant would be given the same code.

The equation in JavaScript:

$$sum\ (rawscoreIBM.shown, Age.shown, 38) * 887 = code$$

The variable rawscoreIBM.shown is the score (out of 10) given to the IBM case study (Case Study 1). The variable Age.shown (the participant's age) and the number 38 (the number was chosen arbitrarily) were added to rawscoreIBM.shown. This sum was then multiplied by the prime number 887. This meant that the code could be easily verified, if it was divisible by 887, it was legitimate.

4.5 The Evaluation Approach Used in This Thesis

This section explains the evaluation approach employed by this thesis, an approach in keeping with the two-dimensional model of XAI evaluation outlined in Section 3.4. Although the XQ Rubric and the XQ Survey were iteratively refined, the evaluation approach behind the iterations remained constant.

4.5.1 Evaluation Methodology: Rubric and 'Surveying People'

The Evaluation Methodology was inspired by two different forms of evaluation methodology. The first was a rubric methodology (Section 3.4.2.2) which underlies the XQ Rubric. The second was a survey methodology. This underlies the XQ Survey (Section 3.4.2.1).

Rubrics are a well-established tool for evaluating human explanations (Sevian & Gonsalves, 2008). The design of the XQ Rubric and its evaluation approaches were developed for and applied to actual explanations generated by AI systems and so can be used confidently in practice.

4.5.1.1 Rubric

The XQ Rubric of this thesis was partly inspired by the rubric used by Sevian and Gonsalves (2008) to evaluate graduate students' explanations of their scientific research, partly by the

results of the literature review (Chapter 3), and partly by the model of 'explanation' which I adopted (Section 2.5.4).

The Sevian and Gonsalves (2008) rubric was designed to evaluate graduate students' explanations of their scientific research. However, it is also both appropriate and generalizable for evaluating the effectiveness of an XAI explanation.

4.5.1.2 Survey

A survey methodology was used to supplement that of the rubric. The methodology recommended by Paolacci et al. (2010) for use with MTurk formed the basis of the XQ Survey designed by this thesis. A survey is a cheap and quick way of gathering a large number of opinions from a variety of people, and so suited my purposes.

4.5.2 Evaluation Concept: Audience Behaviour and Expert Analysis

The Evaluation Concepts employed in this thesis are Audience Behaviour and Expert Analysis (Section 3.4.3). The use of Audience Behaviour was inspired by Ribeiro et al. (2016) and their use of MTurk. However, unlike Ribeiro et al. (2016), this thesis uses members of the explanation's intended audience to assess an explanation (using an Expert Analysis Evaluation Concept). For this, the thesis asks, "How would someone like you view the explanation?"

4.6 Case Studies Used in This Research

The case studies used in this thesis (excluding those developed by Dr Cruz) were found using Google Scholar to search for the strings "Explainable Artificial Intelligence", "XAI", "Case Study", and "Example" in various combinations and permutations (a technique similar that of Vilone and Longo (2021a)). The case studies were drawn from a variety of sources, including journal articles, a Master's thesis, an Ombudsman's Report, and a winning competition entry (the IBM loan case study).

Papers with examples of suggested explanations were shortlisted for further examination. Those that relied completely on mathematical explanations or required expert background knowledge to understand were discarded. The explanations used in the case studies were selected against these criteria: they had to be short or able to be shortened, and they had to be pitched at an appropriate level for the focus group participants. These requirements excluded very technical explanations. With ethical survey considerations in mind, explanations of unpleasant subjects such as warfare and criminal matters were also excluded. The explanations had to be reasonably well-presented, for there was little point in giving the focus group participants a case study so poorly set out as to be incomprehensible.

Since the explanations of the case studies have never been subject to a transparent and public evaluation, citations and critical reactions to the case study were used as a proxy to measure critical evaluation of the explanations. Google Scholar was used to find the citation count of the articles from which these case studies were drawn. Articles which cited the case study articles were read to calibrate the critical response to the case studies. The critical response was compared with the response of the reviewers who used the XQ Rubric and Survey.

Six explanations from the remaining shortlisted papers were selected and became case studies, as outlined below. Full details of each case study are given in Appendix C.

4.6.1 Case Study 1 – The Drone Case Study

Case Study 1 (Appendix C1a) describes a situation in which an after-action review (AAR) board used an XAI algorithm to create an explanation of why an Unmanned Aerial Vehicle (UAV) deviated from its predefined path (Keneni, 2018). An algorithm then creates a reverse model to explain the decision (Keneni, 2018, p. 4). The seven inputs into the model are "*time, x-coordinate, y-coordinate, heading direction, engage in attack, continue mission, and steer UAV*" (Keneni, 2018, p. 4). The two outputs from the model are "*weather conditions and distance from the enemy"* (Keneni, 2018, p. 4).

The wording and graphics from Case Study 1 came from a Master of Engineering thesis by Blen M. Keneni (Keneni, 2018). A follow-up paper was used for more detailed information (Keneni et al., 2019).

Keneni's Master of Engineering thesis (Keneni, 2018) had not been cited at all (21 February 2022). However, a follow-up paper co-authored by Keneni et al. (2019) was cited 35 times at 21 February 2022. Papers that mention the follow-up paper merely note that it exists and do not review it critically. However, since the thesis was accepted, it is reasonable to assume it was considered to be of an adequate standard in the field.

4.6.2 Case Study 2 – The Smart Lego Factory Case Study

The wording and graphics of Case Study 2 (Appendix C1a) came from a paper by Rehse, Mehdiyev, and Fettke (2019). The paper has been cited 39 times at 21 February 2022.

Case Study 2 describes a situation where a manager wished to know what problems might arise in a manufacturing plant and their likelihood of occurrence. The potential problems were identified by an XAI algorithm and presented in a dashboard layout.

Galanti, Coma-Puig, de Leoni, Carmona, and Navarin (2020) criticise this paper for having only one process and five activities: "*The most relevant work is by Rehse et al., which also aims at providing a dashboard to process participants with predictions and their explanation. However, the paper does not provide sufficient details on the actual usage of the explainable-AI literature, and the very preliminary evaluation is based on one single artificial process that consists of a sequence of five activities*" (p. 2). However, Galanti et al. (2020) do not comment on other aspects of the explanation, ignoring the way in which problems and their likelihood of occurrence were communicated to the managers of the manufacturing plant.

4.6.3 Case Study 3 – The Clinical Decision Support Case Study

Case Study 3 (Appendix C1a) outlined an explainable clinical-decision-support visualisation. These visualisations were created to help remove cognitive biases, assisting clinicians to make reliable decisions (Wang et al., 2019). This paper is mentioned in an influential study by Gunning et al. (2019). However, it offered no critical comments on the paper by Wang et al. (2019).

Buçinca, Lin, Gajos, and Glassman (2020); Gade, Geyik, Kenthapadi, Mithal, and Taly (2019); Mohseni, Zarei, and Ragan (2021) all acknowledge this paper. They do not, however, offer any critical comment. This case study is included here, for it has a high citation count (299 at the time of writing, 22 February 2022) and represents an important class of explanation in the field of medical AI.

4.6.4 Case Study 4 – AI-driven medical diagnosis tool

Case Study 4 (Appendix C1a) describes a situation where a clinician requested the reasoning behind an XAI-suggested treatment for a patient's diagnosed condition (Lamy, Sekar, Guezennec, Bouaud, & Séroussi, 2019). It has 145 citations at the time of writing (23 February 2022). Tjoa and Guan (2020) mention this paper in an aside about XAI that employs Case-Based Reasoning (CBR). The design of the explanation is not criticised, and since this paper is so widely cited, it can be concluded that it is also considered to be work of a high standard.

4.6.5 Case Study 5 – The IBM Loan Case Study

Case Study 5 (Appendix C1a) presents the explanation a bank customer received when he asked to be told the reasoning for the bank's rejection of his loan application. The case study was created using the AI Explainability 360 toolkit from IBM (Arya et al., 2020). On 31 January 2019, it was announced that IBM's entry into the FICO (Fair Isaac Corporation) Explainable Machine Learning (XML) Challenge (2019) had won (Jawski, 2019). The FICO XML Challenge (2019) was sponsored by FICO, Google, Imperial College London, MIT, University of Oxford, UC Irvine and UC Berkeley (FICO Community, 2019).

The organisers of the FICO XML Challenge (2019) team explained that:

> *The winning team received the highest score in an empirical evaluation method that considered how useful explanations are for a data scientist with the domain knowledge in the absence of model prediction, as well as how long it takes for such a data scientist to go through the explanations.* (Jawski, 2019)

One aspect of the scoring was "*how much* [the] *explanations help two tasks that data scientists at FICO may routinely conduct*" (FICO Community, 2019). However, there was no documentation of the empirical evaluation method mentioned in the challenge outline (FICO Community, 2019), nor were any scoring sheets made public.

This case study's win in a 2019 XAI competition seems to show that it is a very well-regarded explanation of an AI's decision.

4.6.6 Case Study 6 – The Australian Centrelink 'RoboDebt' Case Study

In July 2016, Centrelink, an Australian Government welfare agency, issued letters to certain welfare recipients informing them that they owed money to the welfare program (Commonwealth Ombudsman, 2017). Centrelink used an algorithm to find welfare recipients deemed to have been overpaid (Commonwealth Ombudsman, 2017). The letters advised recipients that they were suspected of under-reporting, or mistakenly reporting, their income to Centrelink. It asserted that they had been overpaid and consequently owed money to Centrelink (Commonwealth Ombudsman, 2017). These letters and the associated debt recovery became known as the "*RoboDebt*" scandal (Carney, 2019).

The Centrelink letters attempted to explain to recipients of the letter that they owed money and offered advice, if the allegation was mistaken, about how to prove that they had no debt. An example letter, taken from the Commonwealth Ombudsman's report (Commonwealth Ombudsman, 2017), was used to create Case Study 6 (Appendix C1c).

Most people who received the letter were unable to navigate the Centrelink bureaucracy to dispute the debt. People who were unable to dispute the debt were forced to pay. This caused unnecessary pain and suffering (Commonwealth Ombudsman, 2017). Public outcry led to an investigation by the Commonwealth Ombudsman. The "RoboDebt" letter was used as a case study for this thesis because it was widely commentated on by the Ombudsman and the media. The entire RoboDebt letter was evaluated because it offers a full explanation of why the Centrelink customer was found to be in debt to Centrelink, an explanation intended for a non-expert audience.

4.7 EXPERIMENTS

The experiments of this thesis follow an iterative version of the methodology outlined by Peffers et al. (2007). Each experiment went through at least one cycle of Design and Development, Demonstration, and Evaluation (Peffers et al., 2007).

4.7.1 The Focus Group Experiment

This literature review of this thesis was designed to take the first two steps of Peffers et al. (2007), that is, to investigate the research problem and define a solution. The review found that there were few suitable XAI evaluation methodologies. It discovered no evaluation methodology tailored to XAI explanations for non-experts.

The literature review findings were used to create a draft XQ Rubric (Chapter 5.3). This was presented to an online email-based focus group using the Delphi methodology of Skulmoski, Hartman, and Krahn (2007). This methodology is iterative, designed to collect judgements from experts about a product or service. It is different in four ways from other methodologies that gather feedback from groups of participants in that it has: "*anonymity, iteration with controlled feedback, statistical group response, and expert input*" (Goodman, 1987, p. 1).

The Delphi methodology uses a three-stage process of collecting expert judgements from participants. These are then merged, analysed, and returned to the participants for their feedback (Skulmoski et al., 2007).

Because it was less vulnerable to being heavily influenced by a single person, for these tasks the Delphi methodology for focus groups was better suited than similar focus group methodologies. The Delphi methodology is also well suited to long-distance collaboration.

Delphi methodology was used to refine the XQ rubric in a three-stage iterative process. Feedback from the focus group was divided into three types, concerning the quality of the presentation, the text, and the rubric. The XQ rubric was initially presented in a paper format for participants to complete. This was later changed to an online survey. The text and rubric quality were improved from stage to stage. In particular, participants gave valuable feedback about the clarity of rubric categories and gradations.

The original project plan was to follow up the online focus groups with an in-person focus group. Because of COVID-19 restrictions this could not be done.

4.7.2 MTurk Experiments Using the XQ Survey

The next step was to evaluate the effectiveness of the XQ Rubric and validate its use. The online XQ Rubric and the methodology created using focus group research were used to create an online survey designed to evaluate several case studies using a broader audience. This survey became known as the eXplanation Quality Survey (XQ Survey). It was created with the LimeSurvey platform.

The experiments in this stage of the thesis were known collectively as the “MTurk Experiments”. Because they used the same version of the XQ Rubric, and the Amazon Mechanical Turk survey platform was used, they were grouped together.

The XQ Surveys in this section had three parts: the unguided score, the XQ Rubric, and a demographics section. Only the first and third parts were changed, when necessary, to accommodate the case study. The second section did not vary.

The first section of the XQ Survey, Unguided Score, was composed of comprehension questions designed to establish whether the participants had read the case study. The first section also asked participants to rate the explanation on a scale of 0 to 10. In later iterations of this section, participants were asked to justify their rating.

The second section of the XQ Survey was the eXplanation Quality Rubric (XQ Rubric). This section was unchanged from the XQ Rubric finalised with the help of the focus group.

The third section of the XQ Survey, the demographics section, concerned the participants themselves. It was designed to elicit information about a participant's age, location (country), occupation, and knowledge of matters relevant to the explanation.

Three experiments were conducted using the XQ Survey. The participants were recruited from MTurk workers.

4.7.2.1 The First MTurk Experiment: Validation of XQ Rubric and XQ Survey as Evaluation Tools

The first experiment conducted with the XQ Survey was intended to establish a methodology for performing the XQ Survey with participants from MTurk (Chapter 6). The advice of Paolacci et al. (2010) on the best practice for MTurk was followed. Two case studies were presented to the participants. The first was a letter sent to clients of Centrelink, an Australian government welfare agency, asking for information about their employment earnings (Section 4.6.6). The second case study was a notice informing applicants that their requests for a loan had been rejected by an AI algorithm (Section 4.6.5).

4.7.2.2 The Second MTurk Experiment: Validation of XQ Rubrics as a Tool for Constructive Feedback

The second experiment conducted with the XQ Survey was intended to show that its results could be used as feedback to improve an explanation (Chapter 7). Based on the feedback gathered in the first experiment, the IBM loan case study was revised and presented to a different group of MTurk workers. The XQ Surveys were structurally similar, and the demographic profiles of the two groups was similar. The revised case study results were compared with the original case study results to assess the effectiveness of the case study revisions.

4.7.2.3 The Third MTurk Experiment: Independent Validation of the XQ Rubric and the XQ Survey

The third experiment was designed to assess whether someone not involved in creating the XQ Rubric and its associated methodology could nevertheless use it to assess and improve their explanations (Chapter 8). A colleague, Dr Francisco Cruz a Deakin University AI researcher, assisted with this experiment. Dr Cruz created the explanations and collaborated in adjusting the unguided score section and the demographic section to fit his explanations. He then evaluated the XQ Survey results from Round 1 of the XQ Surveys and used them to improve his explanations. The revised explanations were then evaluated a second time by MTurk participants. The Round 2 results were compared with the Round 1 results to assess the impact of Dr Cruz's improvements.

4.8 CONCLUSION

This thesis followed a Design Science methodology as advocated by Peffers et al. (2007). The aim was to produce artefacts (the XQ Rubric and Survey) that could be used to evaluate and provide critical feedback on XAI explanations designed for a non-expert audience.

The Design Science method was suitable for this thesis because its purpose is to help create objects and procedures. The Peffers et al. (2007) methodology was followed because it was tailored to Information Systems design. It provided a comprehensive and clear method for the creation of the XQ Rubric and Survey.

The research conducted a literature review to explore the problem areas of XAI explanation and from this form the basis of the initial rubric. This rubric was revised in three rounds of a Delphi panel. The rubric and survey for evaluation and critical feedback were then validated through three survey-based experiments conducted via MTurk.

5 THE FOCUS GROUP USING THE DELPHI METHODOLOGY

5.1 INTRODUCTION

A Delphi methodology for focus groups (Section 4.7.2) was used to test and refine the XQ Rubric designed to evaluate XAI explanations. The participants chosen to form the focus group were known to me. All are experts in their field of interest. Some were experts in case study topics. Others were educators familiar with rubrics. One was a teacher, and another a researcher in AI reinforcement learning.

The focus group experiments were intended as a pilot study for the MTurk experiments, designed to refine the XQ Rubric and the XQ Surveying method before these were released to a broader audience.

The Ethics Approval for this experiment is presented in Appendix B1.

5.2 METHODOLOGY

The proposed methodology was a three-round Delphi process, a well-established procedure that uses an iterative process to collect and analyse experts' responses to prompts (Skulmoski et al., 2007, p. 2). The Delphi process has four elements: anonymous responses, iteration, controlled feedback, and statistical group response (Dalkey, 1969). The experiment was performed in three stages.

The first stage introduced the topics, that is the marking rubric and several case studies of explanations of decisions produced by XAI systems. Feedback was invited. The second stage summarised feedback from the first stage, invited feedback on the feedback, and reintroduced the topic. The third stage was the same as the second stage. As recommended by Dalkey and Helmer (1963), more emphasis was placed on the feedback summary and less on the case

studies. After the third and final stage, an email was sent to the participants summarising the results of the Delphi experiment. This followed the methodology of Skulmoski et al. (2007).

The four stages of the focus group experiments were designed and implemented iteratively, each informed by the outcome of the previous stage. These can be divided into four parts: pretesting stage, stage one, stage two, and stage three. A face-to-face focus group would probably have been useful, but COVID-19 restrictions made this impracticable.

5.2.1 Pretesting Stage

The pretesting stage included the development of the XQ Rubric design, the ethics application, and the selection of XAI explanations to be used as case studies. The XQ Rubric design emerged from the literature review. The initial layout of the XQ Rubric was based on rubrics used by educators for essay marking. The categories used in the XQ Rubric were based upon the model of a good explanation (Section 2.5.4) combined with the rubric developed by Sevian and Gonsalves (2008) to evaluate graduate students' explanations of their scientific research. This version of the XQ rubric is presented in Appendix D1.

As described in Section 4.6, the five case studies used in this experiment were selected from the literature.

Selecting the Participants

First, it was decided which case studies would be used. Participants were selected for their expertise in one or more of the case study topics. Two of the case studies concerned medical matters, so three doctors were invited to participate in the experiments. Since XAI has a growing role in the administration of the law (see for example, COMPAS (Brennan et al., 2009)) and the ways XAI explanations may be presented (see for example, the GDPR (Wu, 2017)), two lawyers were also invited to participate. Because the rubric is an evaluation method partly developed for use in the education field two teachers were invited. Finally, six experts in Information

Technology (IT) were invited to participate. The participants then sent the email presented in Appendix G1a. A follow up email was sent to the participants confirming their involvement (Appendix G1b).

5.2.2 Stage One

On 8 November 2019, the participants were emailed (Appendix G2a). The original rubric (Appendix D1) and the five case studies (Appendix C1) were attached to the email. The participants were asked to evaluate the attached marking rubric and apply it to the five supplied case studies (Appendix C1).

This is an excerpt from the first email to participants:

> *Firstly, please review the attached marking rubric on its own.*
>
> - *What does it do well?*
> - *Are the criteria clear?*
> - *What changes would improve it?*
>
> *Secondly, please apply the marking rubric to the five attached explanations of artificial intelligence behaviour.*
>
> - *How did each explanation score against the rubric?*
> - *What parts of the rubric helped to evaluate these explanations?*
> - *Now that you have applied it, what changes would improve the rubric?*

On 24 October 2019, an email was sent asking for the completed rubrics to be returned (Appendix G2b). Two months after the first email (8 November 2019), a reminder email was sent (Appendix G2c). Of the nine people invited to participate, six returned surveys.

One participant commented that the experiment took too much work (Appendix A1), so the number of case studies was reduced from five to one. Non-educators reported that they found the XQ Rubric confusing (Appendix A1).

5.2.3 Stage Two

The design of an online survey was finalised a few months after Stage One. In the second experiment, responding to feedback from non-educators that the rubric format was "confusing", it was changed into a multi-choice survey (Appendix D2), where each row was one question. The eXplanation Quality Survey (XQ Survey) was also reduced to just one case study, for participants complained that the first stage took too long to complete. Also in response to feedback, the language and instructions of the rubric were made clearer.

On 22 January 2020, emails were sent inviting participants to complete the online survey. One email sent was sent to each person who participated in Stage One (Appendix G3a). A different email, inviting them to take part in Stage Two (Appendix G3b), was sent to the participants who did not respond to the first survey.

Below is an excerpt from the email sent to participants who participated in Stage One:

> *Thank you for responding to part one of the 2019 Explainable Artificial Intelligence Delphi survey.*
>
> *Five people responded to this survey. Most people found the wording of "target audience" confusing and did not feel qualified to comment on how other people may feel about the explanations. Most people also found some of the language in the rubric confusing. The case studies themselves were criticised for being unclear and confusing.*
>
> *This is the second email in the Research Project titled: Measuring the Effectiveness of Explanations of the Decisions of Artificial Intelligence (AI) Algorithms*

> *The second survey is an online questionnaire that is much shorter and simpler than the first survey because of feedback I received about the first survey. It should only take 10 minutes to complete. It is using the same case study as the previous survey, however, the questions relating to the case study have been changed.*
>
> *I will aggregate everybody's responses and share the results, along with follow-up questions, for the next stage of the Delphi methodology.*

On 11 February 2020, a follow-up email was sent to participants (Appendix G3c). Finally, on 11 June 2020, the results were compiled and analysed (Appendix A2). Of the nine people invited to participate, six responded.

