

# A Cognitive Work Analysis Approach to Explainable Artificial Intelligence in Non-Expert Financial Decision-Making

by

Murat Dikmen

A thesis  
presented to the University of Waterloo  
in fulfillment of the  
thesis requirement for the degree of  
Doctor of Philosophy  
in  
Systems Design Engineering

Waterloo, Ontario, Canada, 2022  
© Murat Dikmen 2022

## Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Ann Bisantz  
Professor, Industrial and Systems Engineering  
University of Buffalo

Supervisor: Catherine Burns  
Professor, Dept. of Systems Design Engineering  
University of Waterloo

Internal Member: Kerstin Dautenhahn  
Professor, Dept. of Electrical & Computer Engineering  
University of Waterloo

Internal Member: Mark Hancock  
Professor, Dept. of Management Sciences  
University of Waterloo

Internal-External Member: Edith Law  
Professor, Dept. of Computer Science  
University of Waterloo

## **Author's Declaration**

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Statement of Contribution

Chapter 2 of this thesis consists of a paper that was co-authored by myself and my supervisor, Dr. Burns:

Dikmen, M., & Burns, C. (2020, December). Abstraction Hierarchy Based Explainable Artificial Intelligence. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 64, No. 1, pp. 319-323). Sage CA: Los Angeles, CA: SAGE Publications.

Chapter 3 of this thesis consists of a paper that was co-authored by myself and my supervisor, Dr. Burns:

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162, 102792.

## Abstract

Artificial Intelligence (AI) is being increasingly used to assist complex decision-making such as financial investing. As most AI systems rely on black-box machine learning models, understanding how to support human decision-makers and gaining users' trust becomes important. Explainable Artificial Intelligence (XAI) has been proposed to address these issues by making the decision-making process of AI systems understandable to users. However, existing XAI approaches fail to take into account users' domain experience, and fail to support users with limited domain expertise. This work aims to fill this gap. We presented an approach to integrate domain expertise into XAI, and showed that this approach can have a number of benefits to users of XAI systems such as improved task performance and better assessment of XAI. The main contributions of this work include identifying the benefits of adding domain knowledge to XAI, demonstrating the usefulness of Cognitive Work Analysis (CWA) in XAI, and developing recommendations for future design of AI systems.

First, through a Work Domain Analysis (WDA) approach, we identified opportunities to improve the existing XAI approaches by augmenting the explanations with domain knowledge and conducted an online study with 100 participants on users' perceptions of AI in a credit approval context. Results showed some benefits in improving user perceptions and highlighted the importance of contextual factors.

Next, we introduced a testbed for exploring user behavior and task performance in a financial decision-making task. We designed decision-support aids based on domain knowledge and explored their effectiveness in an experimental study with 60 participants. In the study, participants engaged with an AI assistant and made investing decisions. Depending on the condition, participants had access to domain knowledge presented on a separate display, domain knowledge embedded in the AI assistant, or no access to domain knowledge. The results showed that participants who had access to domain knowledge relied less on AI when it was incorrect, and obtained better task performance. The effect of domain knowledge on perceptions of AI was limited.

Next, we analyzed the user interviews that were part of the previous study. We identified users' mental models of AI and multiple ways they integrated the AI into their decision-making process. The analysis also revealed the complexity of designing for non-expert users, and we developed recommendations for future research and design.

Finally, we conducted a Control Task Analysis and Strategies Analysis to synthesize the qualitative and quantitative findings and developed decision ladders and information flow maps. The analyses provided insights into the influence of AI on the decision-making

process, challenges associated with non-expert users, and opportunities to improve AI user interface design.

## Acknowledgements

First of all, I would like to thank my supervisor, Catherine Burns, for all the support she provided throughout these years. Since the start of my master's program, she always made sure that I am making meaningful process while continuously learning and growing. She was always there when I needed help. Without her support, this thesis would not be possible.

I would also like to thank my committee members, Dr. Ann Bisantz, Dr. Kerstin Dautenhahn, Dr. Edith Law, and Dr. Mark Hancock, for providing invaluable feedback and recommendations to improve this dissertation.

I would like to acknowledge Natural Sciences and Engineering Research Council of Canada (NSERC) for providing financial support during my PhD. I also want to thank Bryant Whyte for providing subject-matter expertise, Madeleine Kiera Nolan for moderating research sessions and helping me with early investigations into AI, and Sana Allana for helping with study design.

I would like to thank all current and former Advanced Interface Design Lab members who made this journey much more enjoyable and inspiring. Special thanks to Dev Minotra for helping me accelerate my research career and for providing mentorship, and to Yeti Li for exploring the mysterious world of research together.

Finally, I would like to thank my family for always being there. This work would not be possible without their unconditional support throughout these years.

## **Dedication**

To Neriman,  
My amazing wife



# Table of Contents

List of Figures	xiii
List of Tables	xv
List of Abbreviations	xvi
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation and Research Questions . . . . .	2
1.2 Background . . . . .	4
1.2.1 From Human-Automation Interaction to Human-AI Interaction . . . . .	4
1.2.2 Overview of Explainable AI . . . . .	5
1.2.3 Trust in AI . . . . .	6
1.2.4 The Need for a Domain-Centric Approach . . . . .	7
1.2.5 Cognitive Work Analysis as A Domain-Centric Tool . . . . .	8
1.2.6 Conceptualizing Domain Knowledge . . . . .	11
1.3 Contributions . . . . .	11
1.4 Thesis Overview . . . . .	12
<b>2 Study 1: Exploring Abstraction Hierarchy-Based Explanations</b>	<b>14</b>
2.1 Introduction . . . . .	14
2.2 Abstraction Hierarchy Based Explainable Artificial Intelligence . . . . .	15