The second stage was better received than the first, and non-educators found the new format less confusing. Feedback from this stage mostly concerned the wording of the instructions, questions, and answers.

5.2.4 Stage Three

The third stage of the experiment was a review of the XQ Rubric and Survey in preparation for the MTurk experiments. The wording of the XQ Survey was refined, and questions seeking information about the demographic profile of the participants were added (demographic questions covered age, location, occupation and experience with AI and themes in the case study) (Appendix E1b). A section was added inviting participants to give two case studies, the drone Case Study and the Centrelink case study, a score out of 10 to determine how well this score matched the XQ Rubric score (the unguided score section of the XQ Survey).

Following advice on how a MTurk survey should be conducted, reading comprehension questions for both case studies were added (Paolacci et al., 2010). These were designed to determine whether the participants had read the case studies and were not merely selecting answers at random.

This is an excerpt from the email sent to participants:

> *Thank you for agreeing to participate in my Explainable Artificial Intelligence (AI) Delphi Group, part of a study designed to measure the effectiveness of explanations of decisions made by artificial intelligence algorithms.*
>
> *I have aggregated the responses from the first and second questionnaires, and designed a further questionnaire based on this feedback. This third iteration of the questionnaire will be used to survey people on Amazon's Mechanical Turk. I would like you to look over this third survey and confirm that the survey is ready to be sent to a wider audience.*

Because this was a pilot study, when the third experiment was presented to the focus group participants were explicitly requested to concentrate on the wording of the questions (Appendix G4a). To this end, they were not required to use the XQ Rubric in Case Study 2 because they had used it only once, to give feedback. Of the nine people invited to participate, only three responded.

5.2.5 Final Email

A final debriefing email was sent to the focus group participants (Appendix G5). The findings of this experiment were explained to participants, who were given the opportunity to respond.

This is an excerpt of the final email sent to the focus group participants:

> *Thank you for your contribution to the 2019 – 2020 Explainable Artificial Intelligence Delphi Group survey, part of my research project 'Measuring the Effectiveness of Explanations of the Decisions of Artificial Intelligence (AI) Algorithms'.*
>
> *This is the debriefing email for the Research Project titled: Measuring the Effectiveness of Explanations of the Decisions of Artificial Intelligence (AI) Algorithms.*
>
> *This research project allowed me to receive detailed and thoughtful feedback through a series of experiments about my marking method and evaluation rubric.*

The feedback from the experiments led to incremental improvements. I also learned that a short, online survey is the best methodology to gather informative and timely feedback from a range of people. The feedback also led to me altering my original methodology, changing the original rubric format to a survey.

Although there were incremental improvements between experiments, I cannot make any definitive statements until I do a survey with more people. Therefore, I plan to continue my experiments with the Amazon Mechanical Turk website.

5.3 THE DRAFT XQ RUBRIC

The XQ Rubric discussed below (Figure 5-1) was the first draft of the XQ Rubric. It was created with no input from the focus group. This draft rubric was intended to generate discussion from which a better XQ rubric design would emerge.

The model for the XQ Rubric was the rubric of Sevian and Gonsalves (2008) used to evaluate graduate students' explanations of their scientific research. However, the Sevian and Gonsalves (2008) rubric is an education rubric, designed in part to evaluate the presenter of the explanation. The XQ Rubric, by contrast, evaluates the explanation only.

The XQ rubric was divided into families. These were labelled from A to D and coloured in distinctive colours (Figure 5-1). Categories were labelled by their initials, for example the Presentation Clarity category became PC. Each family included several categories that covered the marking criteria of one aspect of the explanation (Figure 5-1).

Family	Category	Excellent (5)	Good (3)	Mediocre (1)	Bad (0)	Comments
Presentation Clarity	Suitable for the audience	Tailored for the audience	The explanation suits the audience	The explanation is written generally but uses words and terms the audience is familiar with	Written generally and uses words and terms the audience is unfamiliar with	
	Obeys the conventions of its medium (eg Grammar and Spelling)	Obeys the conventions of its medium	Mostly obeys the conventions of its medium – one or two small mistakes	Somewhat obeys the conventions of its medium – more than two small mistakes	Does not obey the conventions of its medium	
	Clear wording and unfamiliar terms/symbols defined	Glossary included if new words are introduced. The wording is clear	The wording is clear and term/symbols used are standard, uncommon words are defined	The wording is clear and uncommon words are defined	Terms introduced without introductions, unusual symbols used and not defined	
Content	Decision, Action, or Phenomena is explained	The thing is explained clearly	The thing is explained.	An unsuccessful attempt to explain the thing	The thing is not explained	
	How algorithm works is explained	How algorithm works is explained clearly	The algorithm is explained.	An unsuccessful attempt to explain the algorithm	How algorithm works is not explained	
Satisfaction	The target audience understands the explanation	The target audience understands and accepts the explanation	The target audience mostly understands and accepts the explanation	The target audience is unsure about the explanation	The target audience misunderstands the explanation	
	The explanation is clear and believable to the target audience	The explanation is clear and believable to the target audience	The explanation is mostly clear and believable to the target audience	The explanation is believed with some doubt to the target audience	The explanation is unclear and unbelievable to the target audience	
Truth	Verifiable references and plausible claims	The explanation has more than one verifiable reference and plausible claims	The explanation has plausible claims and references	The explanation has some verifiable references and semi-plausible claims	Unverifiable references and implausible claims	
	Claims reference evidence	The explanation has more than one piece of verifiable evidence	The explanation has plausible evidence	The explanation has some evidence	No evidence presented	
	The explanation is relevant to the discussion	The explanation answers the question (or implied question)	The explanation mostly answers the question (or implied question)	The explanation is only partly answers the question (or implied question)	The explanation is not relevant	
Total						

Figure 5-1 Image of Draft XQ Rubric

5.3.1 Presentation Clarity

The criteria of the 'presentation clarity' family were used to evaluate the explanation's clarity, that is, its presentation, its spelling, and its grammar. The evaluation participants were asked to consider if the explanation could be acted upon or if it was merely a summary of events with no analysis and no path forward.

Sevian and Gonsalves (2008) have similar categories in their rubric: "*Quality*", "*Clear choice of language*", and "*Skill of presentation (technical use of media)*" (p. 1450).

While a poorly presented explanation may be true and may satisfy the target audience, if it is too poorly presented, it will not succeed in explaining what happened and why.

Suitable for the audience

This section evaluated whether the explanation was pitched at the appropriate level for its audience.

Obeys the conventions of its medium (e.g. grammar and spelling)

This section evaluated whether the explanation was well written and presented.

Clear wording and unfamiliar terms/symbols defined

This section evaluated whether the explanation used precise wording and defined unclear and unusual terms and symbols.

5.3.2 Content

The 'Content' family criteria evaluated whether an explanation really was provided. If there was no explanation to evaluate, then there was no need to use the rubric.

The Decision, Action, or Phenomenon is explained

This section evaluated whether an explanation was provided.

How the algorithm works is explained

This section evaluated whether the explanation included information about how the algorithm works.

5.3.3 Satisfaction

The criteria in the 'satisfaction' family evaluated the reactions of the target audience to the explanation. The explanation was evaluated by asking the evaluators whether they understand it clearly. The evaluators were also asked to consider whether the target audience would believe it.

The categories in the Sevian and Gonsalves (2008) rubric upon which this section is modelled are "*use of mental images to support explanation*" and "*scaffolding explanation*" (p. 1454).

The satisfaction categories were included to determine whether the audience failed to understand or refused to believe the explanation. Such an explanation would not be useful (see the established model of good explanation (Section 2.5.4)).

The target audience understands the explanation

This section evaluated whether the target audience understood the explanation.

The explanation is clear and believable to the target audience

This section evaluated whether the target audience would find the explanation clear and whether they would believe it.

5.3.4 Truth

The criteria of the 'truth' family evaluate the truth claims of the explanation. The truth criteria assess whether the explanation has provided references and whether it includes facts relevant to the discussion. The verifiability family of categories also explores whether the claims are plausible.

This ‘truth’ family was included because in explanations truth is fundamentally important (see the established model of good explanation (Section 2.5.4.1)).

Sevian and Gonsalves (2008) have similar categories, "*Factual knowledge*" and "*How knowledge is understood*" in their rubric (p. 1451). These ascertain whether the presenter has given correct and relevant information. In other areas of their rubric Sevian and Gonsalves (2008) acknowledge the importance of references.

Verifiable references and plausible claims

This section evaluates whether the explanation has verifiable references and plausible claims. In the absence of the ability to assess any direct truth of the XAI's claims, this was considered a reasonable proxy.

Claims reference evidence

This section evaluated whether the references provided in the previous section were relevant to the explanation.

The explanation is relevant to the discussion

This section evaluated whether the explanation was relevant to the discussion. An irrelevant explanation would be useless, and its audience would gain nothing from it.

5.3.5 Summary And Conclusion

This draft rubric was revised many times. However, the XQ rubric continued to be concentrated on the presentation and content of the explanation. This emphasis on the audience’s satisfaction and the truth of the explanation are fundamentally important, and made more quantifiable in later iterations.

5.4 RESULTS

5.4.1 Stage One Results

As this was a prototype rubric, the scores given by the five participants were of less importance than their comments and suggestions about the wording and appearance of the rubric itself. Only one of the participants offered no comment.

One of the participants, a medical doctor, marked the medical case studies more harshly than other case studies, perhaps revealing a degree of professional bias. However, a bias of this kind is minimised when many people from different backgrounds use the rubric. To help expose biases of this kind, in later experiments XQ Survey participants were asked about their knowledge of relevant subjects and current jobs as recorded in the Demographic part of the XQ Survey.

Two of the five participants criticised the rubric sections headed "*The target audience understands the explanation*" and "*The explanation is clear and believable to the target audience*". The phrase "*target audience*" was criticised for being too confusing. One participant wrote that because they were not part of the "*target audience*", they could not imagine how the target audience would feel about the explanation. The phrase "*target audience*" was removed in answer to this criticism, and the rubric sections were phrased more straightforwardly.

Two of the five participants criticised the rubric section "*Decision, Action, or Phenomena is explained*". It was felt that the rubric answer referencing "*the thing*" was unclear and imprecise (see Table 5-1 for the phrase in context). The phrase "*the thing*" was removed, and the section was reworded.

Table 5-1 Mark of "Decision, Action, or Phenomena is explained" by case study

	Case Study					Grand Total
	1	2	3	4	5	
The thing is explained clearly	0%	0%	20%	20%	20%	**12%**
The thing is explained	60%	80%	20%	40%	40%	**48%**
An unsuccessful attempt to explain the thing	40%	0%	40%	40%	40%	**32%**
The thing is not explained	0%	20%	20%	0%	0%	**8%**
The thing is explained clearly	0%	0%	20%	20%	20%	**12%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Another participant, who had not filled in the surveys nor commented on them, said that they felt that the XQ Rubric took too long to complete and suggested that using a paper rubric was inefficient. In response to this, an online survey with the same questions as the rubric was created. This survey was not presented in a rubric's traditional grid format. It had instead the appearance of a multiple-choice questionnaire.

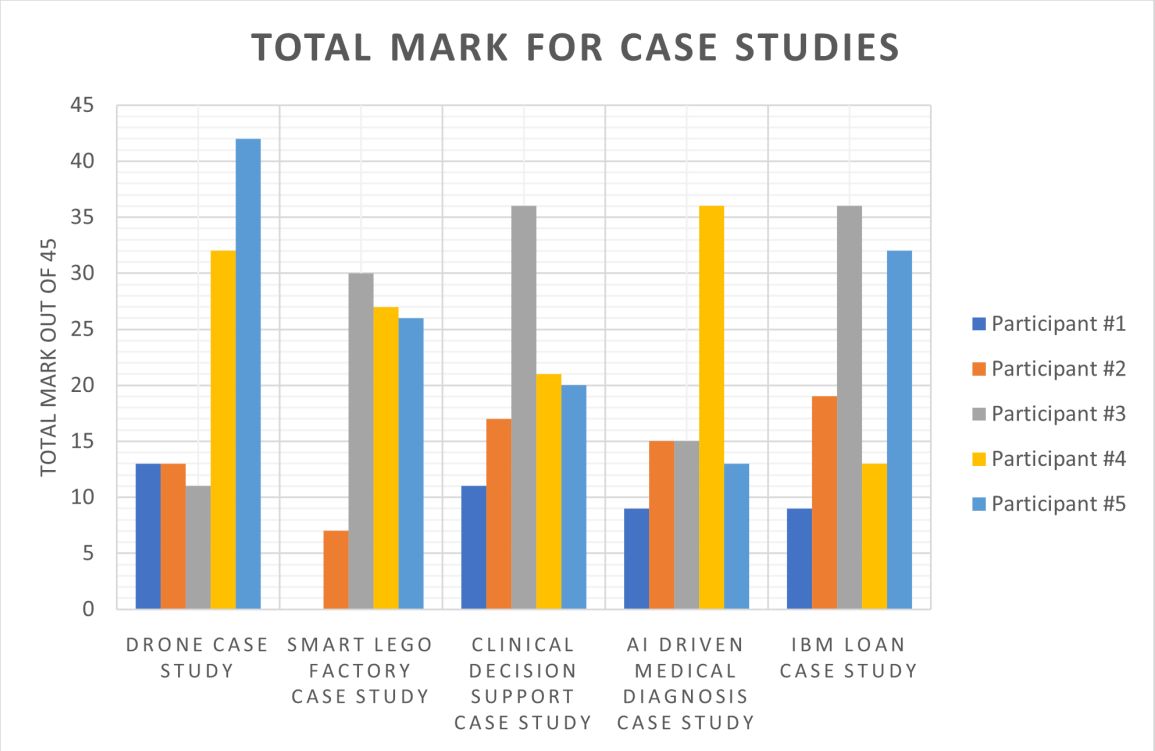

Figure 5-2 Agreement Between Participants (Stage 1)

Figure 5-2 shows an apparent lack of agreement between participants over which case study offered the best explanation. The case studies are marked out of 45 because this was the highest possible score for a case study using the Draft XQ Rubric. If the participants had agreed on

quality but varied in how harshly they marked, the lines would follow the same path; for example, participants #1 and #2 frequently agreed on how the case studies compared to each other in quality, though Participant #1 marked the case studies more harshly than Participant #2. The lack of marking agreement shows that this is a poor rubric, not meeting one of the stated aims of this thesis, the creation of a robust evaluation method for XAI explanations. A good rubric would facilitate and register agreement between evaluators (Moskal & Leydens, 2000).

5.4.2 Stage Two Results

Stage Two was designed to test the participants' responses to the changed wording. It was also intended to test an online system of evaluating explanations, using the XQ Rubric as a base. Participants were asked to review only Case Study 5. The online survey system assigned participant numbers randomly, so a participant identification number for Stage Two was not necessarily assigned to the participant using it in Stage One.

The changes to the rubric and its instructions resulted in greater consistency between participants. Whereas in Stage Two, the total scores (excluding b_user #7) ranged from 8 to 16 out of 45 (Table 5-1), in Stage One, the total scores ranged from 0 to 42 (Figure 5-1). Since a better rubric would produce more consistent results (Moskal & Leydens, 2000), in this regard the Stage Two rubric is better than the Stage One rubric.

Table 5-2 Scores given by Participants (Stage 2)

ID No.	Question Number									Total Score	No. of responses	Score Av.
	1	2	3	4	5	6	7	8	9			
a_User #4	0	1	0	1	0	*	5	1	*	8	7	1.1
b_User #7	3	5	0	1	3	0	3	1	5	21	9	2.3
c_User #9	3	1	1	*	0	3	1	1	1	11	8	1.4
d_User #11	1	1	3	1	0	1	3	3	3	16	9	1.8
e_User #12	3	1	1	*	0	0	1	3	1	10	8	1.3
f_User #13	3	1	1	1	0	1	1	3	3	14	9	1.6

** indicates that no response was given to this question*

Since all participants except b_user #7 commented at least once (Table 5-2), and only b_user #7 responded to question 5 positively (Table 5-1), I believe b_User #7's answers were selected

without care and should be discarded. To determine which participants did not take the survey seriously, comprehension questions were introduced. These were designed to show whether the participant had read the explanation and had participated in good faith.

Table 5-3 Number of Comments by Participants (Stage 2)

User	Number of comments
a_User #4	9
b_User #7	0
c_User #9	8
d_User #11	2
e_User #12	1
f_User #13	3

Feedback on the XQ Rubric

The XQ Survey response from participant a_User #4 was critical of the XQ Rubric, the XQ Rubric's presentation, and the explanation itself. The presentation of the explanation was criticised as confusing, and the participant did not understand where the explanation began and ended. These criticisms were dealt with by making it clear what was background and what was explanation.

Participant C_User #9 found question 5: "*Are references and citations provided?*" unnecessary because references and citations were not needed for this explanation. An option was added for participants to indicate that references and citations were not relevant.

A_User #4 found question 6 unclear: "*Does the explanation refer to facts surrounding the matter being explained?*". The question was changed to: “Are relevant facts mentioned in the explanation?”.

5.4.3 Stage Three Results

Stage Three had significantly fewer participants, with only three people completing the XQ Survey. Despite the low number of participants, there were still enough to act as reviewers for the XQ Rubric and Survey in preparation for the MTurk experiments.

Participant #5 consistently gave the case study a mark of '2' or '0' (Table 5-4). The other two users gave more varied responses.

Table 5-4 Scores given by participants using the XQ Rubric (Stage 3)

Question Number	Participant #2 Scores	Participant #4 Scores	Participant #5 Scores
1	1	0	2
2	1	2	2
3	0	0	0
4	1	1	2
5	2	1	2
6	N/A	0	0
7	3	0	0
8	1	1	2
9	1	1	2
10	0	1	2
Total	**10**	**7**	**14**

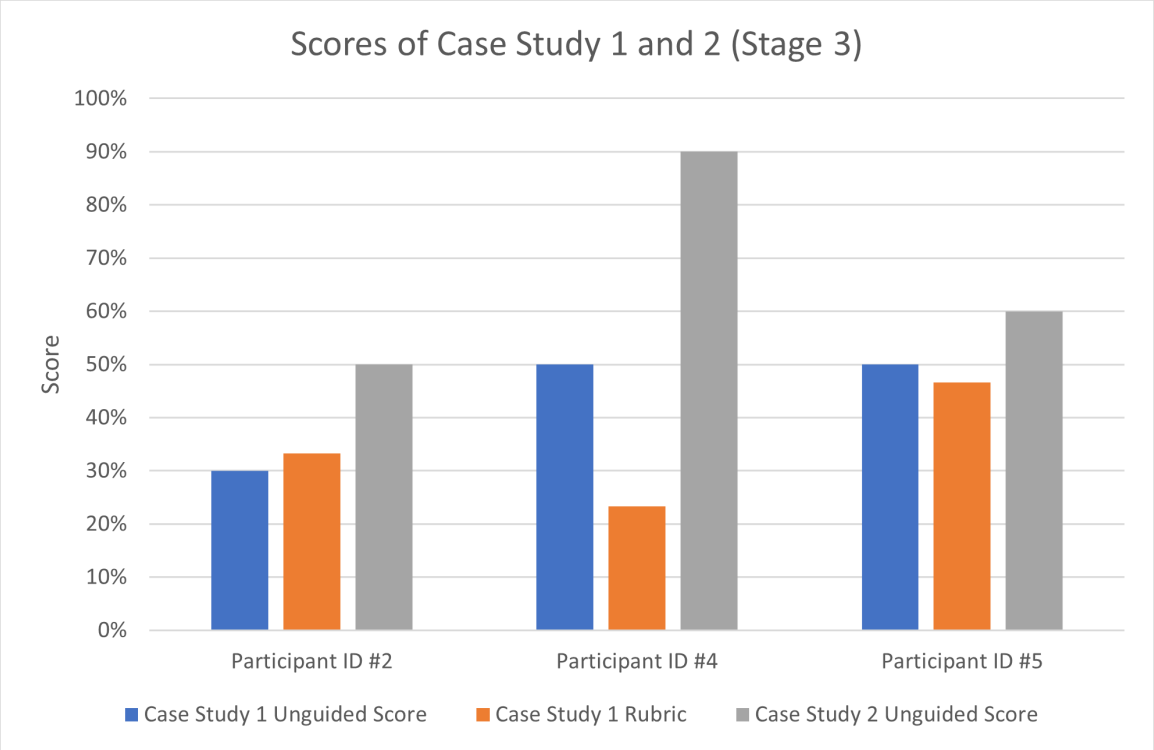

Figure 5-3 Scores given to case studies by participants (Stage 3)

In figure 5-3, the unguided or initial score of Case Study 1 (the drone case study) is similar to the XQ rubric score of Case Study One when expressed as a percentage. Case Study 2 is the Centrelink RoboDebt letter case study. This similarity indicates a significantly more consistent

rubric. This rubric was the most suitable for the MTurk experiments of the rubrics presented in this experiment.

Feedback about the XQ Rubric

The focus group participants offered no feedback on the survey, survey presentation, rubric, and rubric contents of this stage.

5.5 The Updated XQ Rubric

The XQ Rubric discussed below was the final version presented to the MTurk participants. The XQ Rubric was presented as an online survey, and for ease of administration each question was given a code.

The question's code is based on its Family (starting with 'F0') and its line in the XQ Rubric (starting with 'C0'). The 'Family' is a grouping of questions with a similar theme. This layout made it easier to process the large quantity of data from the online survey. The discussion of the updated XQ Rubric is organised by Family. Individual questions are discussed below their Family discussion.

The use of the Likert scale seems to have helped improve answers to the rubric questions. A Likert scale is a survey tool that "*respondents indicate their degree of agreement or disagreement on a symmetric agree-disagree scale for each of a series of statements"* (Burns & Bush, 2007). The scale, widely used, is often phrased as 'Strongly Agree' through three declining levels ('Agree', 'Neutral', and 'Disagree') to 'Strongly Disagree'. A modified version of the Likert scale suggested by Siegle (2015) provided a clear guide for the possible answers of the XQ Survey.

5.5.1 Content Category

The content family was evaluated to determine whether an explanation had been provided. The XQ Rubric, of course, was unnecessary if no explanation was given. Since this category asked redundant questions, it was removed from the Updated XQ Rubric.

5.5.2 Presentation Clarity (F01)

The rationale behind this family is the same as the rationale offered in Section 5.3.1. The questions changed, though not the reason for asking them.

F01C00 Is the explanation clear?

This question was asked to ascertain whether the evaluator could understand the explanation. It provide a response to which other questions could later be compared.

F01C01 Is the explanation written with correct spelling and grammar?

The evaluator was asked to comment on an essential element of the explanation, its spelling and grammar. In order to make its meaning clearer this question rewords the draft rubric category, "*Obeys the conventions of its medium (e.g. Grammar and Spelling)*".

F01C02 Does the explanation define the uncommon words and symbols it uses?

Evaluators were not expected to understand uncommon words and symbols used in the explanation. If they could not, the explanation would be unclear and difficult to comprehend and useless as an explanation. To suit the online version of the XQ Rubric better, this question rewords the draft rubric category, *"Clear wording and unfamiliar terms/symbols defined".*

F01C03 Is the supporting information (or detail) well-presented and understandable?

This question asks about supporting information (or detail). Supporting information is "*additional information that explains, defines or proves the main idea*" (Your Dictionary, 2022). It is important, of course, to consider the role of supporting information in making an

explanation's argument easier to understand. This question did not reword any part of the draft rubric.

F01C04 Can the explanation be acted upon?

This question asks the evaluator to consider whether the explanation can be used. Most case studies in this thesis were written explanations aimed at motivating an audience to act, change their behaviour (IBM loan case study), or accept an AI's decision (drone case study and Dr Cruz's case study). The explanation should be capable of being acted upon, that is, doing something directed by the AI, accepting its conclusions, or challenging them. This question was not included in the draft rubric.

5.5.3 Verifiability (F02)

The rationale behind this family is as it was for Section 5.3.4. The questions changed, though not the reasoning behind them.

The category name was changed because 'Verifiability' was closer to this category's purpose.

F02C01 Are references provided?

This question asks the evaluators to consider whether the explanation cites any references. Following feedback from Stage Two, participants were given the option of marking this question "not relevant".

F02C02 Are relevant facts mentioned in the explanation?

This question asks the evaluators to consider whether the explanation mentions all relevant facts. An explanation that does not include relevant facts is baseless and unverifiable. An unverifiable explanation is untrustworthy and unhelpful.

The wording of this question was changed following feedback from Experiment Two.

5.5.4 Satisfaction (F03)

The rationale behind this family is the same as it was in Section 5.3.3. The questions have changed from their emphasis on the 'target audience', though not their intention.

F03C01 Is the question (or implied question) answered?

This question asks the evaluators to consider whether they are satisfied with the explanation and whether they feel that their question has been answered.

F03C02 Can the explanation's reasoning be followed without difficulty?

This question asks the evaluators to consider whether they can easily understand the explanation's reasoning.

F03C03 Is the explanation convincing?

This question asks the evaluators to consider whether they were convinced by the explanation.

5.5.5 Other (Other01)

In the final question, evaluators were given the option of providing feedback about the explanation. This question was included because the XQ Rubric could not cover every possibility, and evaluators may have wished to comment on some aspects of the explanation not covered.

5.6 Reflections on the XQ Rubric and Survey

These experiments found that short, unambiguous, online, multiple-choice surveys were better than the corresponding alternatives. While surveys can be modelled on rubrics (one row per question), the traditional table format of a rubric is sometimes confusing. An important part of developing and conducting the experiments was the task of expressing the questions, answers, and instructions clearly, simply, and unambiguously.

There was a significant decline in the number of participants. Figure 5-4 shows the dropout rate, which was especially large between Stage Two and Stage Three. It was difficult to manage a

long-distance focus group, especially when participants were restricted in various ways by the COVID-19 pandemic. I had planned to follow this experiment with face-to-face focus group meetings, but COVID-19 considerations prevented it.

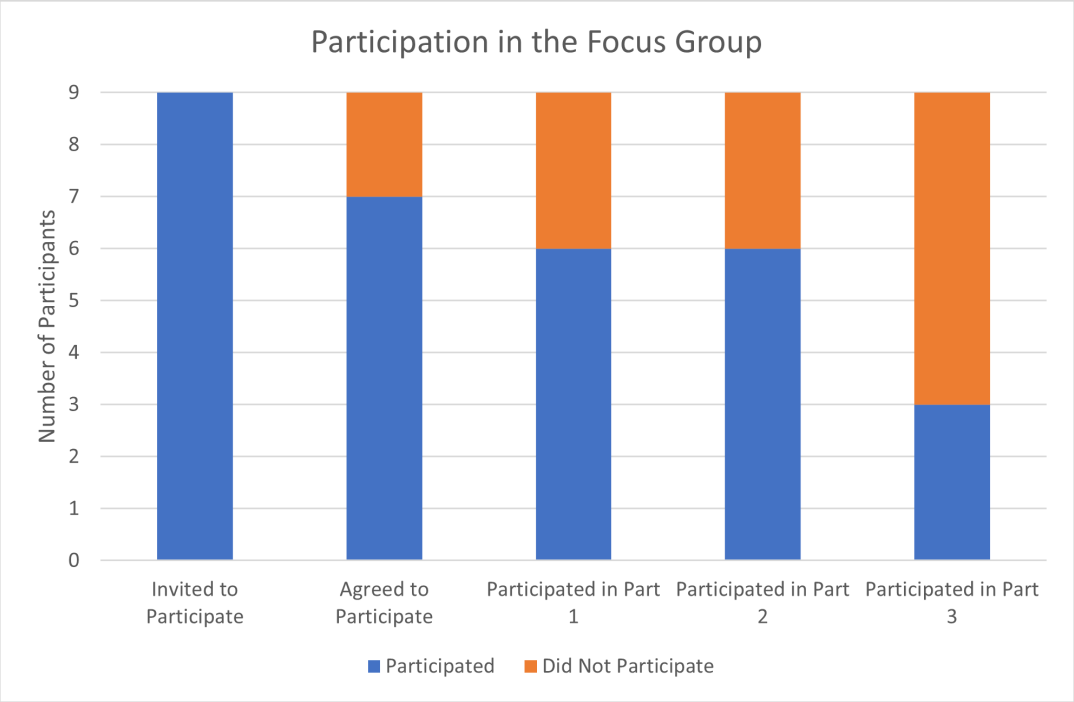

Figure 5-4 Participant Dropout

5.7 Discussion

A significant finding was that all the case-study explanations were regarded as inadequate by members of the focus group. They found that the explanations lacked content and were poorly presented. This was also noticed when the case studies were being developed. It was frequently found that many papers were missing complete examples and relevant information. This lack of information made it hard to compare XAI methodology and explanations between cases.

Focus group participants preferred a short, multiple-choice, online survey to a paper-based rubric. The multiple-choice presentation prevented confusion about how a traditionally presented rubric should be used. Online surveys were easier for participants to access and made comparison between responses easier. With this change in survey methodology, the second and third experiments produced more consistent results.

These experiments were designed to serve as a pilot study for future MTurk experiments. Through a series of reiterative focus groups, the participants refined the XQ Rubric and the XQ Survey method.

In particular, the XQ Rubric changed its wording and its method of presentation. The wording of the rubric was made more precise, and its presentation was changed from a paper rubric to part of an online survey.

The XQ Survey was created to host the new XQ Rubric and, by the use of comprehension and demographic questions, to provide additional context for the answers of its users.

Limitations of these experiments

These focus group experiments were limited in three ways:

1. there were few participants (between three and six)
2. all of them were Australian
3. all of them were well educated

These limitations were removed in the follow-up MTurk experiments, which had more participants, with a broader range of backgrounds.

With the exception of the third stage of the Focus Group experiment, the focus group experiment was also limited by the lack of other scoring methods. This was rectified by including a question in the follow-up experiments prompting participants to score the case study on a scale from 0 to 10 in the ‘unguided score’ section.

6 INITIAL VALIDATION OF XQ RUBRIC AND XQ SURVEY AS EVALUATION TOOLS

6.1 INTRODUCTION

MTurk is an Amazon.com crowdsourcing service commonly used to find participants for surveys (Di Gangi, McAllister, Howard, Thatcher, & Ferris, 2022). As part of this thesis's iterative design methodology, MTurk was used to follow up the results of the Focus Group Experiments. These experiments rectified some of the limitations of the Focus Group Experiment, with its small number of participants and lack of alternative scoring methods. The MTurk experiments surveyed substantially more people than were surveyed by the focus group. The XQ Survey participants were from cultural and educational backgrounds different from those of the focus group. The focus group participants had similar backgrounds: they were all Australian, professional, and educated. The XQ Survey was intended, however, to be appropriate for non-experts of a great variety of backgrounds.

The MTurk experiments were the next step in the iterative design process of creating an evaluation methodology for XAI. The MTurk survey used the XQ Survey created from the Focus Group experiments, largely that of Stage Three.

Because it provided access to a larger subject pool with a more diverse range of participants, including people not familiar with AI, MTurk was used to follow up the Focus Group Experiment.

The Initial Validation Experiment was designed to evaluate the XQ Rubric more closely than it had been in the previous experiment. The new experiment attempted to demonstrate that the XQ Rubric was reliable, easy to use, and accorded with participants' feelings about the case study. The experiment was also used to improve the methodology of the XQ Survey.

The Initial Validation Experiment was part of a more extensive set of experiments, the MTurk experiments, which followed the last four stages of the methodology of Peffers et al. (2007) (Section 4.1).

The Ethics Approval for this experiment is presented in Appendix B2.

6.2 Methodology

6.2.1 Structure of the XQ Survey

The XQ Survey used two case studies. Case Study 1 was the IBM Loan Rejection (Appendix C2a). Case Study 2 was the Centrelink RoboDebt Letter (Appendix C2b). These were chosen for two reasons:

1. of the six case studies available, these were the most widely reviewed, making it possible to compare the XQ Rubric and XQ Survey results against the reviews
2. the IBM loan case study includes figures and scores. The Centrelink RoboDebt case study does not. This allows the use of the XQ Rubric and Survey to be examined in different contexts.

Each case study was presented to the participants twice. First, they were required to answer three questions to establish that they understood what was being presented. Then they rated the explanation, on a scale of 1 to 10, offered by the case study. These four questions comprise the Unguided Score question group.

To allay concerns that MTurk workers would not complete the work as intended and instead choose answers randomly without first reading the Case Studies and the corresponding questions, they were obliged to answer comprehension questions to confirm that they had indeed read the case study (Kittur, Chi, & Suh, 2008). Each case study had three comprehension questions. The first was multiple-choice with a single correct answer. This question served to

check whether the participant had read the case study. The other questions were open-ended with one single correct answer. These were also designed to check whether the participant had understood the case study.

On completing the Unguided Score section, the participant was asked to rate the case study explanations using the survey section based upon the XQ Rubric. This part of the XQ Survey was known as the Rubric Score section. The section's questions were taken from the final version of the XQ Rubric developed in the Focus Group Experiment (Section 5.5). The case study was presented for the second time in the XQ Survey immediately before the XQ Rubric section. This gave participants a copy to refer to when answering the XQ Rubric questions. They may have deleted or forgotten the previous presentation of the case study.

Having completed the Rubric Score section, participants were presented with questions about their age, location and occupation, and questions about how experienced they were with loans, with Centrelink (the Australian welfare agency), pop-culture presentations of AI, and non-fiction discussions of AI (here called "practical AI"). Participants were then given a code used later to prove that they had completed the XQ Survey (see Section 4.4.3).

Best practice for MTurk, as suggested by Paolacci et al. (2010), was followed. An email address, *XAIMTurkSurvey@federation.edu.au*, was created for the MTurk experiments. The MTurk site and the LimeSurvey site for the XQ Survey were also monitored for any problems with the survey and MTurk payments.

Participants were paid US$7.25/hr; this was equivalent to US$4.00 per survey (with an average expected time of 30 to 35 minutes). The payment was based on an average completion time of 30 minutes and an hourly rate of US$7.25 (the US minimum wage). This rate was recommended by Hendrickson et al. (2015).

6.2.2 MTurk Procedure

The MTurk experiments were conducted in two batches. The first was open to 70 participants, though only 69 participants completed the survey. Some of the participant slots were exclusively for MTurk workers designated as “Masters”. Some were available to ordinary participants.

On analysis, it became clear that there were insufficient participants. The data was too scattered, and there was not enough evidence to support firm conclusions. A second batch of participants was recruited to supplement the first.

The participants from the first batch also gave too many responses in bad faith and responses that failed to answer the comprehension questions correctly. Master MTurkers only were recruited to the second batch. The XQ Survey was taken more seriously by the new participants, and more thoughtful responses were provided.

The MTurk monitoring system did not allow participants from the first batch to be recruited for the second batch. Out of a possible 50, 49 people participated in the second batch of the XQ Survey.

6.3 RESULTS

6.3.1 Comprehension Questions

The comprehension questions presented in the Unguided Score section were evaluated. Some wrong answers were deemed to have been given in bad faith. Some answers, though not given in bad faith, were only partly correct or had misinterpreted the question. These answers were deemed to be neither correct nor incorrect.

There were two types of comprehension questions, with the type indicated by the question’s name. Names that started with ‘Gatekeep’ had only one correct answer. Question names that started with ‘Compre’ were open-ended, though there was a correct answer.

6.3.1.1 The IBM Loan Case Study Comprehension Questions

Table 6-1 Type of Answers to IBM Loan comprehension questions

	Gatekeeping IBM	Compre IBM1	Compre IBM2	Compre IBM3
Correct	100%	42%	76%	68%
Neither Correct nor Incorrect	0%	49%	18%	24%
Wrong	0%	9%	6%	8%
Total	**100%**	**100%**	**100%**	**100%**

** Scores in this table have been rounded to whole numbers, so the total is not necessarily exactly 100%*

GatekeepingIBM. What does the "Consolidated Risk Markers" value need to be (approximately)?

The correct answer was "72". All participants answered correctly. The undefined term "Consolidated Risk Markers" was taken directly from the case study.

CompreIBM1. Why did Jason receive this explanation?

Answers that mentioned that Jason's application for a loan had been denied and that Jason wanted to know why were deemed correct (42% of answers). Answers that mentioned why Jason's application for a loan was denied or merely that it had been denied were deemed to be neither correct nor incorrect (49% of answers). Answers that did not mention the denial of Jason's application, or were incomprehensible or irrelevant, were deemed to be wrong and considered to have been offered in bad faith (9% of answers).

CompreIBM2. What does Jason need to do next to be accepted for the loan next time?

Answers that specified what needed to be improved were marked as correct (76% of answers). Answers that mentioned the concepts of "improve", "raise", or "increase" were marked as neither correct nor incorrect (18% of answers). Answers that did not mention what he should improve or that were incomprehensible were deemed to be wrong and have been given in bad faith (6% of answers).

CompreIBM3. What was the main reason Jason was rejected?

Answers that mentioned Consolidated Risk Markers were deemed correct (68% of answers). Answers that mentioned some other reason why Jason was rejected were deemed to be neither correct nor incorrect (24% of answers). Answers that were incomprehensible were deemed to be wrong and were considered to have been given in bad faith (8% of answers).

6.3.1.2 The Centrelink Case Study Comprehension Questions

Table 6-2 Type of Answers to Centrelink comprehension questions

	Gatekeep Centrelink.	**Compre Centrelink1**	**Compre Centrelink2**	**Compre Centrelink3.**
Correct	99%	91%	84%	75%
Neither Correct nor Incorrect	0%	1%	9%	14%
Wrong	1%	8%	7%	12%
Total	**100%***	**100%***	**100%***	**100%***

** Scores in this table have been rounded to whole numbers, so the total is not necessarily exactly 100%*

GatekeepCentrelink. How much money did the recipient of the letter earn from "Super Sparkle Cleaning"?

This question referred to one of the example jobs mentioned in the Centrelink letter. The correct answer was $400. Most participants got this answer correct (99%).

CompreCentrelink1. If the recipient of the letter does not respond to the letter, will they have a debt with Centrelink?

Answers that stated that the recipient of the letter would have a debt were marked correct (91%). One answer was marked as neither correct nor incorrect because it correctly identified that the recipient of the letter would be in debt but was mistaken about who would be responsible for it (1% of answers). Answers that stated that the recipient of the letter would not have a debt with Centrelink were marked as wrong (8% of answers). While these answers were not necessarily given in bad faith, they indicated that the participants had not read the case

study closely. Participants who gave this answer were removed from the results. Clearly, they were not qualified to comment on the case study.

CompreCentrelink2. What does the recipient of the letter need to do next?

Answers that mentioned that the recipient of the letter must update or confirm their Centrelink information were marked as correct (84% of answers). Answers that mentioned only responding to the letter but nothing else were marked as neither correct nor incorrect (9% of answers). Answers that were incomprehensible or did not mention responding were marked as wrong (7% of answers).

CompreCentrelink3. Why did the recipient of the letter receive this letter?

Answers that mentioned the difference between tax records and Centrelink records were marked as correct (75% of answers). Answers that mentioned that the recipients of the letter needed to update their information were marked as neither correct nor incorrect (14% of answers). Answers that were incomprehensible or did not mention that recipients of the letter needed to update their information were marked as wrong (12% of answers).

Table of All Responses Given Marked "Wrong"

Table 6-3 Table of Responses from participants marked 'wrong'

Number of 'bad faith' answers	Number of Participants	Percentage of Participants
0	90	76%
1	16	14%
2	3	3%
3	4	3%
5	3	3%
6	2	2%
Grand Total	**118**	**100%***

** Scores in this table have been rounded to whole numbers, so the total is not necessarily exactly 100%*

Participants could give at most eight answers in bad faith. A response was regarded as being in bad faith if it did not make sense, used insulting language, or was factually wrong. A small number of MTurk participants (2%) randomly answered questions to complete the XQ Survey quickly and get paid (Paolacci et al., 2010). Most participants (76%) did not have answers

marked as in bad faith (see Table 6-3). A small minority of answers (5%) had five or more answers marked as in bad faith. The bad faith answers were discarded and not included in the results.

6.3.2 Participants' Demographics

A total of 118 participants, in two batches, 69 in Batch One and 49 in Batch Two, took part in this survey.

Most participants were in their 40s (37% in their 40s). Declared ages were rounded to the nearest decade (Figure 6-1). Participants were able to keep their age private by answering '0' to the question asking their age.

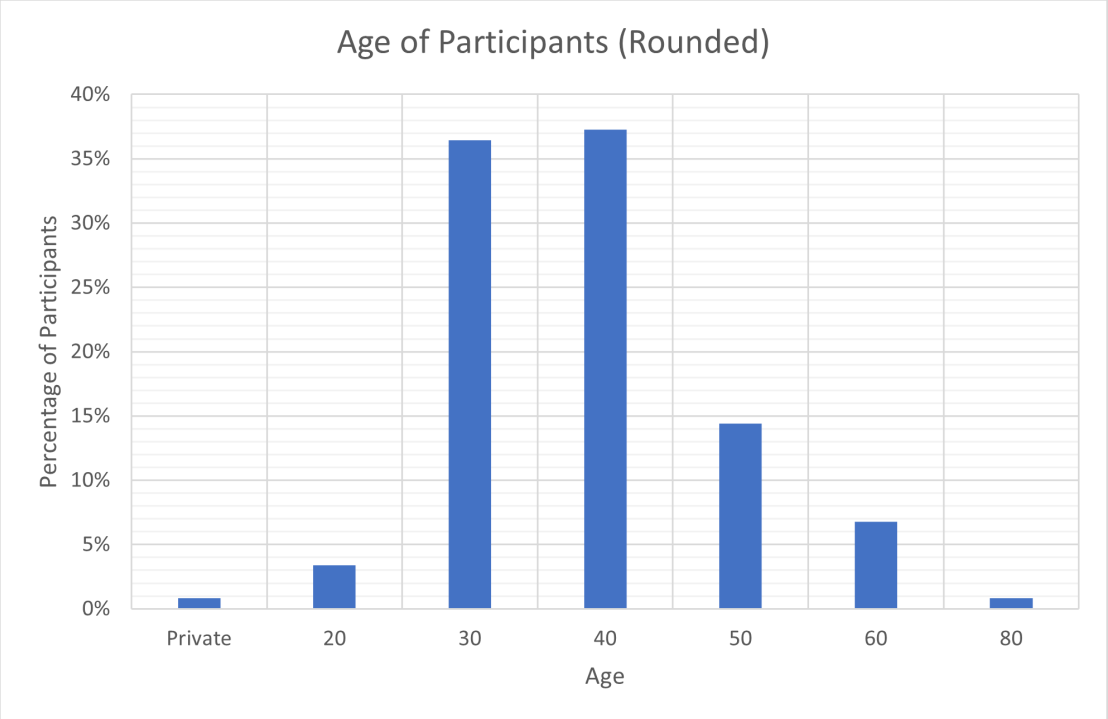

Figure 6-1 Chart of Participant's Age (Rounded)

Most participants were American (USA) (79% of participants). Indians formed 17% of participants (Table 6-4).