2.2.1	Background . . . . .	15
2.2.2	Overview of the System . . . . .	17
2.2.3	Method . . . . .	21
2.2.4	Results . . . . .	24
2.2.5	Discussion . . . . .	26
2.3	Contributions and Conclusion . . . . .	27
<b>3</b>	<b>Study 2: Investigating the Role of Domain Knowledge in Explainable AI</b>	<b>29</b>
3.1	Introduction . . . . .	29
3.2	The Effects of Domain Knowledge on Trust in Explainable AI and Task Performance: A Case of Peer-to-Peer Lending . . . . .	30
3.2.1	Background and Related Work . . . . .	31
3.2.2	Research Questions and Hypotheses . . . . .	35
3.2.3	Method . . . . .	36
3.2.4	Results . . . . .	44
3.2.5	Discussion . . . . .	47
3.3	Contributions and Conclusion . . . . .	52
<b>4</b>	<b>Embedding Domain Knowledge in Explainable AI</b>	<b>54</b>
4.1	Introduction . . . . .	54
4.2	Background . . . . .	54
4.3	Research Questions and Hypotheses . . . . .	56
4.4	Method . . . . .	57
4.4.1	Apparatus . . . . .	57
4.4.2	Sample . . . . .	59
4.5	Results . . . . .	59
4.5.1	Task Performance . . . . .	59
4.5.2	Subjective Ratings . . . . .	60

4.6	Discussion . . . . .	64
4.6.1	The Effect of Embedding Domain Knowledge . . . . .	64
4.6.2	Subjective Measures . . . . .	65
4.6.3	Implications . . . . .	65
4.6.4	Limitations . . . . .	66
4.6.5	Future Directions . . . . .	67
4.7	Contributions and Conclusion . . . . .	67
<b>5</b>	<b>Using Cognitive Work Analysis to Understand Decision-Making with AI</b>	<b>68</b>
5.1	Introduction . . . . .	68
5.2	Method . . . . .	68
5.3	Modeling Workflow with AI: Control Task Analysis . . . . .	69
5.4	Modeling Strategies using Strategies Analysis . . . . .	73
5.5	Discussion . . . . .	81
5.5.1	Control Task Analysis . . . . .	81
5.5.2	Modeling AI on Decision Ladders . . . . .	83
5.5.3	Strategies Analysis . . . . .	84
5.5.4	Future Work . . . . .	88
5.6	Contributions and Conclusion . . . . .	88
<b>6</b>	<b>General Discussion and Conclusion</b>	<b>89</b>
6.1	Introduction . . . . .	89
6.2	Summary of Key Findings . . . . .	89
6.3	Research Questions . . . . .	90
6.3.1	What is the role and importance of domain knowledge in human-AI interaction? . . . . .	90
6.3.2	How can we leverage Cognitive Work Analysis in understanding and designing AI systems? . . . . .	91
6.4	Discussion and Implications . . . . .	92

6.4.1	Domain Knowledge and AI . . . . .	93
6.4.2	Cognitive Work Analysis and AI . . . . .	94
6.4.3	Generalizability . . . . .	95
6.5	Contributions . . . . .	95
6.6	Limitations . . . . .	97
6.7	Future Work . . . . .	97
6.8	Conclusion . . . . .	98
<b>References</b>		<b>99</b>
<b>APPENDICES</b>		<b>117</b>
<b>A Study Materials and Additional Data Analysis</b>		<b>118</b>
A.1	Study 1 . . . . .	118
A.1.1	Study Materials . . . . .	118
A.1.2	Ethics Approval . . . . .	125
A.1.3	Recruitment . . . . .	125
A.1.4	Information Letter and Consent Form . . . . .	127
A.1.5	Procedure . . . . .	130
A.1.6	Data Analysis . . . . .	134
A.2	Study 2 . . . . .	135
A.2.1	Study Materials . . . . .	135
A.2.2	Ethics Approval . . . . .	140
A.2.3	Recruitment . . . . .	140
A.2.4	Information Letter and Consent Form . . . . .	140
A.2.5	Procedure . . . . .	145
A.2.6	Data Analysis . . . . .	152
A.3	Impact of COVID-19 . . . . .	154
A.4	Reflections . . . . .	154

# List of Figures

2.1	One of the fifteen cases used in the study. Information about the loan applicant’s financial situation, loan details, AI system’s decision and explanations are presented on a single page. . . . .	18
2.2	Abstraction hierarchy of loan evaluation system. . . . .	19
2.3	The flow used to generate explanations. In this example, liquid assets (e.g. account balances) achieve the generalized function CAPITAL (1), which can also be achieved using hard assets (2). CAPITAL influences decision (3) such that higher capital leads to better chances of approval. . . . .	20
2.4	AH-based explanations. For each factor mentioned in the data-driven explanations, a description based on AH was presented (dotted box at the bottom). . . . .	21
3.1	List of available loans. . . . .	38
3.2	Financial profile of a borrower. . . . .	39
3.3	Machine learning assistant (on the left) and domain knowledge-based decision aid (on the right). . . . .	41
3.4	The number of risky investments (on the left) and the number of investments among loans where the AI was incorrect (on the right). . . . .	45
4.1	The machine learning assistant in the embedded condition. The red alert box was shown when the risk predicted by the AI was lower than 50% yet the borrower had red flags according to expert domain knowledge . . . . .	58
4.2	Differences in overall portfolio. . . . .	60
4.3	Differences in the number of investments made in loans where the AI made incorrect predictions. . . . .	61

4.4	Differences in the amount of money invested in risky loans (left), and in loans where the AI made incorrect predictions (right).	61
4.5	Trust in AI (trust dimension).	63
4.6	Perceived information amount on the user interface.	63
5.1	Decision ladder depicting how investment decisions can be made on the P2P platform	70
5.2	Decision ladder depicting how investment decisions with the AI assistant can be made on the P2P platform. In this case, two common shortcuts taken by the participants were shown. The shortcut <b>A</b> refers to premature conclusions (mostly low-risk assessments) based on the signs provided by the AI. The shortcut <b>B</b> refers to making a decision when the AI assistant outputted a high default risk percentage	72
5.3	The AI-first strategy employed by participants.	74
5.4	Manual-first strategy employed by participants.	76
5.5	The AI-guided strategy employed by participants.	78
5.6	Information flow map showing the effect of embedding domain knowledge in AI interface on changes in the AI-first strategy.	80
A.1	Research Ethics Committee approval for Study I.	126
A.2	Research Ethics Committee approval for Study II and III.	141
A.3	Home page.	155
A.4	Home page (cont'd).	156
A.5	Home page (cont'd).	157
A.6	Portfolio page.	158

# List of Tables

2.1	Measures used in the experiment. . . . .	23
4.1	Descriptive statistics. . . . .	62
A.1	AH-based explanations . . . . .	123
A.2	Tooltip Descriptions . . . . .	135
A.3	AI Assistant and Key Indicators Panel Descriptions . . . . .	138

# List of Abbreviations

**CWA** Cognitive Work Analysis

**WDA** Work Domain Analysis

**AH** Abstraction Hierarchy

**ConTA** Control Task Analysis

**SA** Strategies Analysis

**P2P** Peer-to-Peer Lending



# Chapter 1

## Introduction

Recent developments in the field of Artificial Intelligence (AI) such as neural networks and deep learning [101] opened up new ways to support people and make the work processes more efficient. From finance to healthcare to defence, almost all fields and industries are currently seeking to take advantage of AI to improve health, safety, and effectiveness. At the core of modern AI lies the recent advancements in machine learning, which are driven by the advancements in computing power and the availability of data to produce AI technologies that learn how to carry out complex tasks that were not possible before.

As AI invades work and everyday life in the form of products and services, understanding how users interact with these systems and addressing challenges related to usability, trust, and adoption becomes important [155, 81, 177]. Similar to many technologies that came before that, understanding the human-AI interaction and identifying issues, and providing solutions is key for the adoption of AI in a healthy way that achieves the goal of augmenting human experience and human cognition [3, 63]. In this dissertation, we present a human factors perspective to human-AI interaction and study how users interact with AI to solve problems.

The field of human-AI interaction is growing rapidly. As the AI systems become more accessible due to developments in a number of technologies (e.g. cloud-based applications) surrounding the AI ecosystem, AI will be part of daily life, and increasingly replace existing technologies. We are already seeing this happening in certain spaces. For example, recommendation systems [20] have been replacing information seeking activities. There are many efforts to bring AI into other spaces as well, including finance [25], transportation [166], healthcare [94], among other others. As the capability to collect data increases (e.g., wearables, mobile sensors, and the internet of things), the application areas for AI are

expected to increase significantly.

One of the key challenges of AI is that due to the complex nature of the underlying processes, understanding and comprehending AI becomes difficult for users [176, 141]. To address this challenge, human-centered approaches to designing AI have been proposed [154, 176, 5]. A key component of these approaches is helping users understand why and how an AI system arrives at a decision (prediction, advice) [69]. There is growing consensus that AI systems that are not able to explain their behaviors will not achieve the desired adoption and proper use [1, 89].

In this work, we investigated how users make sense of AI and integrate it into their decision-making process. Specifically, we focused on the role of domain knowledge as it is one of the key components for making sense of AI and appropriately utilizing it [169, 67, 24] yet an understudied area in AI research. We explored how domain knowledge can be leveraged to improve existing tools and methods that provide explanations regarding an AI agent’s decisions to improve users’ understanding and help them make more informed decisions when using the AI system.

## 1.1 Motivation and Research Questions

As AI technologies are being integrated into different fields, it is important to understand how users will interact with them and identify challenges associated with developing an appropriate understanding of AI. This becomes especially important if the AI is set to solve challenging problems and make high-stake decisions. For example, Dorton et al. [47] describes a real event that happened on an aircraft carrier: Before deploying the aircraft carrier off the coast of Virginia, a junior officer, who was working with an AI system that classified the contacts in the region, reported a dangerous aerial threat to the Tactical Action Officer (TAO) who was in charge of the ship’s weapon systems. Luckily, the description of the threat did not make sense to the TAC, and further inspection was conducted on the AI system’s classification. It turned out that the AI system was set to classify the contacts as highest possible threat as possible, but the junior officer was not aware of this. Since the AI system did not disclose why it classified a contact a high threat and what information it used to make the classification, the junior officer had no way of knowing if the AI should be trusted or not. Had this happened in contested waters, this event would potentially involve engaging weapon systems and have a catastrophic ending [47].

To address the issues with understanding of AI, growing number of researchers have been exploring the concept of Explainable AI (XAI) as a solution. XAI aims to develop

tools and techniques to help users understand and make sense of AI [144]. XAI is an interdisciplinary effort drawing from research in artificial intelligence, psychology, and human-computer interaction. At its core, XAI is concerned with developing techniques that can reveal why an AI made a prediction, and designing user interfaces that convey this information in an understandable way. This work is built on current efforts in the XAI field, however we adopted a slightly different approach to the problem.