Table 6-4 Participants by Country

Country	Participants
Private	1%
India	17%
Ireland	1%
Nigeria	1%
Thailand	1%
United Kingdom	1%
United States of America	79%
Total	**100%***

** Scores in this table have been rounded to whole numbers, so the total is not necessarily exactly 100%*

Participants gave their own job titles. The Australian Bureau of Statistics classification system was used to code these into employment categories (Australian Bureau of Statistics, 2009). Because IT experience might have affected participants' answers, participants who worked in an IT-related occupation were placed in their own category: "*IT Workers*".

Most participants (Table 6-5) that did not work unknown jobs (22%) were 'IT Workers' (14%), 'Managers' (14%) or 'Professionals' (14%).

Table 6-5 Participants in Each Job Category (by batch)

Job category	Participants
Clerical and Administrative Workers	6%
Community and Personal Service Workers	3%
IT Workers	14%
Machinery Operators and Drivers	1%
Managers	14%
Private	6%
Professionals	14%
Sales Workers	11%
Technicians and Trades Workers	8%
Unknown	22%
Grand Total	**100%***

** Scores in this table have been rounded to whole numbers, so the total is not necessarily exactly 100%*

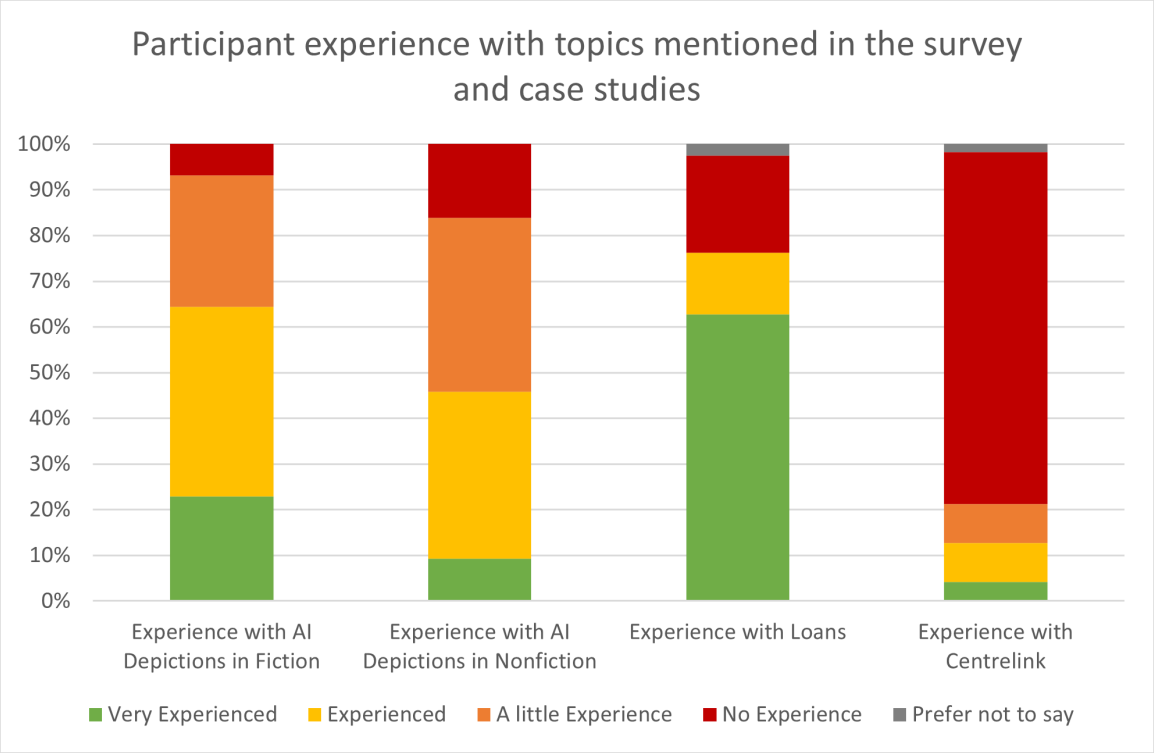

Figure 6-2 Chart of Experience Levels

Most participants (77%) were not familiar with the Centrelink 'RoboDebt' scandal (Figure 6-2). This was to be expected; no members of the participant group were from Australia.

Most participants (63%) declared that they were very experienced with loans (Figure 6-2). Most participants (42%) were at least a little familiar with pop-cultural depictions of AI and most (46%) had read or watched non-fiction discussion of AI (Figure 6-2).

6.3.3 Case Study 1 – IBM Automated Loan Denial Explanation

A discussion of this case study may be found in Section 4.6.5. The case study is in Appendix C2a.

It was important to remove responses from participants who gave answers in bad faith. These responses are not reliable and do not reflect the true value of the explanation.

Figure 6-3 Unguided Score Boxplot for IBM Loan Case Study

Figure 6-3 shows that participants who answered comprehension questions correctly rated the IBM case study higher, on average. The range of marks was between 1 and 10 in the Unguided Score Section (Table 6-6).

Table 6-6 Unguided Score Descriptive Statistics

	N	Mean	SE Mean	StDev	Minimum	Median	Maximum
Unguided Score	118	7.153	0.213	2.315	1	8	10

The MTurk participants found the explanation clear (F01C00) and convincing (F03C03) (Table 6-7). However, they would have preferred better use of references (F02C01) and more definitions (F01C02) (Table 6-7).

Table 6-7 Most Common Answers in MTurk Surveys for the IBM Loan Surveys

Question Code	Question	Most Common Answer	Percentage of Participants
F01C00	Is the explanation clear?	The explanation is clear	44%
F01C01	Is the explanation written with correct spelling and grammar?	The explanation has no grammatical errors and no spelling mistakes	86%
F01C02	Does the explanation define the uncommon words and symbols it uses?	All unusual and uncommon words and symbols are defined	28%
F01C03	Is the supporting information (or detail) well-presented and understandable?	The information supplied to support the explanation is clearly presented and understandable	45%
F01C04	Can the explanation be acted upon?	It is not clear how to act upon the explanation, but it can be acted upon	40%
F02C01	Are references provided?	No references are provided even though there are claims that need references	35%
F02C02	Are relevant facts mentioned in the explanation?	There are frequent references to relevant facts	41%
F03C01	Is the question (or implied question) answered?	The explanation answers the question (or implied question)	44%
F03C02	Can the explanation's reasoning be followed without difficulty?	The reasoning of the explanation is laid out clearly, and it is possible to follow the argument	50%
F03C03	Is the explanation convincing?	The explanation is convincing	51%

F01C02 was the most inconsistently answered question (Table 6-7 and Table 6-8). The MTurk participants did not agree on any answer.

Table 6-8 Responses to Question F01C02 (IBM loan case study)

F01C02. Does the explanation define the uncommon words and symbols it uses?	**Percentage of Participants**
No definitions or keys were needed	23%
All unusual and uncommon words and symbols are defined	28%
Only some words and symbols are defined	25%
No words and symbols are defined, even though the explanation is unclear without them	25%
Total	100%

The FICO Explainable Machine Learning (XML) Challenge (2019) gave this case study a winning score (FICO Community, 2019). Its marking guidelines define the expected user of any explanation created for the challenge as "*a data scientist with the domain knowledge*" (FICO

Community, 2019). However, the briefing text introducing the IBM section chosen to be Case Study 1 defines the user as *"a bank customer*". This definition implies that a non-expert should understand it. However, a common complaint from MTurk participants was that some words and phrases used to explain the AI's decision were not clearly defined (Table 6-8). This matches similar concerns about this case study raised by the focus group participants (Table 6-9).

Table 6-9 Responses from Focus Group Part 1

Clear wording and unfamiliar terms/symbols defined	No. participants who selected this option
Glossary included if new words are introduced. The wording is clear	0
The wording is clear, and the terms/symbols used are standard. Uncommon words are defined	1
The wording is clear and uncommon words are defined	2
Terms introduced without introductions, unusual symbols used and not defined	2
Grand Total	**5**

Given that the MTurk experiment evaluated only one part of IBM's winning entry, IBM's win in the FICO XML Challenge (2019) does not invalidate the scores. This gap in marking shows that the evaluator's assumptions and background may have a considerable effect on how the explanation is received.

6.3.4 Case Study 2 – The Centrelink Automated Assessment and Letter

A discussion of this case study is in Section 4.6.6. The case study itself is in Appendix C2b.

Like Case Study 1, in Case Study 2, participants who answered the questions correctly, on average marked the case study higher (Figure 6-4).

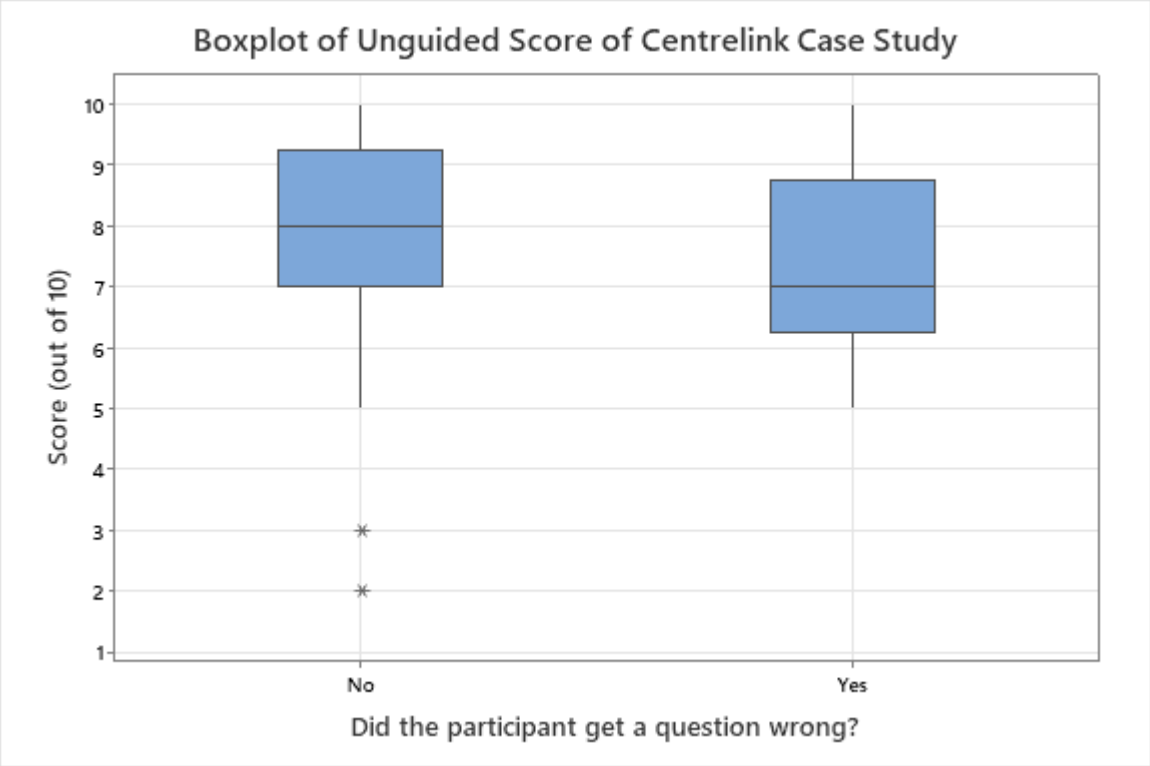

Figure 6-4 Boxplot of Centrelink Unguided Score

Table 6-10 shows that MTurk participants used the XQ Rubric to rate it no less than 2 out of 10 and gave it an average rating of 8. This score is much higher than expected, given the hostile Ombudsman's report and the critical media coverage of the letter (Farrell & McDonald, 2019). Table 6-11 shows that the rubric responses show the same trend.

Table 6-10 Marks of the Centrelink Letter

	N	Mean	SE Mean	StDev	Minimum	Median	Maximum
Unguided Score of the Centrelink Letter	118	7.9	0.2	1.7	2	8	10

Table 6-11 Table of rubric responses to the Centrelink case study

Question Code	Question	Most Common Answer	Percentage of Participants
F01C00	Is the explanation clear?	The explanation is clear	56%
F01C01	Is the explanation written with correct spelling and grammar?	The explanation has no grammatical errors and no spelling mistakes	86%
F01C02	Does the explanation define the uncommon words and symbols it uses?	All unusual and uncommon words and symbols are defined	44%
F01C03	Is the supporting information (or detail) well-presented and understandable?	The information supplied to support the explanation is clearly presented and understandable	63%
F01C04	Can the explanation be acted upon?	It is clear how to act upon the explanation and there are instructions for the reader on what to next	81%
F02C01	Are references provided?	All claims that need it are referenced	69%
F02C02	Are relevant facts mentioned in the explanation?	There are frequent references to relevant facts	58%
F03C01	Is the question (or implied question) answered?	The explanation answers the question (or implied question)	67%
F03C02	Can the explanation's reasoning be followed without difficulty?	The reasoning of the explanation is laid out clearly, and it is possible to follow the argument	61%
F03C03	Is the explanation convincing?	The explanation is convincing	69%

There are at least three reasons why the XQ Rubric was rated so highly by MTurk participants:

1. MTurk participants were not the target audience
2. The letter was presented differently to MTurk participants
3. The media and Royal Commission discussed all aspects of the debt recovery scheme, not just the letter

The Royal Commission criticised the letter's accessibility, suggesting that vulnerable people might find it difficult to understand (Commonwealth Ombudsman, 2017). Even among the MTurk participants, 17% of all participants did not correctly identify critical aspects of the letter (Figure 6-5). MTurk participants in the second batch, which only contained participants rated 'Master', answered more questions correctly than participants in the first batch.

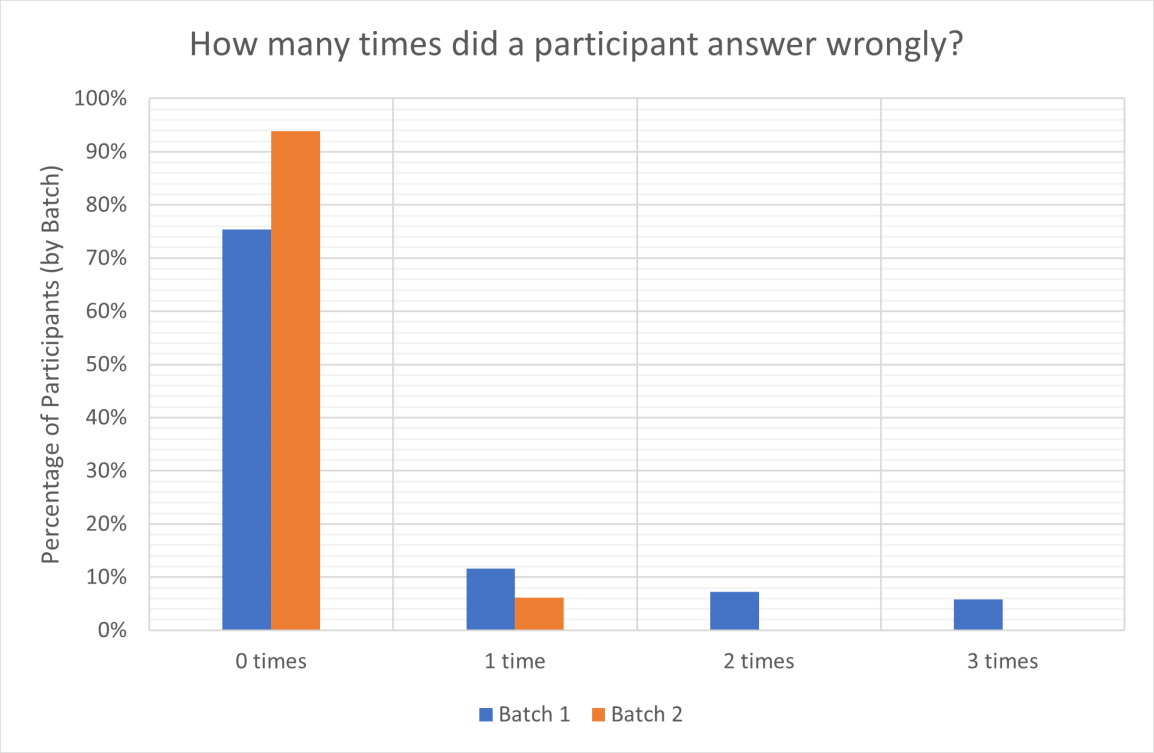

Figure 6-5 Distribution of participants who answered comprehensions wrongly (by batch)

By the very nature of the MTurk survey, the Centrelink letter was presented differently to MTurk participants than it had been to Centrelink customers. The letter did not come as a surprise in the mail, for example, and MTurk participants had no reason to be shocked by its contents, unlike the letter's recipients. MTurk participants were given a form letter, not a letter addressed to them directly. The MTurk participants were not the target audience. No MTurk participants were from Australia, nor were any MTurk participants in danger of getting into debt with Centrelink. This difference in presentation may also have allowed the MTurk participants to view the letter more dispassionately.

MTurk participants were mainly employed in jobs that require a good education (see Section 6.3.2), and some technological literacy and broader education is needed to participate in MTurk surveys. Centrelink customers did not necessarily share these characteristics.

When examining the media commentary and later the Ombudsman's report on the "RoboDebt" fiasco, it is important to remember they concern all aspects of the debt recovery scheme, not just the letter. The explanation in the letter may be reasonable, but the added demands, such as

proof of employment from previous years, were difficult to meet for some people. The entire scheme was not user-tested properly, and the people affected did not receive adequate advice and guidance about the new compliance system.

6.3.5 Conclusion

In our evaluation survey, the average mark of Case Study 1 was considerably lower than that of Case Study 2. The difference could be that more participants were familiar with loans (Section 6.3.2) and were more likely to be critical of what was familiar (Norton, Frost, & Ariely, 2007). However, the average mark of Case Study 1 could be lower for a simpler reason: it is just a worse explanation, though this is unlikely considering Case Study 1 won an award and Case Study 2 was heavily criticised by the Australian media.

6.4 Discussion

6.4.1 Evaluating the XQ Survey

Moskal and Leydens (2000) stressed the need for a reliable and valid rubric. A reliable rubric will generate the same score whoever the marker, and a valid rubric reflects the quality of the explanation assessed in its scoring. Very little research has been done on the validation of rubrics, though interestingly, Arcuria, Morgan, and Fikes (2019) recently developed a method to validate rubric scores, comparing the scores given to the same work by rubrics and the scores given by a panel of experts.

6.4.2 The Experiment's Results

These MTurk experiments were designed to improve the XQ Rubric's capabilities as an evaluation method. They showed that the XQ Rubric was reliable, intuitive, and adequately represented participants' feelings about the case study.

This experiment evaluated two case studies: the IBM loan and the Centrelink Letter case studies. The IBM automated loan denial case study had scored well in the FICO XML Challenge (2019), so the critical responses from the participants were not expected. Only half the participants found the explanations convincing and easy to follow. 35% of participants regarded the number of references as insufficient.

The FICO Challenge did not release its marking for the IBM loan denial case study, so it is possible that problems detected by participants were also detected by the FICO Challenge. Feedback about the IBM case study was used to revise the case study for use in a future experiment (Chapter 7).

Feedback from participants on the Centrelink RoboDebt case study was less critical than was expected. Most participants (86%) believed that the explanation's grammar and spelling were good. Participants regarded the explanation as easy to understand, and they were clear about what action should be taken next. However, the Ombudsman's report criticised the clarity of the letter and the difficulty of understanding it. The differences in the reception of the letter may be both due to the manner in which people received it and to the socio-economic circumstances of the recipients (Section 6.3.2).

6.4.3 MTurk Recommendations

Participants designated as "Masters" were more reliable than ordinary participants, this is demonstrated in Figure 6-5. Comprehension questions helped to filter out answers given in bad faith and the answers of participants who did not understand the explanations.

7 VALIDATION OF XQ SURVEYS AS A TOOL FOR CONSTRUCTIVE FEEDBACK

7.1 INTRODUCTION

This experiment was one of three experiments that followed the last four stages of the methodology of Peffers et al. (2007) (Section 4.1).

This Validation Experiment was designed to show that the XQ Rubric was more than an evaluation method. It could also be used to give feedback about an explanation and improve its quality. These experiments provided more feedback, more data about the XQ Rubric, and a better understanding of what categories were important to evaluators.

When the revised explanation was evaluated, it was expected that it would score better than the original on both the unguided and rubric scores. The experiments in Chapter 6 were extended to test the hypothesis that the explanation would score better on the XQ Survey if evaluated and revised in the light of previous survey results.

The Ethics Approval for this experiment is presented in Appendix B3.

7.2 METHODOLOGY

These experiments were designed to test the hypothesis that a revised explanation would score higher than the original explanation on the unguided and rubric scores. One case study was chosen and revised using feedback from the Initial Validation experiment (Chapter 6). The case study was the IBM loan case study (Appendix C3) (IBM Research, 2019). This is further discussed in the methodology chapter (Section 4.7.2.2).

The revised case study was presented to newly recruited MTurk participants. Each MTurk participant saw one case study. They were paid US$3.00. Since the survey was completed more

quickly this was a higher hourly rate. The results were analysed to determine whether the participants ranked the revised explanations higher than the original explanations.

The XQ Survey also had some minor improvements. The most significant change was the addition of drop-down lists in the demographics section, designed to standardise the job descriptions given by participants. Participants were also asked to justify their unguided score.

7.2.1 Improvements

This section details the improvements made to the case study based on feedback from the Focus Group experiment and the Initial Validation experiment.