One of the premises of modern AI is to provide decision-making support to individuals who deal with complex problems. These individuals are usually professionals who are considered domain experts (e.g., financial analyst, physician) and the AI systems that are developed are expected to augment their decision making [73]. However, there is growing evidence suggesting that even for professionals, lack of domain expertise can lead to over-reliance on AI and incorrect judgments. For example, less experienced physicians tend to perceive the quality of AI advice (even if it is incorrect) higher than more experienced physicians [60]. Similarly, less experienced physicians were more likely to adhere to erroneous AI advice [116] than more experienced physicians. These findings suggest that domain knowledge is an important factor for appropriately incorporating AI advice into decision-making. XAI approaches, on their own, do not close this gap, and may lead to misjudgments of the capability of the AI system [169, 150].

Additionally, recent developments suggest that AI that was developed for complex decision-making is making its way to everyday life, allowing non-expert users to make complex decisions without the need for a domain expert. Examples include AI-based investing advisors [174], mental health assistants [52], and personalized diet assistants [98], among others. In these cases, the problems associated with understanding the AI will remain an important challenge, therefore the need to research and implement XAI becomes essential. Research shows that among non-expert users (non-professionals), the level of domain expertise affects how AI is utilized. For example, users with less domain experience are less likely to identify incorrect AI advice [91], more likely to rely on AI advice [150], and trust AI more than users with more domain experience [129].

To address these issues, we argue that understanding the role of domain knowledge when interacting with XAI systems and integrating domain knowledge into XAI systems is critical. In this work, we focused on integrating domain knowledge into XAI and explored opportunities to support users and increase their ability to make sense of AI through domain knowledge. The overarching research questions in this thesis were:

- What is the role and importance of domain knowledge in human-AI interaction?
  - In what ways can domain knowledge be used to support users to make better decisions when working with an AI system?

- How does domain knowledge affect perceptions and use of an AI system?
- How can we leverage Cognitive Work Analysis in understanding and designing AI systems?
  - How can CWA be used to gain a deeper understanding of a human-AI interaction?
  - How can CWA be used to design better AI systems?

## 1.2 Background

In the following sections, we present a brief high-level overview of the main concepts that are the focus of this work and provide justification for research questions. Note that following chapters present more in-depth literature reviews pertaining to the work presented in the corresponding chapter.

### 1.2.1 From Human-Automation Interaction to Human-AI Interaction

Human-automation interaction has been a subject of attention in human factors research since the 1960s [74]. At that time, the primary function of automation was to carry out physical tasks. However, with the computerization of work, the focus of human-automation interaction evolved into supervisory control [151, 131, 173]. Typically, the problems of interest were in the form of levels of automation or proper role allocation between humans and automation. Early forms of automation were mostly used in complex technical systems such as power plants and aircraft. However, with recent advancements in AI, there is a new form of human-automation interaction that has much broader applicability than traditional automation systems. These aspects of automation have been captured, to some extent, in the expert systems literature [107].

As the focus of this work is AI systems, it is important to identify the differences between current AI systems and traditional automation. First, the type of automation typically studied in the literature was mostly concerned with types of systems that are designed for expert users, or for complex systems, such as power plants. However, current AI systems that are being built have many end-user properties, from recommendation systems to decision-making aids, such as intelligent investing assistants, personal healthcare assistants, and self-driving cars. Second, the traditional automation systems were mostly built

around rules and were mostly rule-based systems. Given an input (or multiple inputs), these systems followed certain decision paths constrained by hard-coded rules to provide an output. In most cases, these systems were less probabilistic and easier to understand. If the user knows the rules, they could understand how the input is transformed into the output. Of course, there were uncertainties, for example, sensor noise and similar environmental factors that added (external) uncertainty. Machine learning-based AI systems are probabilistic by nature. Usually, there are no explicit rules (or little explicit rules), and the goal is to train the machine learning models by letting them observe data that exists in the real world to learn the patterns and associations, and produce a reasonable input-output relationship. A classic example is an image detection algorithm that is able to detect faces by training over a large sample (could range from hundreds to millions) of images that may or may not include a face. Such algorithms learn, over time, to differentiate an image where there is a human face present, from an image where there is no human face.

Compared to rule-based systems, the learning process of current AI systems is not bounded by rules, and learning the patterns does not necessarily constitute rule-based reasoning. Since the model learns associations between data points, and since these tend to be more correlational in nature, the input-output relationship ultimately becomes probabilistic. This probabilistic nature has implications for users, as it becomes relatively difficult to understand how the model came up with a particular input-output relationship. Compared to traditional automation systems, this may require a different way of thinking and mental model of the system. In traditional automation, failures or breakdowns in the system can be traced easily by examining the rules that led to that failure, and appropriate action can be taken. In an AI system, however, finding faults in the system requires re-examining the algorithm itself, or the training data that was fed into the algorithm, and this process is relatively more complex. The uncertainty can also show itself in the form of ambiguity which is an essential part of these complex AI systems [148]. To deal with these issues, the field of XAI emerged in recent years, which we will discuss next.

### 1.2.2 Overview of Explainable AI

The term explainable AI (XAI) has become popular after a DARPA (Defense Advanced Research Projects Agency) program [70]. The goal of XAI research is to fill the gap in understanding the behavior of a machine learning model. In its conception, XAI was defined as “AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future.” ([70], p.44). XAI is concerned with both technical aspects (e.g., how to produce models that are easier to explain) and human-centered aspects (e.g., how to design an

explanation interface). Since the program launched, there has been a growing interest in developing techniques and approaches (see [1, 89, 119] for reviews).