The IBM loan case study was previously evaluated by the Focus Group and MTurk users. It was chosen for this experiment because it had been used previously, and there was clear feedback about how it could be revised. Ignoring focus group feedback, the case study was revised only on feedback from the previous MTurk experiments. To improve the case study, the meaning of "*consolidated risk markers*" was explained.

Placeholder contact information was added, and a placeholder graph illustrating which consolidated risk factors influenced the decision of the AI to reject Jason's loan application was also included.

See Appendix C3 for the updated IBM loan case study.

7.2.1.1 Survey Improvements

The XQ Survey itself was based on the survey used for the IBM case study in Chapter 6. For demographic data, a drop-down list replaced a text box. This was to standardise the participants' job description and location by having them to choose the most correct option. The change did not affect the evaluation of the case studies because the demographic information was requested after the case study evaluation had been completed.

A question was added asking participants to justify their unguided score rating. This addition should not affect the evaluation of the case study because it sought only to clarify the unguided score, not adjust it.

7.3 Results of the IBM Loan Case Study

So that comparisons could validly be made between the original case study and the revised case study, it needed to be established that the demographic make-up of the two survey groups was similar. If the two groups were significantly different, then changes in the score might be attributable to factors not related to the case study. Special attention was given to the demographic make-up of the groups.

In this section, Round 1 refers to the survey answers associated with the original case study. Round 2 refers to the survey answers associated with the revised case studies.

These results compare the evaluation of the original IBM loan case study in the first MTurk experiments with the evaluation of the revised IBM loan case study.

7.3.1 Demographics

The median age of the participants was 40 years (rounded to the nearest 10), the same in both rounds (Table 7-1). Two participants chose the age of 0 to keep their age private (as permitted by the XQ Survey instructions).

Table 7-1 Descriptive Statistics for Age (rounded)

Round	N	Mean	SE Mean	StDev	Minimum	Median	Maximum
1	117	37.7	0.9	10.3	20	40	80
2	50	40.3	1.8	12.6	20	40	80
Total	**167**						

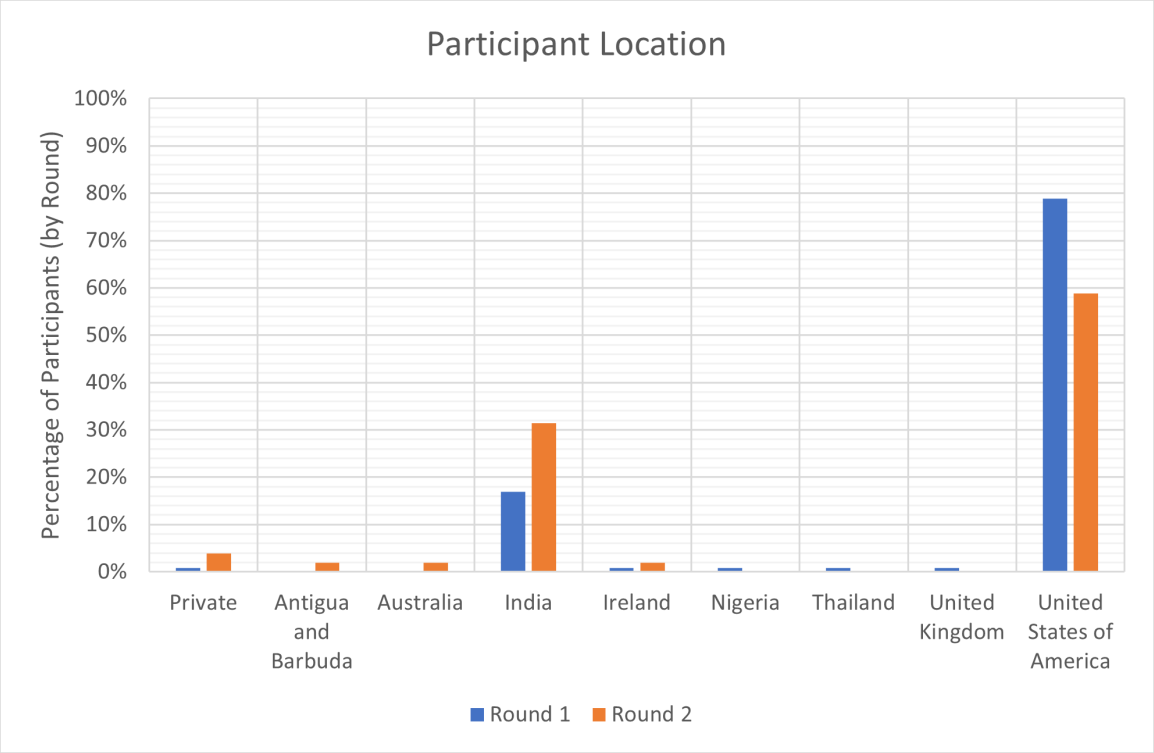

Figure 7-1 Location of Participants

The most common country for participants to be located in both rounds was the United States of America (72% of all participants, Figure 7-1). The second most common was India.

The most common occupation type was 'manager' when the rounds were counted together (20% of the population, Table 7-2). However, in the Initial Validation Experiment of the XQ Survey, participants were asked to describe their job type rather than choose from a drop-down list (the XQ Survey used in this chapter). As a result, many participants in Round 1 were categorised "unknown" (22% of the population). In Round 2, when participants chose their job type from a drop-down list, 'manager' was the most common job type (31% of the population).

Table 7-2 Job Titles of Participants

Row Labels	Round 1	Round 2	Total of Both Rounds
Managers	14%	31%	20%
Professional	14%	20%	16%
Unknown	22%	0%	15%
IT Workers	14%	16%	15%
Sales Workers	11%	4%	9%
Technicians and Trades Workers	8%	10%	8%
Clerical and Administrative Workers	6%	12%	8%
Private	6%	0%	4%
Community and Personal Service Workers	3%	0%	2%
Private	0%	6%	2%
Labourers	0%	2%	1%
Machinery Operators and Drivers	1%	0%	1%
Total	**100%***	**100%***	**100%***

** Scores in this table have been rounded to whole numbers, so the total is not necessarily exactly 100%*

In both surveys, most participants agreed with the statement: "I sometimes consume entertainment that features AI" (41% of the population in Round 1, 49% of the population in Round 2, Table 7-3).

In both surveys, most participants had a little experience with AI non-fiction, such as academic articles and newspaper reports (38% of the population in Round 1 and 45% in Round 2, Table 7-3).

In both surveys, most participants were very experienced with loans (60% of the Round 1 population and 54% of the Round 2 population, Table 7-3).

Table 7-3 Participant Experience

	AI Pop Culture		AI Practical Matters		Financial Experience	
	Round 1	**Round 2**	**Round 1**	**Round 2**	**Round 1**	**Round 2**
Very Experienced	23%	12%	9%	12%	63%	55%
Experienced	42%	49%	36%	22%	14%	24%
A little Experience	29%	31%	38%	45%	0%	0%
No Experience	7%	8%	16%	22%	21%	16%
Prefer not to say	0%	0%	0%	0%	3%	6%
Total	**100%***	**100%***	**100%***	**100%***	**100%***	**100%***

** Scores in this table have been rounded to whole numbers, so the total is not necessarily exactly 100%*

7.3.2 Comprehension Questions

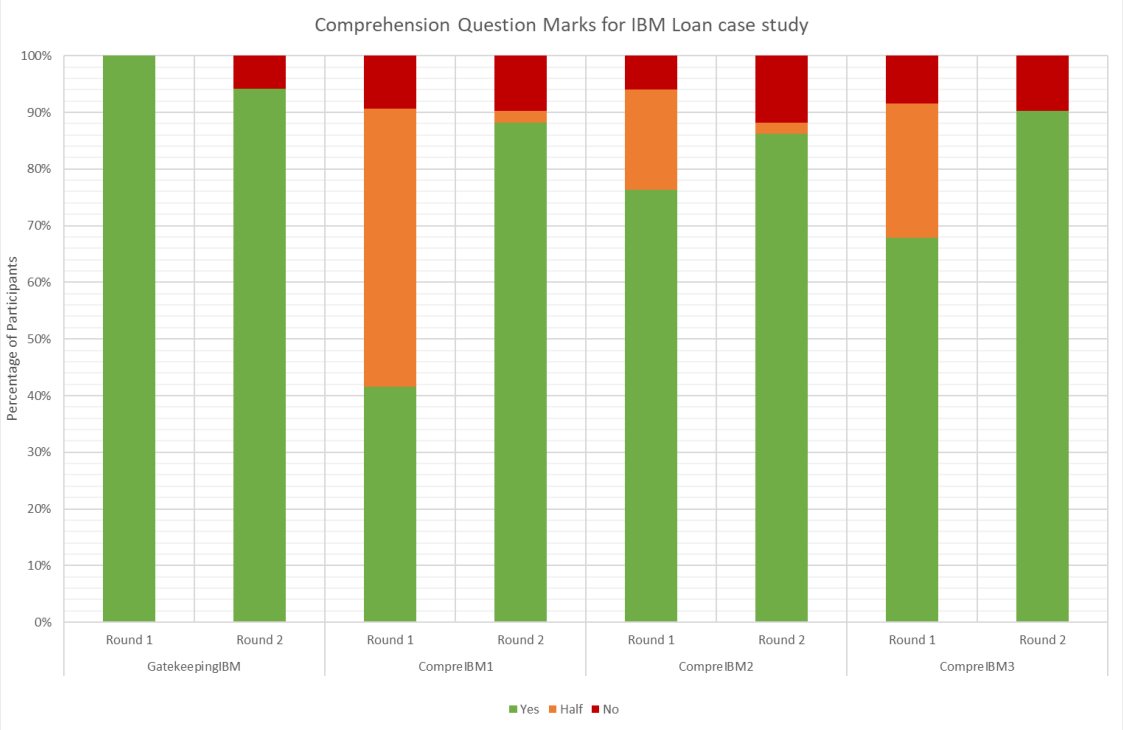

Figure 7-2 Comprehension Question Marks for IBM loan case study

The comprehension questions for Round 2 were the same as those of Round 1 (Section 6.3.1). Figure 7-2 shows the distribution of Comprehension Question marks for the IBM loan case study. While the number of participants who answered comprehension questions wrongly increased between Round 1 and Round 2 in every question, the number of participants who got questions half-right decreased. This indicates a better case study, for participants who got the question half-right had obviously misunderstood something. Since there were fewer half-right answers, then a higher proportion of participants understood the case study. The questions themselves were not changed.

7.3.3 Unguided Score

Table 7-4 Unguided Score Marks for IBM Loan Case Study

Round	N	SE Mean	StDev	Minimum	Median	Max	Mean	Bootstrapped 95% CI
1	118	0.2	2.3	1	8	10	7.15	6.75 - 7.57
2	51	0.3	1.9	1	8	10	7.69	7.20 - 8.22

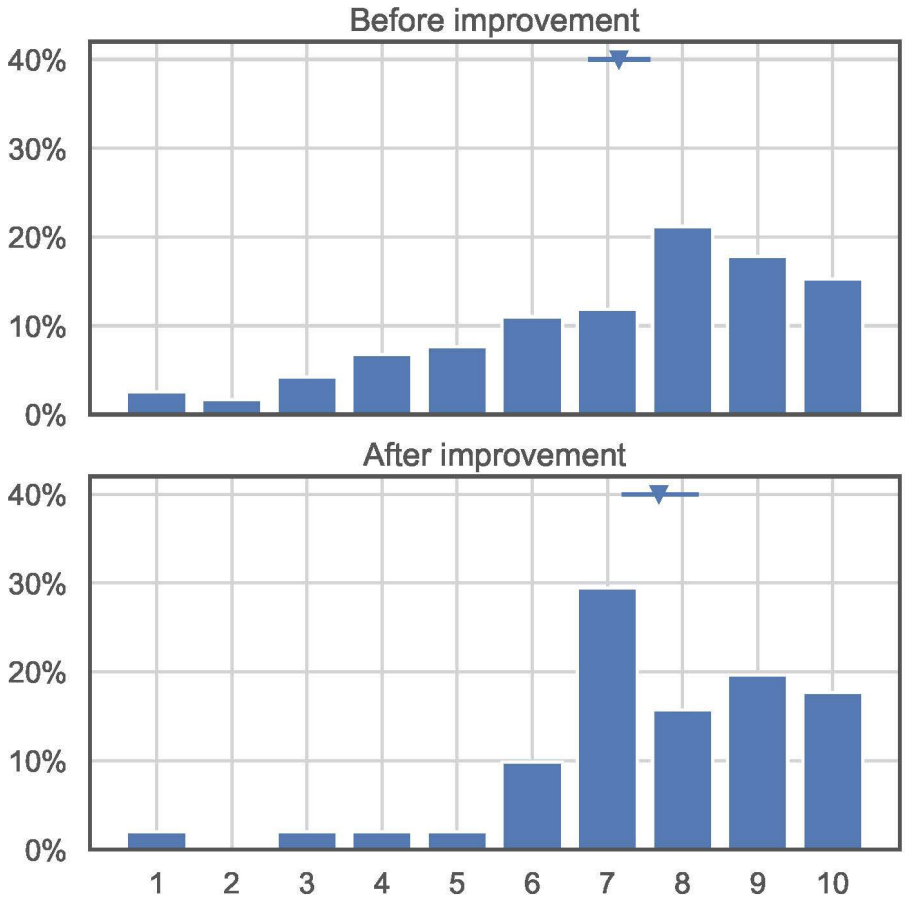

Figure 7-3 Histogram of Unguided Score for the IBM case study

Figure 7-3 is a histogram of the unguided score for the IBM case study. A bootstrapping technique was used to determine a 95% confidence interval (Bootstrapped 95% CI) for the Unguided Score. Since there is some overlap between the two bootstrapped confidence intervals, it cannot be determined if the improvements made to the case study made a significant difference to the Unguided Score.

7.3.4 Rubric Evaluation

Question Code	Question	Most Common Answer in Batch 1	Percentage in Batch 1	Most Common Answer in Batch 1	Percentage in Batch 2
F01C00	Is the explanation clear?	The explanation is clear	44%	The explanation is clear	53%
F01C01	Is the explanation written with correct spelling and grammar?	The explanation has no grammatical errors and no spelling mistakes	86%	The explanation has no grammatical errors and no spelling mistakes	96%
F01C02	Does the explanation define the uncommon words and symbols it uses?	All unusual and uncommon words and symbols are defined	28% *See Section 6.3.2 for more detail*	All unusual and uncommon words and symbols are defined	47%
F01C03	Is the supporting information (or detail) well-presented and understandable?	The information supplied to support the explanation is clearly presented and understandable	45%	The information supplied to support the explanation is clearly presented and understandable	61%
F01C04	Can the explanation be acted upon?	It is not clear how to act upon the explanation, but it can be acted upon	40%	The explanation can be acted upon	47%
F02C01	Are references provided?	No references are needed to support this explanation, and so no references are provided	51%	No references are needed to support this explanation, and so no references are provided	57%
F02C02	Are relevant facts mentioned in the explanation?	There are frequent references to relevant facts	41%	There are frequent references to relevant facts	47%
F03C01	Is the question (or implied question) answered?	The explanation answers the question (or implied question)	44%	The explanation answers the question (or implied question)	61%

Question Code	Question	Most Common Answer in Batch 1	Percentage in Batch 1	Most Common Answer in Batch 1	Percentage in Batch 2
F03C02	Can the explanation's reasoning be followed without difficulty?	The reasoning of the explanation is laid out clearly and it is possible to follow the argument	50%	The reasoning of the explanation is laid out clearly and it is possible to follow the argument	51%
F03C03	Is the explanation convincing?	The explanation is convincing	51%	The explanation is convincing	65%

In Round 1, 25% of participants were dissatisfied with the definitions. This significantly decreased, by 8%, in Round 2 (Chi-Squared test, $p < 0.05$).

In Round 1, most participants were satisfied with the supporting information. Overall less than 12% were dissatisfied with the supporting information.

In Round 1, 40% of participants were not clear about how to act on the information, though they believed they could. In Round 2, 47% of participants felt they could act on the information.

There was a large improvement in the participants' responses in the matter of references. 34% of participants in Round 1 thought there was a significant number of missing references. In Round 2, 37% of participants were satisfied.

Most participants were satisfied with the references to relevant facts. Less than 13% overall were dissatisfied.

In Round 1, 20% of participants felt that the explanation answered only part of the question. In Round 2 this decreased to 8%, a large improvement.

These results show that improvements made to the loan case study based on feedback from the previous survey made a substantial difference to the quality of the explanation. (The

demographic make-up of participants was similar to that of the previous loan case study, which means that these two surveys can be meaningfully compared.)

7.3.5 Survey Design

The improved design of the survey used by this experiment resulted in more precise demographic information and easier processing of data. The MTurk participants' answers were kept uniform by drop-down selection lists. The drop-down lists led to increased demographic precision.

This improvement is best demonstrated by the occupation answers. Previously the occupations were sorted and grouped using the Australian Bureau of Statistics (2009) guide. This introduced difficulties, for many participants did not answer clearly. Insufficient information was gathered for their job to be classified in the correct Australian Bureau of Statistics category. However, by asking participants to group themselves using the occupation drop-down list, they (presumably) selected the most suitable answer. This led to more precise information and easier processing.

7.4 Discussion

These experiments show that rubric feedback can be used to improve explanations. The IBM loan case study improved. The minimum scores increased most (by 20%), showing that even participants who disliked and were disapproving of certain aspects of it were more positive than those evaluating the original case study. Because the minimum rubric score also increased by the same amount as the total rubric score, the score improvement was not related to the new question which required participants to justify their unguided score.

8 Independent Validation of the XQ Rubric and the XQ Survey

8.1 Introduction

This, the final experiment, builds on the results of the experiments in Chapter 6 and Chapter 7. While these chapters demonstrated that the XQ Rubric and Survey could be used to evaluate and improve explanations, the experiments described were conducted by the author of this thesis. To remove biases that this might introduce, a series of experiments were conducted with an independent collaborator, Dr Francisco Cruz of Deakin University. Dr Cruz was not involved in the creation and early exploration of the XQ Rubric and Survey. This round of experiments was designed to evaluate and gather feedback on the use of the XQ Rubric and Survey by someone not involved in their design, testing, and implementation.

These experiments demonstrated that someone who had not been involved in the design and creation of the methodology could nevertheless use it to evaluate and improve XAI explanations. The XQ Rubric is meant for public use by people unfamiliar with its origin and workings: this was a test of its capabilities in that regard.

The experiment was part of a set of experiments that followed the last four stages of the methodology of Peffers et al. (2007) (Section 4.1), demonstrating and evaluating the final versions of the XQ Rubric and XQ Survey.

The Ethics Approval for this experiment is presented in Appendix B4.

8.2 Methodology

The experiment had two rounds, both designed to study the capacity of the XQ Rubric to improve the target explanation. The experiments were performed in five stages:

1. development of the experiment

2. MTurk Round 1
3. processing data and improving explanations
4. MTurk Round 2
5. reprocessing the data

These stages follow the methodology outlined by Peffers et al. (2007) (see Chapter 4).

8.2.1 Stage 1 – Development of the Experiment

In the first stage of the experiment, I approached Dr Cruz, proposing a research collaboration. His research concerned explainable reinforcement learning (XRL) methods for application in scenarios requiring collaboration between an autonomous robot and its human team-mate (Bignold, Cruz, Dazeley, Vamplew, & Foale, 2022; Cruz, Dazeley, Vamplew, & Moreira, 2021). These explored the use of explanations in a scenario where a human and an autonomous drone collaborated to reveal the location of an opponent in an environment divided by barriers.

The scenario given to participants in this experiment was a variant of a scenario Dr Cruz had developed for an unrelated AI-explanation research project. It shows, in essence, a game of hide-and-seek. The green-shirted person works with the drone. Together they look for the red-shirted person. An example of one of the scenarios is shown in Figure 8-1.

This scenario was selected because hide-and-seek is a familiar childhood game, and participants would be able to draw upon their personal experiences. Participants were also expected to be familiar with the capabilities of consumer quadcopter drones.

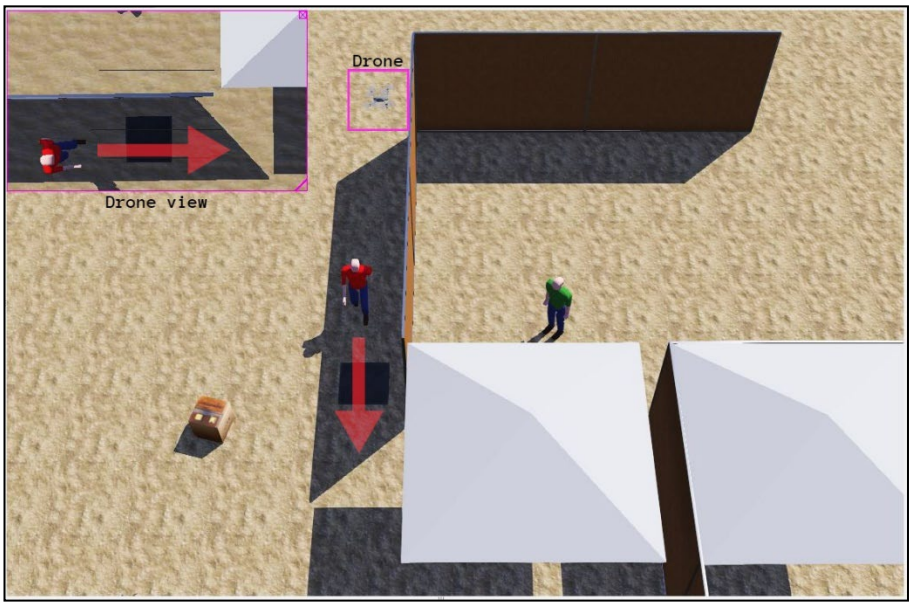

Figure 8-1 Image of the Scenario as presented to participants

Dr Cruz wrote a series of explanations based on six permutations of the scenario described above. While an XAI did not generate the explanations, they represent the explanatory style of an XAI. The hide-and-seek case studies may appropriately be investigated as examples of the use of the XQ rubric by independent researchers.