In a review by [89], several explanation approaches have been identified, including using interpretable methods such as decision trees, model-agnostic methods, and example-based explanations, among others. Explaining the behavior of an AI model can take many forms. The explanations can be global (i.e., explaining the logic of the entire model) and local (i.e., explaining the logic of a particular prediction or decision). The explanations can also be model-specific, limited to certain models, or model-agnostic, applicable to many different models [1]. Studies that examined XAI techniques so far yielded mixed results. While in some cases, XAI seems to provide clear benefits [128, 108, 180], other studies found negative consequences of XAI [34, 96, 23]. It appears that there is more work to do, as acknowledged by DARPA in their retrospective analysis of the XAI program [72].

In this work, we took a functional approach to XAI. Instead of focusing on the AI technology itself, our focus was on the function it serves, and the role it plays in a decision-making situation. While we were interested in evaluating the AI, similar to the typical approaches used in the XAI literature, we were mostly concerned with the human performance at the end. Therefore, the focus was not on specific XAI tools or methodologies. We acknowledge that the current tools will evolve, and better techniques will be developed. However, we believe that two challenges will persist: First, what does it all mean for the user? How will it affect the way they solve problems? Second, how will a user make sense of the AI itself and any proposed explanation approach?

### 1.2.3 Trust in AI

One of the key motivating factors behind Explainable AI is to increase user trust in the AI system, especially when high stake decisions are on the table [70]. Ideally, the user should have appropriate levels of trust in the AI [103]. In human-automation interaction, trust in automation has been a fundamental concept [77, 103, 131]. Failure to calibrate trust in an automated system properly may result in misuse (overreliance) and disuse (underreliance) of automation [131], decreased performance and less adoption. Trust in automation has been extensively studied (See [77], for a review; [147] for a meta-analysis on factors influencing trust). One of the most influential approaches to studying trust in automation comes from Lee and See [103]. According to Lee and See, three factors are critical in trusting an automated agent: performance, process, and purpose. Performance is defined as a human’s observation of results, the process is defined as a human’s assessment of how the system works, and purpose is defined as the intention of the system. To establish

appropriate levels of trust, these three factors should match with each other in a human’s mind. For example, if the observed performance matches the human’s understanding of the system (process), then appropriate levels of trust can be developed. Trust and reliance on automation increase with the perceived reliability of the automation [145, 142, 120]. Trust can mediate the relationship between beliefs and reliance [51, 167]. It decreases with automation error [102, 18], but providing explanations of why the error occurred (observing the process; [103]) can increase trust and reliance despite the errors. Explaining behavior increases trust especially when the automation is less reliable [168]. Overall, the goal of system designer should not increase or decrease users’ trust without providing matching levels of automation reliability. This is key to achieving optimal performance.

In the context of XAI, trust in automation can be translated into trust in the AI system where the user must understand the process, observe the performance and evaluate if they match the intention of the system in order to trust in the system appropriately. Overtrust and overreliance on AI, similar to overtrust in automation, can lead to misuse of the system. Having appropriate levels of trust in the capabilities of AI algorithms would benefit users when interacting with the content suggested by these systems. For example, by calibrating their trust towards a particular AI decision-maker more appropriately, a user can decide when to rely on the AI, and when to ignore its suggestions. Moreover, if the user can make sense of why the AI system makes a prediction or a decision, they will be more likely to (1) calibrate their trust based on the explanations provided by the AI system (i.e. Is the explanation sound?), and (2) develop better mental models of the AI system (i.e. What is it sensitive to?). It is also worth noting that interacting with AI systems may not always lead to binary situations, i.e., whether to trust the AI or not. Since most AI systems are probabilistic in nature, trust can also affect the extent to which a user’s decision process is influenced by AI.

While there has been extensive research on trust in automated systems [77], and in trust in XAI [99], there is a lack of research on how domain knowledge influences trust in black-box machine learning algorithms. This thesis, therefore, aims to fill this gap by looking at how trust in AI is impacted when interacting with XAI.

#### 1.2.4 The Need for a Domain-Centric Approach

Regardless of how good an AI system is at explaining its behavior, model, and process, interpreting these outputs requires some domain expertise [11]. Most XAI approaches focus on a group of users who are experts in the task domain the AI is operating in [93] and make clear distinctions between domain experts and non-experts [4]. However,

domain expertise is not binary [157], and studies have shown that users with varying levels of domain expertise interact differently with XAI, and in most cases, non-expert users demonstrated poorer task performance and unwarranted trust in AI [60, 116, 91, 150], as they may struggle to make sense of the explanations [169]. Therefore, it is important that XAI systems go beyond explanations and provide reasoning facilities [46] as otherwise the user’s background domain knowledge can lead to inaccurate representations about why the AI makes a decision [93, 61].

Here, we argue that a domain-centric perspective can be useful, and this is the approach we adopted in this work. As the name suggests, the domain-centric approach starts with understanding the task domain, the expertise required to perform at an optimal level, and the challenges associated with decision-making in the task domain. This perspective is then used to understand and evaluate how the current users work, and to explore ways to apply the insights, findings, and recommendations to technology design (e.g., user interfaces) to support decision-making and task performance. We argue that this approach is critical in systems where non-expert decision-makers may rely on AI for decision support. By understanding what really matters in the task domain, we can develop tools to support non-expert decision-makers to (1) gain a deeper understanding of the problem, and (2) make better sense of the AI system they are interacting with. Our first research question is concerned with this approach:

**Research Question 1:** What is the role and importance of domain knowledge in human-AI interaction?

- In what ways can domain knowledge be used to support users to make better decisions when working with an AI system?
- How does domain knowledge affect perceptions and use of an AI system?

### 1.2.5 Cognitive Work Analysis as A Domain-Centric Tool

In this work, we used a CWA approach to develop a domain-centric approach and to understand and improve human-AI interaction. CWA is a framework to analyze complex systems and focuses on identifying the information support that helps users to become “flexible, adaptive problem solvers” ([164], p.136). A key characteristic of CWA is to understand the constraints of a system that reduces the degree of freedom while at the same time presenting opportunities for flexible adaptation.