The explanations were used to create six different case studies for use in each round of this experiment. Dr Cruz and I then developed comprehension questions (GatekeepingDrone0, CompreDrone1, CompreDrone2, and CompreDrone3 (Appendix E4a)) to check whether survey participants understood the case study. Dr Cruz was included in the creation of the comprehension questions because it was intended that, where possible, this experiment would be based on his decisions.

While Dr Cruz has occasionally collaborated with my thesis supervisors, he is not a member of my supervision team and had no input in the development of the XQ Rubric and XQ Survey. In

these experiments he had the role of the independent XAI researche. He followed my methodology for creating and posting XQ Surveys on MTurk.

The original case studies used in this experiment are in Appendix C4a.

8.2.2 Stage 2 – Round 1 MTurk Experiments

This round of the experiment established a baseline assessment for Dr Cruz's original explanations and gathered feedback about improving them. To eliminate bias, participants who had participated in previous MTurk surveys were excluded from taking part.

This first batch of surveys had 92 participants out of a possible 120. After some days, fewer new participants joined the survey. To remedy this, I uploaded the job posting again to attract more participants. Nothing else was changed. The second batch completed 28 out of 28 surveys.

While 2 hours were allotted per worker, Dr Cruz expected that a survey would take about 20 minutes to complete, making the pay rate US$7.25/hr, an amount in line with the payment standards of the other experiments.

8.2.3 Stage 3 – Processing Data and Revising Explanations

After Round 1 of MTurk evaluation had run, the results and data were downloaded from the survey platform. From this data, bar charts were generated. These were given, with the data, to Dr Cruz to process and improve his explanations. To avoid bias, Dr Cruz carried out the XQ Rubric data analysis and modified the explanations without my input.

The spelling and grammar of the case studies had been identified by the participants as areas for potential improvement. After Dr Cruz had revised the explanations, they were reviewed and corrected by an independent proof-reader. Dr Cruz approved the changes.

After the explanations had been updated in response to the feedback of participants, the MTurk experiment was re-run with a new set of participants, using the same survey procedures.

The revised case studies used in this experiment are in Appendix D4b.

8.2.4 Stage 4 – Round 2 MTurk Experiments

After uploading the XQ Survey with the revised case studies to MTurk on 28 November 2021, the MTurk interface was monitored to confirm that the script randomly assigning surveys (Appendix F2) was working correctly and that MTurk workers were accepting the job.

I increased the pay from US$2.00 to US$2.50 for Round 2 of MTurk surveys, hoping that this would be a more attractive pay rate. The acceptance rate lagged again, however, so the job posting was uploaded again with the pay increased from US$2.50 to US$3.00 in order to attract more participants. Nothing else was changed.

This increase in the pay rate is not likely to have induced participants to evaluate the case studies more favourably. A study by "*Buhrmester, Kwang, and Gosling (2016) found that changing the pay for their personality surveys from 2 to 10 to 50 cents produced responses with nearly identical alpha values, and the only factor affected was the time for the researcher to collect a complete sample.*"(Grysman, 2015). The risk of positive bias was minor and having more participants was worthwhile. The possibility that increased pay rates may have affected the evaluation was tested during the analysis of the results (Section 8.3).

Participants from previous surveys, including participants who had participated in the previous round, were not eligible to participate in this round of the experiment. This exclusion was to prevent a participant's interpretation from being affected by their exposure to earlier case studies.

The response rate for this round was noticeably lower than that of previous experiments. The job posting was re-uploaded several times to increase its visibility. On each occasion, this resulted in a short-lived boost in the number of responses.

Case Studies 4 and 5 were evaluated fewer times than the other Case Studies in the earlier batches. To correct this, duplicate entries for Case Study 4 and 5 were added to the randomising script so that they had a higher chance of being chosen.

8.2.5 Stage 5 – Reprocessing Data

The data from the two rounds were then processed. The rounds were compared to establish whether the revised explanations had better unguided and XQ Rubric scores than the original explanations.

As the first step in this analysis, the demographics of the participant cohorts for each round were compared to make sure that there were no significant variations that might skew the observed results in the participant groups in earlier and later batches and rounds.

8.3 RESULTS

The results of this experiment are presented in four sections. The first three are similar to the three sections of the XQ Survey: the Unguided Score, the XQ Rubric, and the demographics sections. These sections are not presented in order of the XQ Survey. Before commencing the comparison, it was important that the two surveys had been shown to be comparable.

Participants were asked to complete all sections of the survey, including the demographics section, the unguided score section, and the rubric section.

8.3.1 Demographics

In Round 1, 119 people participated in the XQ Survey. In the second, 103.

A two-sample t-test was used to test whether the mean age of participants (*Age*) in each round was equal. It was found that there was no significant difference in the mean age of participants ($p > 0.05$).

A Chi-Square test was used to test if the participants' answers about their occupation, knowledge of AI in pop culture, knowledge of more academic AI matters, and their experience with drones (*Occupation*, *AIPopCulture*, *AIPracticalMatters*, and *DroneExperience)* depended on whether they had participated in Round 1 or 2. It was found that the answers were not dependent on the round in which they participated ($p > 0.05$).

Most participants of both rounds were located in the USA and India (see Figure 8-1). This did not change enough to affect the scores.

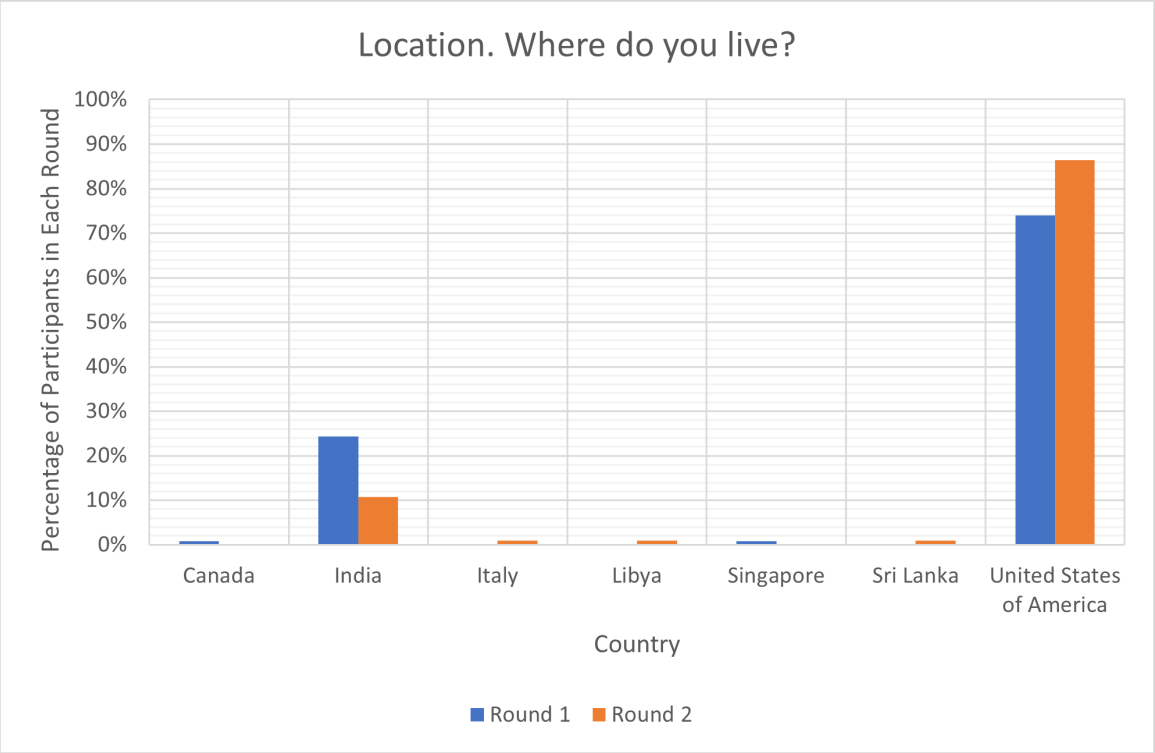

Figure 8-1 Location of Participants

In conclusion, the age, nationality, job, and other personal data for participants from both rounds were similar enough that any difference in the Unguided Score and Rubric sections was most likely due to the differences in the explanations, not their demographic makeup.

8.3.2 Unguided Score

GatekeepingDrone0: What number scenario have you been given?

Table 8-1 Results from GatekeepingDrone0

	GatekeepingDrone0

	Round 1	Round 2	All Rounds
Yes	77%	82%	79%
Half	19%	5%	13%
No	4%	14%	9%

Most participants correctly identified the case study. Participants who got the question half correct answered '2'. They were marked as half correct since they may have read the question number, which was '2', to find the scenario number thereby misinterpreting the question.

CompreDrone1: What colour shirt is the opponent team wearing?

Table 8-2 Results from CompreDrone1

	CompreDrone1		
	Round 1	**Round 2**	**All Rounds**
Yes	97%	100%	98%
No	3%	0%	2%

Participants were asked to pick the correct colour from a dropdown list. Most participants (98%) answered correctly (Table 8-5). In Round 2, all participants got the colour correct.

CompreDrone2: What colour shirt is the team with the drone wearing?

Table 8-3 Results from CompreDrone2

	CompreDone2		
	Round 1	**Round 2**	**All Rounds**
Yes	98%	97%	97%
No	3%	3%	3%

As with CompreDrone1, participants were asked to pick the correct colour from a dropdown list. However, unlike CompreDrone1, CompreDrone2 had more wrong answers in Round 2. This may have been because information about the coloured shirt worn by the drone's team is presented later in the case study. Perhaps some participants did not read far enough to obtain the information.

CompreDrone3: What game are the two humans and the drone playing?

Table 8-4 Results from CompreDrone3

	CompreDone3		
	Round 1	**Round 2**	**All Rounds**
Yes	98%	95%	97%
Half	1%	1%	1%
No	1%	4%	2%

The correct answer was “hide and seek”. I accepted all spelling and grammar variants. Those who described what was happening without using the name ‘hide-and-seek’ were marked half correct. Most participants in both rounds correctly identified what game the two humans and the drone were playing.

Comprehension Question Comparison

The XQ Survey asked participants to read the explanation and respond to four comprehension questions. This was to test their understanding of the scenario and to make sure they had read the explanations being evaluated. The number of incorrect answers to the comprehension questions remained the same for both rounds (*GatekeepingDrone0, CompreDrone1, CompreDrone2, CompreDrone3*) (two-sample t-test, $p > 0.05$). However, the proportion of partly correct answers was significantly less (two-sample t-test, $p < 0.05$). This improvement suggests that participants understood the explanation better and were better able to answer the comprehension questions. The number of wrong answers was primarily due to bad-faith responses, not to inadequate comprehension of the case study. This was demonstrated by the low number of nonsense answers.

After rating the explanation out of 10 (*RawScoreDrone)*, participants were asked to justify it (*ExplanationScoreDron*). Their ratings were grouped according to whether the participant had understood the explanation (they understood the explanation, found it hard to understand, or did not understand it). Participants who understood the explanation were further grouped into those who had positive comments about the explanation, those who criticised aspects of the

explanation, and those who did not comment on the explanation content (*Positive Comment and Understood Explanation, Criticism and Understood Explanation, Understood Explanation*).

Participants' comments that could not be categorised by this system were categorised instead according to the content of the answers. If the answer was given in bad faith, it was categorised as "*Nonsense*". If the answer was given in good faith but its meaning was unclear, it was marked "*Unclear*". Finally, if the answer merely restated the explanation, it was categorised as "*Scenario Information Only*".

A Chi-Square test was used to determine whether the type of comment depended on the round ($p < 0.05$). The most significant change between rounds was the increase in *Positive Comment and Understood Explanation* comments (28% in Round 1 responded in this way, and 47% in Round 2). This improvement indicates that the second explanation was better and less confusing and that participants were less critical of the explanation and understood it better.

Figure 8-2, below, shows the contribution each type of comment made to the Chi-Square test result. Positive Comment and Understood Explanation and Nonsense contributed the most to this result.

The modal Unguided Section score for the explanation (out of 10) was 8 for Round 1 and 10 for Round 2. However, the population means were equal across both rounds (two-sample t-test, $p > 0.05$). The absence of significant difference between Round 1 and Round 2 suggests that even though more participants rated the case study very highly in Round 2, it is difficult to tell whether improvements to the explanation affected its quality. The XQ Rubric category scores, discussed in the next section, are more explicit about the effect on the explanation Dr Cruz's revisions had.

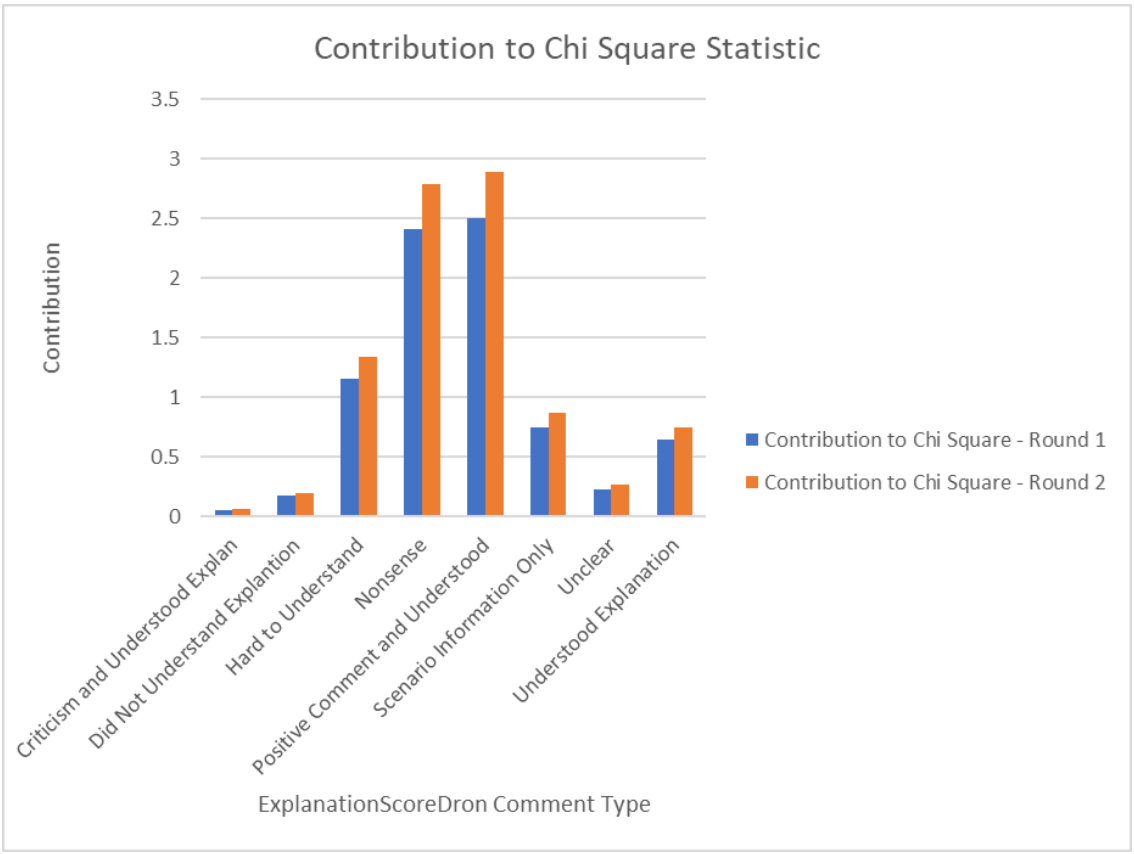

Figure 8-2 Comment Type Chi-Square Test

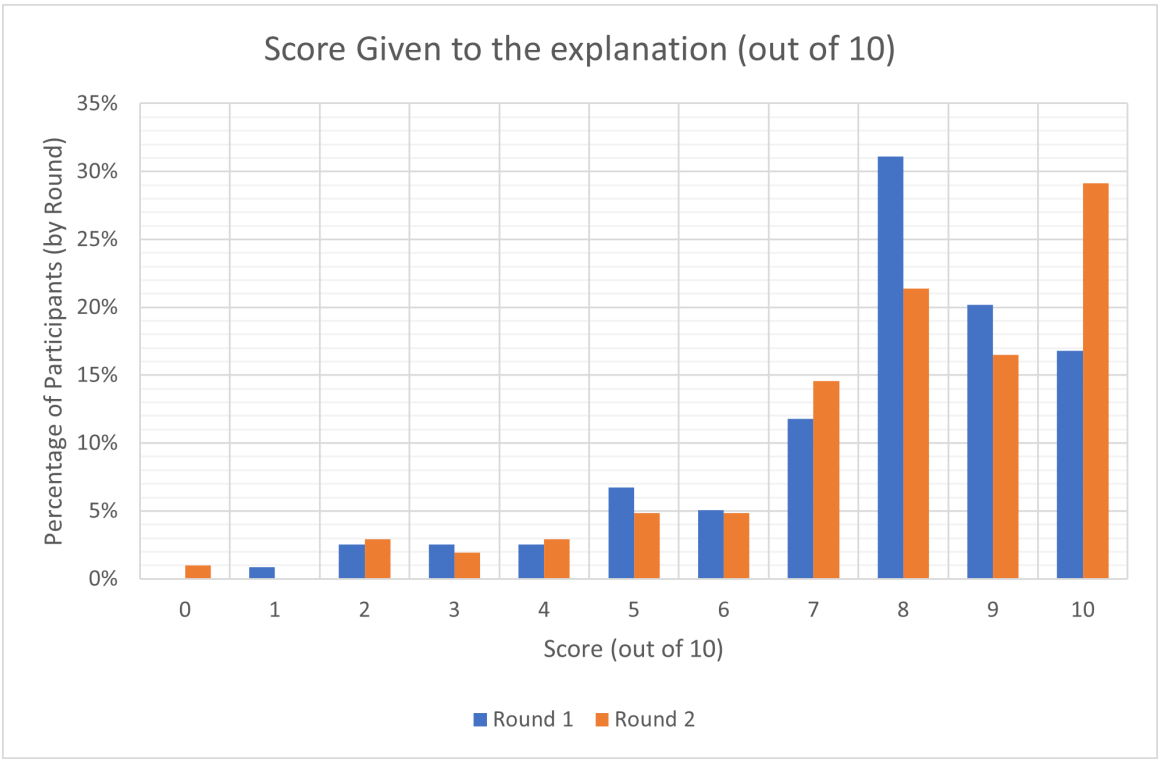

Figure 8-3 Score (out of 10) given to Explanations (Round 1 and Round 2)

8.3.3 Rubric

Question Code	Question	Most Common Answer in Batch 1	Percentage in Batch 1	Most Common Answer in Batch 1	Percentage in Batch 2
F01C00	Is the explanation clear?	The explanation is clear	50%	The explanation is clear	63%
F01C01	Is the explanation written with correct spelling and grammar?	The explanation has no grammatical errors and no spelling mistakes	72%	The explanation has no grammatical errors and no spelling mistakes	85%
F01C02	Does the explanation define the uncommon words and symbols it uses?	No definitions or keys were needed	55%	No definitions or keys were needed	69%
F01C03	Is the supporting information (or detail) well-presented and understandable?	The information supplied to support the explanation is clearly presented and understandable	49%	The information supplied to support the explanation is clearly presented and understandable	53%
F01C04	Can the explanation be acted upon?	The explanation can be acted upon	44%	The explanation can be acted upon	42%
F02C01	Are references provided?	No references are needed to support this explanation, and so no references are provided	51%	No references are needed to support this explanation, and so no references are provided	57%
F02C02	Are relevant facts mentioned in the explanation?	There are some references to relevant facts	48%	There are some references to relevant facts	40%
F03C01	Is the question (or implied question) answered?	The explanation answers the question (or implied question)	61%	The explanation answers the question (or implied question)	61%

Question Code	Question	Most Common Answer in Batch 1	Percentage in Batch 1	Most Common Answer in Batch 1	Percentage in Batch 2
F03C02	Can the explanation's reasoning be followed without difficulty?	The reasoning of the explanation is laid out clearly and it is possible to follow the argument	47%	The reasoning of the explanation is laid out clearly and it is possible to follow the argument	55%
F03C03	Is the explanation convincing?	The explanation is convincing	53%	The explanation is convincing	56%

In Round 2, more participants found the explanation clear and well-written (*F01C00*, *F01C01*). More participants did not need definitions and keys (*F01C02)*, and they also did not need further references (*F02C01)*. Participants were also more likely to find that the information was clearly presented *(F01C03)*, convincing *(F03C03),* and understandable (*F03C02)*.

The number of participants who felt that the explanation answered the question or implied question did not change *(F03C01)*.

These results, coupled with the increase in *Positive Comment and Understood Explanation* comments, strongly indicate that the XQ Rubric successfully guided Dr Cruz to improve his explanation.

8.3.4 Debrief of the Independent Expert

When the XQ Survey was completed, a debriefing survey was sent to Dr Cruz to discuss his experience of the experiment (Appendix E4c). Dr Cruz's comments were largely positive. However, he found the question: "*F01C04. Can this explanation be acted upon?*" unclear (Appendix A6c). He did not understand how best to edit his explanation using the results from question F01C04.