CWA is a comprehensive framework that was developed to analyze and improve complex socio-technical systems [137, 163]. It was initially developed to understand the complexity



of work environments such as power plants, however the flexibility of the approach led to explorations of CWA in many different domains. For example, Read et al. [139] identified that CWA has been used many diverse fields including aviation, ground transportation, defence applications, healthcare, and finance. The unique characteristic of CWA is its focus on constraints of the work environment that ultimately determines the behavior of the actors (humans and automation) while giving them the flexibility to achieve the goals in multiple ways, which is key to deal with unexpected events. CWA offers five phases that address different types of constraints and dimensions of the work environment [122]. These include Work Domain Analysis (WDA), Control Task Analysis (ConTA), Strategies Analysis (SA), Social Organization and Cooperation Analysis (SOCA), and Worker Competencies Analysis (WCA). Below is a short summary of what each phase aims to accomplish:

- **Work Domain Analysis (WDA):** WDA aims to identify physical and goal-related constraints of the work environment, and has been the most influential phase in the history of CWA. WDA uses abstraction-decomposition space (ADS) to represent the functional relationships between elements in a task domain. ADS consists of a decomposition hierarchy and an abstraction hierarchy (AH, Figure 2.2). The decomposition hierarchy describes the part-whole relationships of a domain (i.e., systems and subsystems). The AH focuses on identifying the constraints at different levels of abstractions, including functional purpose (purpose of the system), abstraction function (rules and laws that limit the action), generalized function (functions that are achieved), physical function (physical, tangible components) and physical form (descriptions and appearance of physical characteristics) levels. According to Vicente [164], the number of levels of abstraction and the content is domain-dependent, however, it appears that historically these five levels have been useful to describe a wide variety of domains.
- **Control Task Analysis (ConTA):** ConTA aims to identify the constraints and requirements related to known, recurring situations [163]. Using decision ladders, ConTA describes the cognitive steps that are taken to achieve a specific goal. It also aims to identify expert performance by identifying particular steps the experts took or skip. ConTA has been used to understand working with automation in the past [104].
- **Strategies Analysis (SA):** SA aims to identify strategies that can be used to achieve certain goals. These goals are usually identified in the ConTA. For example, a ConTA can reveal what task needs to be done (e.g. risk assessment), and SA can

reveal different ways of accomplishing the task (e.g. various ways of conducting risk assessment). Compared to WDA and ConTA, SA is an understudied phase of CWA. The primary tool used in a SA is information flow maps. Compared to WDA and ConTA, there are fewer instances of SA in the literature [38].

- **Social Organization and Cooperation Analysis (SOCA):** SOCA deals with requirements associated with the organizational structure. It aims to identify how different actors (humans and machines) can coordinate and share responsibilities.
- **Worker Competencies Analysis (WCA):** WCA is akin to a user-centered approach where the requirements associated with users are identified. Through Skills, Rules, Knowledge (SRK) Taxonomy, WCA aims to map the constraints and requirements identified in other phases of CWA to human capabilities and limitations. WCA has direct implications for design, as it concerned with how information should be represented (e.g., on a user interface) that allows appropriate level of cognitive engagement (skill-based, rule-based, or knowledge-based).

While CWA has traditionally been applied to complex socio-technical systems such as power plants [163], transportation [15], and healthcare [16], it is a versatile framework that is applicable to a number of different fields, and to solve a number of different problems, including interface design, work design, and training requirements.

Ultimately, CWA is about understanding the work and identifying insights and opportunities to improve work. Therefore, the usefulness of this approach, similar to other approaches, is bounded by its utility and influence on improving the understanding of the context, and the ability to find opportunities to improve the tools, technologies, and workflows. This work presents an exploration of using CWA in the context of AI/XAI, and while using CWA to understand the dynamics of interacting with an XAI system and to explore the design opportunities, it also aims to present a case study for evaluating its usefulness. The following research question was developed with regards to CWA:

**Research Question 2:** How can we leverage Cognitive Work Analysis in understanding and designing AI systems?

- How can CWA be used to gain a deeper understanding of a human-AI interaction?
- How can CWA be used to design better AI systems?

## 1.2.6 Conceptualizing Domain Knowledge

What constitutes “domain knowledge” has been discussed extensively in the learning literature [2] and numerous ways to conceptualize domain knowledge has been proposed. For example, one way to conceptualize it is to consider declarative, procedural, and conditional knowledge about a subject matter. This could also include topic knowledge, discipline knowledge, and so on. This approach resembles the “textbook” knowledge [175] where domain knowledge is more about the body of knowledge about a particular field.

Another approach is to define domain knowledge from expert-novice differences in a task domain [33], which is the approach we adopted in this work. In this approach, domain knowledge is conceptualized by observing how experts reason and behave, and how they differ from novices. These differences provide insights into what constitutes domain knowledge in a field.

In this work, we conceptualized domain knowledge as the conceptual understanding, reasoning, and strategic actions that experts rely on to do their job. We operationalized it as “the decision rules and criteria used by experts to make credit and lending decisions”. When we mention “domain knowledge” in studies presented in Chapters 2, 3, and 4, we refer to a set of rules and guidelines that expert credit underwriters rely on when making credit approval decisions. These guidelines and rules can be external (e.g., set by an association in the lending industry) or internal (developed based on personal experience). In the CWA literature, these are usually described as “expert knowledge” [16, 163] and one of the goals of CWA is to elicit this knowledge by conducting field studies with experts and making it explicit through the analysis of the work domain.

## 1.3 Contributions

The work has a number of contributions to research and design. The main contributions included the following:

- We demonstrated that adding domain knowledge to XAI has a number of benefits to users. Through experimental studies, we explored how domain knowledge affects perceptions and use of AI in a financial decision-making task. We showed that non-expert users can benefit from having access to domain knowledge when interacting with an AI system. While the role of domain expertise in the context of AI and XAI have been studied in the past, to our knowledge, this is one of the first attempts

at supporting users to overcome the limitations of the lack of domain expertise. We also demonstrated that embedding domain knowledge into XAI is valuable in helping users avoid relying on AI when it makes a mistake. Taken together, these findings present opportunities to improve future AI systems (through integration of domain knowledge), as well as contribute to the growing body of research on domain expertise in XAI.

- We presented a novel use case of Cognitive Work Analysis (CWA). We leveraged Work Domain Analysis (WDA) to improve the existing explainable AI techniques. Furthermore, we applied Control Task Analysis and Strategies Analysis to understand how AI influences non-expert users' decision-making process. This work builds on and extends the CWA research on reasoning about automation [19] and modeling automation [104]. This work made a unique contribution to CWA research by investigating an understudied area (XAI) and by demonstrating the usefulness an under-utilized component of CWA (strategies analysis) while providing unique insights that can be leveraged to design future AI systems.
- Through user interviews, we identified perceptions of AI, and how AI is being integrated into decision-making by non-expert users. We identified a number of issues pertaining to XAI, and opportunities to improve design of XAI. Furthermore, our analysis revealed the complexity of designing for non-expert users, and recommendations for future research and design were developed.
- We introduced a testbed that simulates a realistic financial investing decision-making situation that can be used to explore future research questions in the investing space or human-AI interaction. We demonstrated different use cases for the testbed and explored multiple aspects of the human-AI interaction.
- This work also provides insights about how non-expert users engage in investing, and has implications for integrating AI systems to Peer-to-Peer lending platforms and similar investing contexts. We demonstrated how information presented on the user interface influences investors' decision-making processes and their investment decisions.

## 1.4 Thesis Overview

In the first half of this thesis, we explored how domain knowledge can be added to XAI and provided results of a series of experimental studies that investigated various aspects of

decision-making performance and perceptions of AI. In the second half, we took a broader approach to human-AI interaction and provided insights into how non-experts integrate AI into their decision-making process. We also utilized CWA to synthesize the findings and provided recommendations for design and future research. In addition to the brief overview of the literature presented in this chapter, each chapter includes a background literature section that focuses on the subject and the study presented.

This thesis is organized into six chapters, excluding this introduction chapter. A brief overview of each chapter is presented below:

- In Chapter 2, we present results from an online experiment that looked at the effects of augmenting explanations with domain knowledge on users' perceptions of XAI. This study also presents an exploration of utilizing WDA to produce domain knowledge explanations.
- In Chapter 3, we present an experimental study that investigated the effects of providing domain knowledge on task performance and perceptions of XAI in a complex financial decision-making context.
- In Chapter 4, we discuss an extension to the study presented in Chapter 3. An additional condition was tested using the same procedure and material, and the results were compared to the findings presented in Chapter 3.
- In Chapter 5, we present a CWA approach to synthesize the findings and insights gained so far and discuss implications of applying CWA to the current context.
- In Chapter 6, we provide a general discussion, contributions to research and design, the limitations we encountered, and our recommendations for future work.

# Chapter 2

## Study 1: Exploring Abstraction Hierarchy-Based Explanations

### 2.1 Introduction

In this chapter, we present a study that served as an initial exploration of CWA in the XAI context. The primary motivation for this study was to address the issue that users of XAI systems may lack the required domain expertise needed to make sense of the AI explanations, as the current XAI techniques provide only data-driven explanations for predictions without explaining the domain relevance. One of the immediate application areas of a WDA is that it provides a model of the domain knowledge, and in this way could serve as a way to identify domain knowledge that might be helpful to add to AI explanations.

In the following study, we explored the application of CWA in the context of XAI. We built an AI system using loan evaluation data set and applied an XAI technique to obtain data-driven explanations for predictions. Using an AH, we generated explanations that convey domain knowledge to accompany data-driven explanations. An online experiment was conducted to test the usefulness of AH-based explanations. Participants read financial profiles of loan applicants, the AI system’s loan approval/rejection decisions, and explanations that justify the decisions. Presence or absence of AH-based explanations was manipulated, and participants’ perceptions of the explanation quality was measured. The results showed that providing AH-based explanations helped participants learn about the loan evaluation process and improved the perceived quality of explanations. We conclude

that a CWA approach may increase understandability in explaining the decisions made by AI systems. This work was published as a conference article [44].

In the following sections, we present the study and discuss the key findings. More details about the study, including additional analyses, can be found in the Appendix.

## 2.2 Abstraction Hierarchy Based Explainable Artificial Intelligence

As Artificial Intelligence (AI) becomes integrated into complex socio-technical systems such as healthcare, finance and defence, there is an increasing concern that the black-box nature of these systems may result in adverse situations where the AI misbehaves or makes incorrect predictions in high-stake decisions. Explainable Artificial Intelligence (XAI) field emerged to combat these concerns and make the black-box AI systems more understandable by providing explanations that justify the decisions made by AI systems [70].

This work extends these efforts and explores a domain-driven approach to assist current XAI techniques by utilizing Cognitive Work Analysis (CWA). By combining domain knowledge from the Abstraction Hierarchy (AH) and data-driven explanations from current XAI methods, we present a unified solution that can provide better support to users of AI systems.