Apart from question F01C04, Dr Cruz found the XQ Rubric very useful when editing his explanation. He said that "*the XQ Rubric questions were quite useful to assess the explanations in*

terms of their quality" (Appendix A6c). Dr Cruz felt that "*using the XQ Survey to improve the explanations instead of trying by myself was a better approach since it allowed me to identify gaps that would be very difficult to find. For instance, the use of technical terms in the explanations*" (Appendix A6c).

8.4 Discussion

This experiment was designed to test the hypothesis that researchers who had not participated in the development of the XQ Rubric and the XQ Survey could nevertheless successfully use them to evaluate and improve their explanations. To do this, I created an explanation, assessed it using the XQ Rubric and the XQ Survey, revised the explanation, and then assessed how well the explanation had improved as a result of the XQ Rubric and survey responses.

The lack of a significant difference in the unguided score demonstrates that merely asking for a rating out of 10 is inadequate. However, the comments explaining why the participant gave the rating match the XQ Rubric's rating. The XQ Rubric enabled Dr Cruz to improve his explanation (Appendix A6c).

9 DISCUSSION

9.1 INTRODUCTION

This chapter is an overview of the methods and findings of this thesis. It discusses the research questions, aims, and objectives, the means employed in the construction of the XQ Rubric and Survey, the results, and the implications for further research.

9.2 THEORY AND METHODS

The methodology of this thesis is adapted from that of Peffers et al. (2007):

1. Define the specific research problem and justify the value of a solution
2. Define the objectives for a solution
3. Design and development
4. Observe and measure how well the artefact supports a solution to the problem
5. Demonstration
6. Evaluation

The method was a combination of iterative design-science and User-Centred Design (UCD) (Abras et al., 2004). These methods were chosen to suit the investigation and its primary construct, the rubric.

9.2.1 Findings of the Literature Review

The literature review defined the research problem as the absence of a robust evaluation scheme for XAI explanation and found justification for a solution. Gaps in the literature suggested objectives. The effect of poorly constructed XAI explanations – sometimes their complete omission – by creators and users of AI algorithms suggested the importance of a solution. Several case studies were described to illustrate this.

9.2.2 The Experiments

The experiments followed steps three to six of the Peffers et al. (2007) methodology. Each set of experiments (Chapters 5 to 8) used an iterative design-science methodology.

9.2.2.1 The Focus Group Using the Delphi Methodology Survey

The first set of experiments, the Focus Group, building on objectives defined by the literature review, concerned the design and development (Peffers et al., 2007) of an XQ rubric and XQ survey using feedback and advice from a focus group selected for the purposes.

The focus group participants were asked to consider how well the XQ Rubric and, later, the XQ Survey performed as evaluation schemes. The participants were given the task of using it to evaluate a series of created case studies to demonstrate and evaluate the XQ Rubric and XQ Survey. The focus group participants were debriefed after three rounds of improvement. The next set of experiments (this time on MTurk) was begun.

9.2.2.2 Validation of XQ Rubric and XQ Survey as Evaluation Tools

The second set of experiments – the first Amazon MTurk experiment – was also part of "design and development" (Peffers et al., 2007). This time the experiments were intended to help design and develop a methodology for using MTurk in the evaluation of XAI explanations which employed the XQ Rubric and the XQ Survey.

9.2.2.3 Validation of XQ Rubrics as a Tool for Constructive Feedback

The third set of experiments – the second MTurk experiments – were intended to "observe and measure how well the artefact supports a solution to the problem" (Peffers et al., 2007). This experiment was designed to evaluate the use of the artefacts (the explanation method) as a means of gathering feedback to improve an XAI explanation (the XQ Rubric and the XQ Survey). An explanation was revised using feedback from the XQ Rubric and the XQ Survey. This explanation was presented again to a different group of MTurk participants. It was hypothesised

that a good evaluation scheme would successfully guide the process of improvement to produce a better explanation.

9.2.2.4 Independent Validation of the XQ Rubric and the XQ Survey

This experiment was intended to demonstrate that the XQ Rubric and Survey could be used by someone unfamiliar with them (in this case, an independent AI researcher, Dr Cruz of Deakin University). Dr Cruz was invited to participate in an experiment to demonstrate and evaluate the final versions of the XQ Rubric and the Survey. Dr Cruz followed the methods employed in the two previous experiments to evaluate an explanation using the XQ Rubric and Survey on MTurk. He then used the MTurk evaluations to improve his explanation. Finally, he presented the revised explanation to different MTurk participants.

The results from this survey were analysed. It was confirmed that the XQ Rubric feedback had successfully improved the explanation. The fourth set of experiments – the third MTurk experiments – were intended to demonstrate and analyse the XQ Rubric and the Survey (Peffers et al., 2007).

9.3 The Developed Methodology

This section explains the methodology developed by this thesis. It has three parts: the XQ Rubric, the XQ Survey, and an outline of best practice for using the developed methodology.

9.3.1 The XQ Rubric

The XQ Rubric is the main research contribution of this thesis. It began as a paper rubric (figure 9-1), but soon became part of an online survey system (the XQ Survey). This presents the rubric in a survey format, where each row is one question, and the row's cells are the possible answers (figure 9-2).

Family	Category	Excellent (5)	Good (3)	Mediocre (1)	Bad (0)	Comments
Presentation Clarity	Suitable for the audience	Tailored for the audience	The explanation suits the audience	The explanation is written generally but uses words and terms the audience is familiar with	Written generally and uses words and terms the audience is unfamiliar with	
	Obeys the conventions of its medium (eg Grammar and Spelling)	Obeys the conventions of its medium	Mostly obeys the conventions of its medium – one or two small mistakes	Somewhat obeys the conventions of its medium – more than two small mistakes	Does not obey the conventions of its medium	
	Clear wording and unfamiliar terms/symbols defined	Glossary included if new words are introduced. The wording is clear	The wording is clear and term/symbols user are standard, uncommon words are defined	The wording is clear and uncommon words are defined	Terms introduced without introductions, unusual symbols used and not defined	
Content	Decision, Action, or Phenomena is explained	The thing is explained clearly	The thing is explained.	An unsuccessful attempt to explain the thing	The thing is not explained	
	How algorithm works is explained	How algorithm works is explained clearly	The algorithm is explained.	An unsuccessful attempt to explain the algorithm	How algorithm works is not explained	
Satisfaction	The target audience understands the explanation	The target audience understands and accepts the explanation	The target audience mostly understands and accepts the explanation	The target audience is unsure about the explanation	The target audience misunderstands the explanation	
	The explanation is clear and believable to the target audience	The explanation is clear and believable to the target audience	The explanation is mostly clear and believable to the target audience	The explanation is believed with some doubt to the target audience	The explanation is unclear and unbelievable to the target audience	
Truth	Verifiable references and plausible claims	The explanation has more than one verifiable reference and plausible claims	The explanation has plausible claims and references	The explanation has some verifiable references and semi-plausible claims	Unverifiable references and implausible claims	
	Claims reference evidence	The explanation has more than one piece of verifiable evidence	The explanation has plausible evidence	The explanation has some evidence	No evidence presented	
	The explanation is relevant to the discussion	The explanation answers the question (or implied question)	The explanation mostly answers the question (or implied question)	The explanation is only partly answers the question (or implied question)	The explanation is not relevant	
Total						

Figure 9-1 Paper Rubric Layout

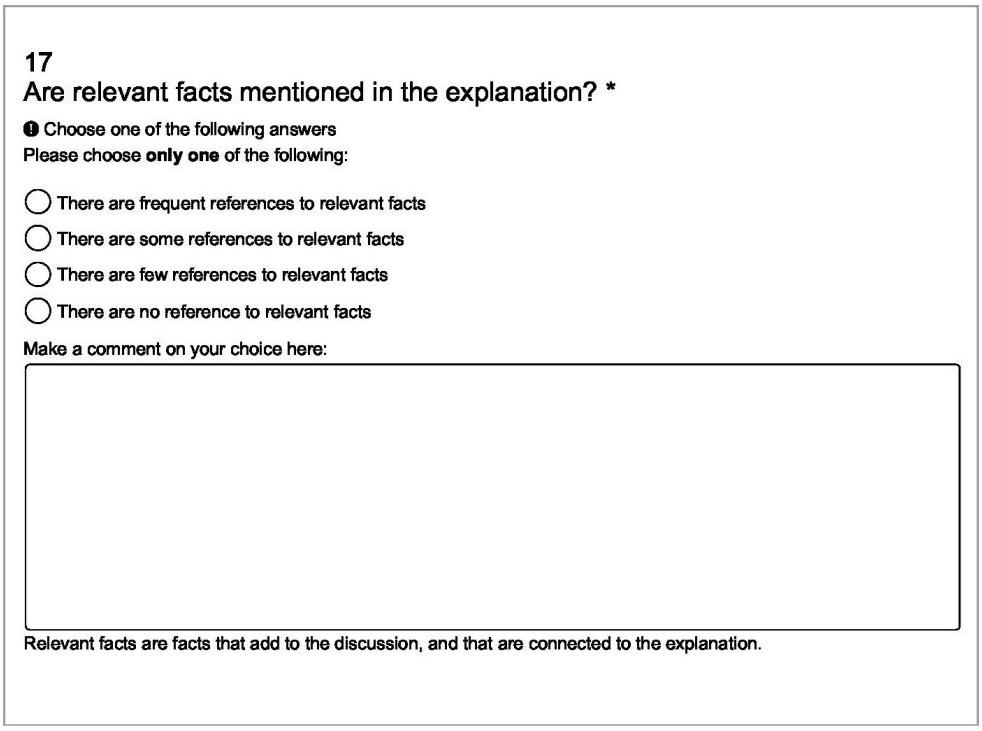

Figure 9-2 Rubric in Survey Format

9.3.2 The XQ Survey

The XQ Survey was a three-part online questionnaire run on LimeSurvey. It was first developed in the Focus Group Experiment and used in tandem with the XQ Rubric in every subsequent experiment. The XQ Survey was an easy-to-understand and accessible way to gather case study scores from the survey participants.

Unguided Score Section

This has two parts: a series of comprehension questions and a section for the participant to rate the case study out of 10 and justify their rating.

The use of comprehension questions follows the practice of Paolacci et al. (2010). The questions were designed to identify participants who had not given sufficient attention to the case study and those who selected answers to the survey at random ('bad-faith' participants). The questions were tailored to the case study. The correct answer was intended to be obvious but not easy to guess. For example, the comprehension question in the Independent Validation Experiment was, "*What colour shirt is the opponent team wearing?*"

The second part of the 'Unguided Score' section required participants to score the case study out of 10 to provide supplementary information about its quality. Participants were then asked to justify their scores. This provided another opportunity to receive unguided feedback about the case study.

It was important that there was just one correct answer to the comprehension questions. The answer had to be findable in the case study.

The XQ Rubric Section

The second section is the XQ Rubric section. This is covered above in Section 9.3.1.

The Demographic Section

The third section is the 'Demographic' section. This asked for information from participants about their age, occupation, and their country location. Participants were also questioned about their knowledge of key concepts in the explanation and case study. This section was designed to allow case studies with different populations of respondents to be compared.

Importantly, participants were given the opportunity to opt out of answering demographic questions such as their age, occupation, or location, keeping this information private. Questions about their knowledge of key concepts in the explanation or case study were presented as multiple-choice, giving participants the opportunity to indicate their level of familiarity with the material being examined.

9.3.3 Best Practice in Using MTurk to Evaluate Explanations

In using the MTurk service to evaluate XAI explanations with the XQ Rubric and XQ Survey it is recommended that only 'Master-level' workers are employed. Although more costly to employ, 'Master-level' MTurk workers are less likely to give bad faith answers, and more likely to be conscientious in their responses.

It is important to pay the MTurk workers at least the US Federal minimum wage. Paying less encourages participants to be careless with their answers. It is also advisable to treat MTurk workers as employees. This leads to more conscientious work, and workers are more likely to participate in another study of this kind. Workers should also be given a way to contact the survey organiser about problems.

10 CONCLUSION

10.1 INTRODUCTION

This thesis sought to answer these research questions:

> *What would a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations look like?*
>
> *How can a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations be created?*
>
> *Can such a rigorous, empirically justified, human-centred scheme for evaluating AI-decision explanations be used to improve explanations?*

By following the methodology outlined by Peffers et al. (2007) to create an artefact (in this case, an evaluation scheme), the thesis successfully constructed an empirically justified and rigorous human-centred scheme for evaluating AI-decision explanations.

The thesis exposed three deficiencies in the literature on explainable AI (XAI):

1. There is no method to gather non-expert evaluations of XAI explanations
2. There is no scheme for non-experts to use to evaluate XAI explanations
3. There is no clear definition of a good explanation for use in XAI research

In response:

1. A method to gather non-expert evaluations of XAI explanations was created (Section 9.3)
2. An evaluation scheme for non-experts to use in evaluating XAI explanations was created (XQ Rubric and XQ Survey)
3. A definition of a good explanation for use in XAI research was proposed (Section 2.5.4)

10.2 Contributions to the Thesis

10.2.1 Method to gather non-expert evaluations of XAI explanations

There is no generally accepted method for evaluating an XAI explanation, especially evaluation by a non-expert. XAI research mostly concerns developing XAI explanations, not their evaluation. This thesis took steps towards a solution, a method to gather and assess non-expert evaluation of XAI explanation.

The evaluation method developed is designed for the use of non-experts, not AI professionals. As AI becomes more widespread and more people are increasingly affected by AI technologies, non-experts will need good explanations of AI decisions.

10.2.2 Evaluation scheme for non-experts to use to evaluate XAI explanations

At present, there is no clear and well-developed method for use by non-experts in their evaluation of XAI explanations. The XQ Rubric and Survey offer a starting point.

The evaluation scheme developed in this thesis is stable and robust, with a well-demonstrated capability to evaluate very different case studies. Its successful use in different scenarios (the Focus Group experiment and the MTurk experiments) was demonstrated.

10.2.3 Model of "good explanation" for use in XAI research

Evaluation schemes developed without a proper understanding of what constitutes a good explanation are clearly inadequate (see Section 3.4.1). Only with reference to a good explanation can candidate explanations be compared. The '*know it when you see it*' view of explanation lacks objectivity. The definition employed by this thesis gives a reference point for a good explanation.

10.3 Findings from the Focus Group Experiments

Following a review of the literature, I selected case studies to be used with the XQ Rubric and gathered together a focus group of colleagues and others (Chapter 5). This group provided feedback about the draft rubric and helped to improve the wording of the rubric and its gradations.

10.4 Findings from the MTurk Experiments

When the focus group had completed its revisions of the XQ Rubric, it was validated using participants recruited through MTurk. The first MTurk experiment expanded the work of the focus group. This MTurk experiment had more participants than the focus group and a wider variety of ages, nationalities, and occupations. The MTurk participants were not known to me.

The second MTurk experiment extended the work of the first. It compared two explanations, one original and one revised. This experiment established that the findings of the XQ Rubric could be used to improve an explanation.

The third MTurk experiment, conducted in the same way as the second, developed the work of the second. It compared two explanations, one original and one revised. The experiment was also an independent validation of the XQ Rubric and Survey. Dr Cruz, an AI researcher who had not been involved in the creation and development of the XQ Rubric and Survey, was invited to collaborate in conducting the experiment. Participants who had not participated in earlier experiments were employed to evaluate and improve the explanation. This experiment demonstrated that person independent of its developer could use the XQ Rubric and Survey to evaluate and improve an explanation without the developer's assistance.

10.5 Limitations Of The Study

The thesis was limited in two ways: in its scope and in its analysis of the XQ Rubric and Survey. (It was also somewhat constrained by the closures and lockdowns of the COVID-19 pandemic.)

10.5.1 Thesis Scope

The thesis concentrated on assessment by a non-experts; it did not include expert perspectives on the explanation. This limitation (Section 1.4) was suggested by the Bhatt et al. (2020) interviews of XAI stakeholders. There have been few evaluation methods useful for non-experts: the thesis sought to consider and meet their interests and concerns.

The thesis is limited to the evaluation of explanations of XAI decisions.

10.5.2 Analysis of the XQ Rubric and XQ Survey

Analysis of the XQ Rubric and the XQ Survey was somewhat restricted by the number and composition of survey recipients. The focus group participants were all professional Australians who were known to me. They had a variety of backgrounds and ages. MTurk participants had similar backgrounds and ages to each other. All MTurk participants in every experiment were computer literate. The thesis would have benefited from more participants and a broader array of participants; limited resources prevented this.

10.5.3 COVID-19 Pandemic

Restrictions imposed during the COVID-19 pandemic interrupted and disrupted some practical, public parts of the thesis. I had hoped to do additional in-person focus groups, but COVID-19 restrictions prevented my meeting people in person.

10.6 Implications for Future Research

10.6.1 Tools Used to Evaluate the XAI Explanations

I used three survey tools for these experiments: LimeSurvey, MTurk, and Google Drive. LimeSurvey is an online survey platform. MTurk is a crowdsourcing work website. Google Drive, a file-sharing server, provided an easy way for survey participants to access the case studies. For this, LimeSurvey was unreliable.

10.6.2 The Audience

Participation Rates

I was disappointed by the small number of participants. At the focus group stage, the number of people who replied dwindled with each round of emails. In the MTurk experiments, it took longer and longer to find participants, and there were fewer overall. Ordinary users had evidently become increasingly reluctant to participate in the surveys. In addition, the pool of potential MTurk participants began to shrink because people who had joined earlier surveys were excluded from those conducted afterwards.

The number of participants diminished with each MTurk round in the third experiment, and I was forced to run the MTurk job advertisement again. Using MTurk's classification of 'new' was intended to catch the attention of people searching for jobs.

In Round 2 of the third experiment, I increased the remuneration of participants, first from US$2.00 to US$2.50, then from US$2.50 to US$3.00. The increased pay rate did not attract substantially more people. It could be that this job was unattractive to MTurk participants. Perhaps I had simply exhausted the supply of people willing to do it.

Bad-Faith Participants

I was disappointed in the high proportion of bad faith participants, that is, participants who failed the comprehension questions or gave nonsensical responses. Behaviour of this sort underscores the need for thorough screening by comprehension questions.

Non-Experts and the General Audience

The audience, for both the focus group and the MTurk participants, was deliberately chosen from people not specialists in XAI. However, the MTurk participants were from various backgrounds, and it is possible that a proportion had some acquaintance with XAI research. By necessity, people using MTurk were computer literate. Nevertheless, as far as possible, the testing was performed on a non-expert audience.

10.6.3 Case Studies

It was difficult to find material for use as example explanations for case studies. Selection of the case studies is covered in Section 4.6.

I was surprised by the MTurk participants' mild reaction to the Centrelink 'RoboDebt' letter, which attracted strong condemnation by the Ombudsman and the Australian public.

10.6.4 Time Taken

I was surprised at how long the XQ Surveys took to complete. I had expected that most people would finish a case study survey in 15 minutes. The survey seems to have taken considerably longer to complete.

10.6.5 Avenues for Future Research

There are three avenues for future research on the XQ Rubric, XQ Survey, and the method developed by this thesis. An opportunity for significant further research lies in extending the scope of the XQ Rubric and Survey to make it useful for experts. This would expand the scope of the XQ Rubric and Survey and extend the range of its practical use.

The XQ Rubric and Survey could be augmented and extended for use in all explanations, not merely XAI explanations. This would be useful when evaluating explanations in other fields, for example in medicine or education.

The XQ Rubric could be extended to deal with legal matters and social-justice concerns. Although there is a need for better XAI explanations concerning legal matters, the present XQ Rubric is not designed for this. The legal status of AI, AI explanations, and XAI is complex and beyond the scope of this thesis. Legal explanations, governed by different rules, are fundamentally different from explanations of an AI's decisions (Doshi-Velez et al., 2017).

This thesis also does not consider social-justice concerns about AI and XAI mentioned in Chapter 2. However, the XQ Rubric and Survey could be adapted to give more attention to social-justice matters. Preliminary work has been done by Gebru et al. (2021) to assess the quality of the data used in XAI explanations. New questions could be added to assess the social harm of incorrect explanations, for example.

10.7 Conclusion

The literature review identified three significant gaps in the explainable AI (XAI) literature.

1. There is no method to gather non-expert evaluations of XAI explanations
2. There is no evaluation scheme that a non-expert could use to evaluate an XAI explanation
3. There is no clear definition of a good explanation for use in XAI research

In response to these shortcomings, this thesis provided:

1. a model of a good explanation for use in XAI research (Section 2.5.3)
2. a methodology of gathering non-expert evaluations of XAI explanations
3. an evaluation scheme for non-experts to utilise in evaluating XAI explanations (XQ Rubric and XQ Survey).

This thesis set out a clear model of a good explanation and developed and explained how an XQ Rubric and an XQ Survey could be used for the evaluation of explanations by non-experts. The thesis also demonstrated that the XQ Survey and Rubric could be used to provide constructive feedback to improve explanations.

The rubric, and the associated survey and method, aim to lessen the use of inadequate, incomplete, and false explanations. Governments worldwide, notably the European Union, are beginning to pass laws requiring explanations for AI decisions. To satisfy the law, the explanations need to be good explanations.

This thesis created an XQ Rubric and an XQ Survey as tools for the evaluation of XAI explanations. The tools enable the creators of an XAI explanation to assess it. This thesis also outlined a model of good explanation, useful as a starting point for future research into the issues addressed by this thesis.

11 APPENDICES

Appendix A – Results

Appendix A1 – Focus Group Results from Part 1

Appendix A2 – Focus Group Results from Part 2

Appendix A3 – Focus Group Results from Part 3

Appendix A4 – Initial Validation Results

Appendix A5 – Constructive Feedback Results

Appendix A6 – Independent Validation Results

- Appendix A6a – Independent Validation Results - Part 1
- Appendix A6b – Independent Validation Results - Part 2
- Appendix A6c – Independent Validation Debriefing Survey

Appendix B – Ethics Approval

Appendix B1 – Focus Group Approval
Appendix B2 – Initial Validation Approval
Appendix B3 – Constructive Feedback Approval
Appendix B4 – Independent Validation Approval
Appendix B5 – Human Research Ethics Committee Applications

Appendix C – Case Studies

Appendix C1 – Focus Group Case Studies

- Appendix C1a – Case Studies from Part 1
- Appendix C1b – Case Study from Part 2
- Appendix C1c – Case Study from Part 3

Appendix C2 – Initial Validation Case Studies

- Appendix C2a – Initial Validation Case Studies Centrelink Letter
- Appendix C2b – Initial Validation Case Studies IBM Loan

Appendix C3 – Constructive Feedback Case Study

Appendix C4 – Independent Validation Case Studies

- Appendix C4a – Independent Validation Case Studies - Scenarios 1-6
- Appendix C4b – Updated Independent Validation Case Studies - Scenarios 1-6

Appendix D – Rubrics

Appendix D1 – Focus Group Part 1 Rubric
Appendix D2 – Focus Group Part 2 Rubric

Appendix E – Surveys

Appendix E1 – Focus Group Surveys
- Appendix E1a – Focus Group Round 2
- Appendix E1b – Focus Group Round 3

Appendix E2 – MTurk #1 Survey

Appendix E3 – MTurk #2 Survey

Appendix E4 – MTurk #3 Surveys
- Appendix E4a – MTurk #3 – Independent Validation Round 1
- Appendix E4b – MTurk #3 – Independent Validation Round 2
- Appendix E4c – MTurk #3 – Independent Validation Debriefing Survey

Appendix F – Scripts

Appendix F1 – MTurk #3 – Independent Validation Round 1 Randomisation Script
Appendix F2 – MTurk #3 – Independent Validation Round 2 Randomisation Script

Appendix G – Letters to Focus Group Participants

Appendix G1 – Delphi Invitation Emails
- Appendix G1a – Delphi Invitational Email
- Appendix G2a – Delphi Invitational Follow Up Email

Appendix G2 – Delphi Part 1 Emails
- Appendix G2a – Delphi Part 1 Initial Email
- Appendix G2b – Delphi Part 1 Follow Up Email
- Appendix G2c – Delphi Part 1 Second Follow Up Email

Appendix G3 – Delphi Part 2 Emails
- Appendix G3a – Delphi Part 2 Initial Email to People Who Did the First Survey
- Appendix G3b – Delphi Part 2 Initial Email to People Who Did NOT Do the First Survey
- Appendix G3c – Delphi Part 2 Follow Up Email

Appendix G4 – Delphi Part 3 Emails
- Appendix G4a – Delphi Part 3 Email

Appendix G5 – Final Emails

Appendix G5a – Final Email to Delphi Participants

12 REFERENCES

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda.* Paper presented at the SIGCHI Conference on Human Factors in Computing Systems. CHI'18.

Abdul, A., von der Weth, C., Kankanhalli, M., & Lim, B. Y. (2020). *COGAM: measuring and moderating cognitive load in machine learning model explanations.* Paper presented at the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.

Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications, 37*(4), 445-456.

Adom, D., Yeboah, A., & Ankrah, A. K. (2016). Constructivism philosophical paradigm: Implication for research, teaching and learning. *Global journal of arts humanities and social sciences, 4*(10), 1-9.

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. (2021). Algorithmic bias in data-driven innovation in the age of AI. In (Vol. 60, pp. 102387): Elsevier.

Amazon Web Services. (2020). Amazon Mechanical Turk Requester UI Guide. Retrieved from https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/Introduction.html

Amazon Web Services. (2022). Amazon Mechanical Turk Homepage. Retrieved from https://www.mturk.com/

Andrade, H. G. (1997). Understanding rubrics. *Educational Leadership, 54*(4), 14-17.

Angwin, Larson, Mattu, & Kirchner. (2016, 23 May 2016). Machine Bias. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arcuria, P., Morgan, W., & Fikes, T. G. (2019). *Validating the Use of LMS-Derived Rubric Structural Features to Facilitate Automated Measurement of Rubric Quality.* Paper presented at the Proceedings of the 9th International Conference on Learning Analytics & Knowledge.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115.

Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., . . . Mojsilovic, A. (2020). AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *J. Mach. Learn. Res., 21*(130), 1-6.

Australian Bureau of Statistics. (2009). *Australian and New Zealand Standard Classification of Occupations*. (1220). Online Retrieved from https://www.abs.gov.au/ausstats/abs@.nsf/Previousproducts/1220.0Main%20Features1First%20Edition,%20Revision%201?opendocument&tabname=Summary&prodno=1220.0&issue=First%20Edition,%20Revision%201&num=&view=#:~:text=1220.0%20%2D%20ANZSCO%20%2D%20Australian%20and%20New%20Zealand%20Standard%20Classification%20of%20Occupations%2C%20First%20Edition%2C%20Revision%201%C2%A0%C2%A0

Best, J. (2018). Numbers Games: Review of The Tyranny of Metrics by Jerry Z. Muller (2018). *Numeracy: Advancing Education in Quantitative Literacy, 11*(2).

Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Machine Learning Explainability for External Stakeholders. *arXiv preprint*.

Bignold, A., Cruz, F., Dazeley, R., Vamplew, P., & Foale, C. (2022). Human engagement providing evaluative and informative advice for interactive reinforcement learning. *Neural Computing and Applications*, 1-16.

Biran, O., & Cotton, C. (2017). *Explanation and justification in machine learning: A survey.* Paper presented at the IJCAI-17 Workshop on Explainable AI (XAI).

Boston, C. (2002). *Understanding Scoring Rubrics: A Guide for Teachers*. University of Maryland: ERIC Clearinghouse on Assessment and Evaluation.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior, 36*(1), 21-40. doi:10.1177/0093854808326545

Brown, G. (2006). Explaining. In O. Hargie (Ed.), *The handbook of communication skills* (3 ed., pp. 195 - 228): Routledge.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). *Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems.* Paper presented at the Proceedings of the 25th international conference on intelligent user interfaces.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research*: American Psychological Association.

Burns, A. C., & Bush, R. F. (2007). *Basic marketing research using Microsoft Excel data analysis*: Prentice Hall Press.

Canadian Marketing Association. (2022). *Privacy Law Pitfalls: Lessons Learned from the European Union*. Retrieved from Online: https://thecma.ca/docs/default-source/default-document-library/cma-2022-report-privacy-legislation-pitfalls.pdf

Carenini, G., & Moore, J. D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence, 170*(11), 925-952. doi:10.1016/j.artint.2006.05.003

Carney, T. (2019). Robo-debt illegality: The seven veils of failed guarantees of the rule of law? *Alternative Law Journal, 44*(1), 4-10.

Carstensen, A.-K., & Bernhard, J. (2019). Design science research – a powerful tool for improving methods in engineering education research. *European Journal of Engineering Education, 44*(1-2), 85-102.

Cialdini, R. B. (2007). Authority: Directed Deference. In *Influence: The psychology of persuasion* (Revised ed., pp. 208-236). New York: Collins Business.

Commonwealth Ombudsman. (2017). Centrelink's automated debt raising and recovery system (02|2017).

Cooley, M. (2000). Human-centered design. *Information design*, 59-81.

Cruz, F., Dazeley, R., Vamplew, P., & Moreira, I. (2021). Explainable robotic systems: Understanding goal-driven actions in a reinforcement learning scenario. *Neural Computing and Applications*, 1-18.

Dalkey. (1969). *The Delphi method: An experimental study of group opinion*. Retrieved from Santa Monica Calif:

Dalkey, & Helmer. (1963). An experimental application of the Delphi method to the use of experts. *Management science, 9*(3), 458-467.

Dastin, J. (2018, 11 October). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters Online*.

David, M. (2020). The Correspondence Theory of Truth. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Online: Metaphysics Research Lab, Stanford University.

Di Eugenio, B., Glass, M., & Trolio, M. (2002). *The DIAG experiments: Natural language generation for intelligent tutoring systems.* Paper presented at the Proceedings of the International Natural Language Generation Conference.

Di Gangi, P. M., McAllister, C. P., Howard, J. L., Thatcher, J. B., & Ferris, G. R. (2022). Can you see opportunity knocking? An examination of technology-based political skill on opportunity recognition in online communities for MTurk workers. *Journal of Internet Research*.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv e-prints*. Retrieved from https://arxiv.org/abs/1702.08608

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., . . . Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *arXiv preprint arXiv:1711.01134*. Retrieved from https://arxiv.org/abs/1711.01134

du Boulay, B., O'Shea, T., & Monk, J. (1981). The black box inside the glass box: presenting computing concepts to novices. *International Journal of Man-Machine Studies, 14*(3), 237-249.

Elizalde, F., Sucar, E., Noguez, J., & Reyes, A. (2009). *Generating explanations based on Markov decision processes*. Paper presented at the MICAI 2009: Advances in Artificial Intelligence: 8th Mexican International Conference on Artificial Intelligence, Guanajuato, México, November 9-13, 2009 Proceedings, Mexico. https://books.google.com.au/books?id=3SdqCQAAQBAJ

European Union. (2017). Robots: Legal Affairs Committee calls for EU wide rules [Press release]. Retrieved from http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+IM-PRESS+20170110IPR57613+0+DOC+PDF+V0//EN&language=EN

Farrell, P., & McDonald, A. (2019, 27 June 2019). Centrelink robo-debt system is 'extortion', says former tribunal member. *ABC News*. Retrieved from https://www.abc.net.au/news/2019-06-27/centrelink-robo-debt-system-extortion-former-tribunal-member/11252306

FICO Community. (2019). Explainable Machine Learning Challenge. Retrieved from https://community.fico.com/s/explainable-machine-learning-challenge

Fowler Jr, F. J. (2013). *Survey research methods*: SAGE publications.

Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., & Taly, A. (2019). *Explainable AI in industry.* Paper presented at the Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.

Galanti, R., Coma-Puig, B., de Leoni, M., Carmona, J., & Navarin, N. (2020). *Explainable predictive process monitoring.* Paper presented at the 2020 2nd International Conference on Process Mining (ICPM).

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the Acm, 64*(12), 86-92.

Goodman. (1987). The Delphi technique: a critique. *Journal of advanced nursing, 12*(6), 729-734.

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine, 38*(3), 50. doi:10.1609/aimag.v38i3.2741

Google. (2022). Google Drive. Online. Retrieved from drive.google.com

Grysman, A. (2015). Collecting narrative data on Amazon's Mechanical Turk. *Applied Cognitive Psychology, 29*(4), 573-583.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37), eaay7120.

Hendrickson, D., Navarro, D., Langsford, S., Kennedy, L., & Perfors, A. (2015). *Session 2: Ethical issues surrounding online experiments - Some reflections on trying to be ethical on Mechanical Turk.* Paper presented at the Australasian Experimental Psychology Conference (EPC), Online.

Hevner, A., & Chatterjee, S. (2010). Design science research in information systems. In *Design research in information systems* (pp. 9-22): Springer.

Hind, M., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., & Varshney, K. R. (2018). Increasing Trust in AI Services through Supplier's Declarations of Conformity. *arXiv preprint arXiv:1808.07261*.

Holzinger, A. (2017). Transparency & Trust in Machine Learning: Making AI interpretable and explainable. *HCI-KDD*. Retrieved from https://hci-kdd.org/2017/10/09/transparency-trust-machine-learning-making-ai-interpretable-explainable/

Hutson, M. (2021). The opacity of artificial intelligence makes it hard to tell when decision-making is biased. *IEEE Spectrum, 58*(2), 40-45.

IBM Research. (2019). AI Explainability 360 - Demo. Retrieved from https://aix360.mybluemix.net/explanation_cust#

Ishii, H., Kobayashi, M., & Arita, K. (1994). Iterative design of seamless collaboration media. *Communications of the Acm, 37*(8), 83-97.

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*: Cambridge University Press.

Jawski, G. (2019). FICO Announces Winners of Inaugural xML Challenge [Press release]. Retrieved from https://www.fico.com/en/newsroom/fico-announces-winners-inaugural-xml-challenge

Johannesson, P., & Perjons, E. (2014). Research strategies and methods. In *An Introduction to Design Science* (pp. 39-73): Springer.

Keneni. (2018). *Evolving Rule Based Explainable Artificial Intelligence for Decision Support System of Unmanned Aerial Vehicles.* (Masters of Engineering Degree). University of Toledo, United States of America.

Keneni, Kaur, D., Al Bataineh, A., Devabhaktuni, V. K., Javaid, A. Y., Zaientz, J. D., & Marinier, R. P. (2019). Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. *IEEE Access, 7*, 17001-17016.

Kim, M.-Y., Atakishiyev, S., Babiker, H. K. B., Farruque, N., Goebel, R., Zaïane, O. R., . . . Yao, H. (2021). A Multi-Component Framework for the Analysis and Design of Explainable Artificial Intelligence. *Machine Learning and Knowledge Extraction, 3*(4), 900-921.

Kittur, A., Chi, E. H., & Suh, B. (2008). *Crowdsourcing user studies with Mechanical Turk.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Kotrlik, J., & Higgins, C. (2001). Organizational research: Determining appropriate sample size in survey research appropriate sample size in survey research. *Information technology, learning, and performance journal, 19*(1), 43.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology 50*(1), 537-567.

Kuwajima, H., & Ishikawa, F. (2019). *Adapting SQuaRE for Quality Assessment of Artificial Intelligence Systems.* Paper presented at the 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW).

Laidlaw, J. (2014). Expert Panel. Retrieved from https://www.betterevaluation.org/en/evaluation-options/expert_panel

Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., & Séroussi, B. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial intelligence in medicine, 94*, 42-53.

Lane, H. C., Core, M., Lent, M. V., Solomon, S., & Gomboc, D. (2006). *Explainable Artificial Intelligence for Training and Tutoring*. Retrieved from Defense Technical Information Center: http://www.dtic.mil/dtic/tr/fulltext/u2/a459148.pdf

Langfeldt, L. (2004). Expert panels evaluating research: decision-making and sources of bias. *Research Evaluation, 13*(1), 51-62.

Layton, R. (2022, 22 February). New Model Code For Personal Data Protection Is Better Than GDPR. *Forbes*.

Lester, J. C., & Porter, B. W. (1997). Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics, 23*(1), 65-101.

LimeSurvey. (2021). LimeSurvey Survey Tool. Online. Retrieved from https://www.limesurvey.org/

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

Lizotte, D. J., Harris, C. J., McNeill, K. L., Marx, R. W., & Krajcik, J. (2003). *Usable assessments aligned with curriculum materials: Measuring explanation as a scientific way of knowing.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020). *Explainable artificial intelligence: Concepts, applications, research challenges and visions.* Paper presented at the International Cross-Domain Conference for Machine Learning and Knowledge Extraction.

Malle, B. F. (1999). How People Explain Behavior: A New Theoretical Framework. *Personality and Social Psychology Review, 3*(1), 23-48. Retrieved from https://pdfs.semanticscholar.org/645a/f3039537e1ab2419a621b62cc17006a15729.pdf

McNeill, K. L., & Krajcik, J. (2006). *Supporting students' construction of scientific explanation through generic versus context-specific written scaffolds.* Paper presented at the annual meeting of the American educational research association, San Francisco.

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. *CoRR*. Retrieved from http://arxiv.org/abs/1706.07269

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum (Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences). *ArXiv e-prints*. Retrieved from http://adsabs.harvard.edu/abs/2017arXiv171200547M

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3-4), 1-45.

Moradi, M., & Samwald, M. (2021). Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications, 165*, 113941.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical assessment, research & evaluation, 7*(10), 23-31.

Muller, J. Z. (2019). *The tyranny of metrics*: Princeton University Press.

Nielsen, J. (1993). Iterative user-interface design. *Computer, 26*(11), 32-41.

Nieveen, N., & Folmer, E. (2013). Formative evaluation in educational design research. *Design Research, 153*, 152-169.

Norton, M. I., Frost, J. H., & Ariely, D. (2007). Less is more: the lure of ambiguity, or why familiarity breeds contempt. *Journal of personality and social psychology, 92*(1), 97.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., . . . Krasanakis, E. (2020). Bias in data-driven artificial intelligence systems - An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1356.

Paolacci, Chandler, & Ipeirotis. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411-419.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184-188.

Patton, M. Q. (2008). *Utilization-focused evaluation*: Sage publications.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for Information Systems Research. *Journal of Management Information Systems, 24*(3), 45-77. doi:10.2753/Mis0742-1222240302

Polonski, V. (2018). People don't trust A.I.? Here's how we can change that. In *The Conversation*.

Rehse, J.-R., Mehdiyev, N., & Fettke, P. (2019). Towards explainable process predictions for industry 4.0 in the dfki-smart-lego-factory. *KI-Künstliche Intelligenz, 33*(2), 181-187.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*: Cambridge university press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why should I trust you?: Explaining the predictions of any classifier.* Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). *Overtrust of Robots in Emergency Evacuation Scenarios*. Paper presented at the The Eleventh ACM/IEEE International Conference on Human Robot Interaction, Christchurch, New Zealand.

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems, 33*(6), 673-705.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*: Malaysia; Pearson Education Limited.

Sandelowski, M. (2000). Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed‐method studies. *Research in nursing & health, 23*(3), 246-255.

Sanneman, L., & Shah, J. A. (2020). *A Situation Awareness-Based Framework for Design and Evaluation of Explainable AI.* Paper presented at the International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems.

Santa Cabrera, J. R., Castillo, P., & Jimenez, J. (2017). Implementing rubrics to assess writing skills in an Adults Advanced EFL (English as a Foreign Language) Class at ICDA (Instituto Cultural Dominico Americano). *International Journal of Innovation and Applied Studies, 20*(2), 681-710.

Sevian, H., & Gonsalves, L. (2008). Analysing how Scientists Explain their Research: A rubric for measuring the effectiveness of scientific explanations. *International Journal of Science Education, 30*(11), 1441-1467. doi:10.1080/09500690802267579

Simpson, T. W. (2012). What Is Trust? *Pacific Philosophical Quarterly, 93*(4), 550-569. doi:10.1111/j.1468-0114.2012.01438.x

Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The Delphi Method for Graduate Research. *Journal of Information Technology Education: Research, 6*(1), 1-21.

Sternberg, R. J., & Frensch, P. A. (1992). On being an expert: A cost-benefit analysis. In *The psychology of expertise* (pp. 191-203): Springer.

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *SSRN Electronic Journal*. Retrieved from http://dx.doi.org/10.2139/ssrn.3263878

Tangermann, V. (2018). To Build Trust In Artificial Intelligence, IBM Wants Developers To Prove Their Algorithms Are Fair. *futurism.com*. Retrieved from https://futurism.com/trust-artificial-intelligence-ibm/

Thellman, S., Silvervarg, A., & Ziemke, T. (2017). *Lay Causal Explanations of Human vs. Humanoid Behavior.* Paper presented at the International Conference on Intelligent Virtual Agents.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward Medical XAI. *IEEE transactions on neural networks and learning systems, 32*(11), 4793-4813.

Turley, E. D., & Gallagher, C. W. (2008). On the" uses" of rubrics: reframing the great rubric debate. *English Journal*, 87-92.

United States Department Of Labor. (2022). Minimum Wage. Retrieved from https://www.dol.gov/general/topic/wages/minimumwage#:~:text=The%20federal%20minimum%20wage%20for,of%20the%20two%20minimum%20wages.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Journal of Artificial Intelligence, 291*.

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Vilone, G., & Longo, L. (2021a). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction, 3*(3), 615-661.

Vilone, G., & Longo, L. (2021b). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion, 76*, 89-106.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). *Designing theory-driven user-centric explainable AI.* Paper presented at the Proceedings of the 2019 CHI conference on human factors in computing systems.

Wang, D., Zhang, W., & Lim, B. Y. (2021). Show or suppress? Managing input uncertainty in machine learning model explanations. *Artificial Intelligence, 294*, 103456.

Whitby, B. (2009). *Artificial Intelligence*: Rosen Publishing Group.

Winfield, A. F., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., . . . Theodorou, A. (2021). IEEE P7001: a proposed standard on transparency. *Frontiers in Robotics and AI*, 225.

Wu, P. (2017). GDPR and its impacts on machine learning applications. Retrieved from https://medium.com/trustableai/gdpr-and-its-impacts-on-machine-learning-applications-d5b5b0c3a815

Your Dictionary. (2022). Supporting Detail. In *Your Dictionary*. Online.

Zhang, & Han. (2022). *Algorithms Have Built Racial Bias in Legal System-Accept or Not?* Paper presented at the 2021 International Conference on Social Development and Media Communication (SDMC 2021).

Zhang, & Lim. (2022). *Towards Relatable Explainable AI with the Perceptual Process.* Paper presented at the CHI Conference on Human Factors in Computing Systems.

Zhang, Wu, & Zhu. (2018). *Interpretable convolutional neural networks.* Paper presented at the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zoltowski, C. B., Oakes, W. C., & Cardella, M. E. (2012). Students' ways of experiencing human-centered design. *Journal of Engineering Education, 101*(1), 28-59.