### 2.2.1 Background

When building user-facing AI systems, providing appropriate information regarding why the system behaves in a particular way is key to achieve trustworthiness and evaluation capability [117], which became more important after European Union’s regulations on “right to explain” [64]. While there are many definitions of explanations in the AI context, for the purpose of this work, we used explainability as “the ability to explain or to present in understandable terms to a human” ([48], p.2). XAI is a recent and an active research area, and several review articles have already been published (e.g. [119]). Currently, the most common method to achieve explanations is looking at the existing relationships between data attributes and finding evidence for the AI’s decision. Extracting these relationships is easier in linear models (ante-hoc explanation models; [82]) and more difficult in black-box models. However, recently, researchers started to unpack the black-box algorithms by

applying explanation models on top of black-box methods (post-hoc explanation models; [82]). These techniques (e.g. LIME; [140]) can be model-agnostic and locally faithful to the prediction model. Other approaches use different methods, however the common thread among these techniques is that they are data-driven. Below is a sample result from applying such an approach in a loan evaluation context.

**User Profile:**

Loan Amount = \$20,000, Duration = 24 Months, Savings Account Balance = \$5000.

**Decision:**

AI rejected the loan because Savings Account Balance was \$5000. If the user had at least \$10,000 in savings, the loan would be approved.

While these approaches can produce understandable explanations, as other researchers have pointed out [143, 82], explaining an AI agent’s decision using only data-driven approaches is not sufficient without the context, the domain relevance, and the reasoning behind the observed relationships. As AI systems are getting more involved in high-stakes decisions, providing domain-driven and comprehensive explanations becomes more important.

## Cognitive Work Analysis to Understand the Domain

Data-driven methods, in most cases, do not provide the reasoning behind the decisions. For example, a person may be denied a loan because the loan duration was longer than 24 months, however from a user’s perspective the importance of this may be difficult to grasp. A user facing such explanations will rely on their existing knowledge which may be inaccurate [119]. Therefore, providing the reasoning behind decisions by drawing from the domain knowledge, like an expert human would, can help users understand the decisions better. CWA [163], especially the first step, work domain analysis, can be particularly useful in analyzing and mapping the work domain and identifying the underlying reasons behind data-driven explanations. The “Why-How” structure of the abstraction hierarchy can be used as a knowledge representation framework [19] that provides rich and meaningful descriptions of how elements in a system are related. Such descriptions can explain why the data that is used as justification for AI decisions matters and the functions it serves. We argue that combining such descriptions with data-driven explanations can support users in (a) making sense of the explanations (b) considering alternative strategies when making action decisions. In the following sections, we present a prototype system and an experiment to explore how CWA can be used in the context of explaining decisions made by an AI system.



## 2.2.2 Overview of the System

To explore how AH can be applied to XAI, we created a prototype using a data set well known in machine learning research, the “German Credit” data set [49], which features 1000 loan applicant profiles and 20 variables such as account balance and loan purpose and labels for each profile (good or bad applicant). We used good and bad as “loan approved” and “loan rejected” in this study. To create the AI system, we trained a random forest classifier on the data (80/20 train-test split, AUC score = .65). Then, we applied the LORE technique (Local rule-based explanations; [68]) which samples data points that are close to an instance that needs to be explained, builds a decision tree, and extracts the decision rules for that instance and the conditions that would change the decision of a system (counterfactuals; [165]). In total, we selected 15 cases from the data set to explain and use in the experiment. The outcome of this process is shown in Figure 2.1.

Next, we built an AH to better understand the loan evaluation context, as shown in Figure 2.2. Our goal was to create a model that supports explaining the concepts, therefore the AH was strictly limited to the features of interest. The features in the data set were considered as physical functions. The generalized function level was the main emphasis in this model, which resulted in identifying various assessment processes and sub-processes. We used 5C Framework of Credit [8] as an approach to customer risk assessment. Abstract function and functional purpose levels were modeled but not utilized.











To generate explanations, for each factor mentioned in data-driven explanations (Figure 2.1), we created explanations based on the AH using the following logic: Find the element in the AH, find the function it serves, search for one other factor that achieves the same function, and finally, describe how these are connected to the decision (i.e. approval or rejection). Figure 2.3 illustrates the flow used to generate AH-based explanations. Once the explanations were generated, they were attached to the data-driven explanations and served as additional clarifications, as shown in Figure 2.4. Note that the descriptions provided by the AH are domain related but not necessarily AI related, and they serve to provide better understanding of the factors the data-driven techniques extract from the model.

## Overview of the Study

To examine the usefulness of using AH-based explanations, we designed an online experiment, manipulated the presence of AH-based explanations, and measured users’ perceptions of explanation quality.





## FINANCIAL PROFILE

---

 Chequing Account Balance.....	Less than \$2000
 Savings Account Balance.....	Less than \$1000
 Credit History.....	Existing credits at other banks
 Credits at this Bank.....	3
 Other Installment Plans.....	None
 Housing.....	Own
 Employment.....	Employed, more than 7 years
 Job.....	Unskilled work
 Most Valuable Asset.....	Real Estate
 Dependents.....	1 person

## LOAN DETAILS

---

 Loan Amount	\$11540
 Loan Duration	9 Months
 Loan Purpose	Appliances
 Other Debtors	None

## AI PROGRAM'S DECISION

---

 Loan denied. The probability of Approval was 10%.

## EXPLANATIONS

### MOST IMPORTANT FACTORS, ACCORDING TO THE AI PROGRAM

---

 Loan duration was **LESS THAN 21 MONTHS**.

### IF THE FOLLOWING WERE TRUE, AI PROGRAM WOULD APPROVE.

---

 Loan duration is **MORE THAN 21 MONTHS**.

Figure 2.1: One of the fifteen cases used in the study. Information about the loan applicant's financial situation, loan details, AI system's decision and explanations are presented on a single page.

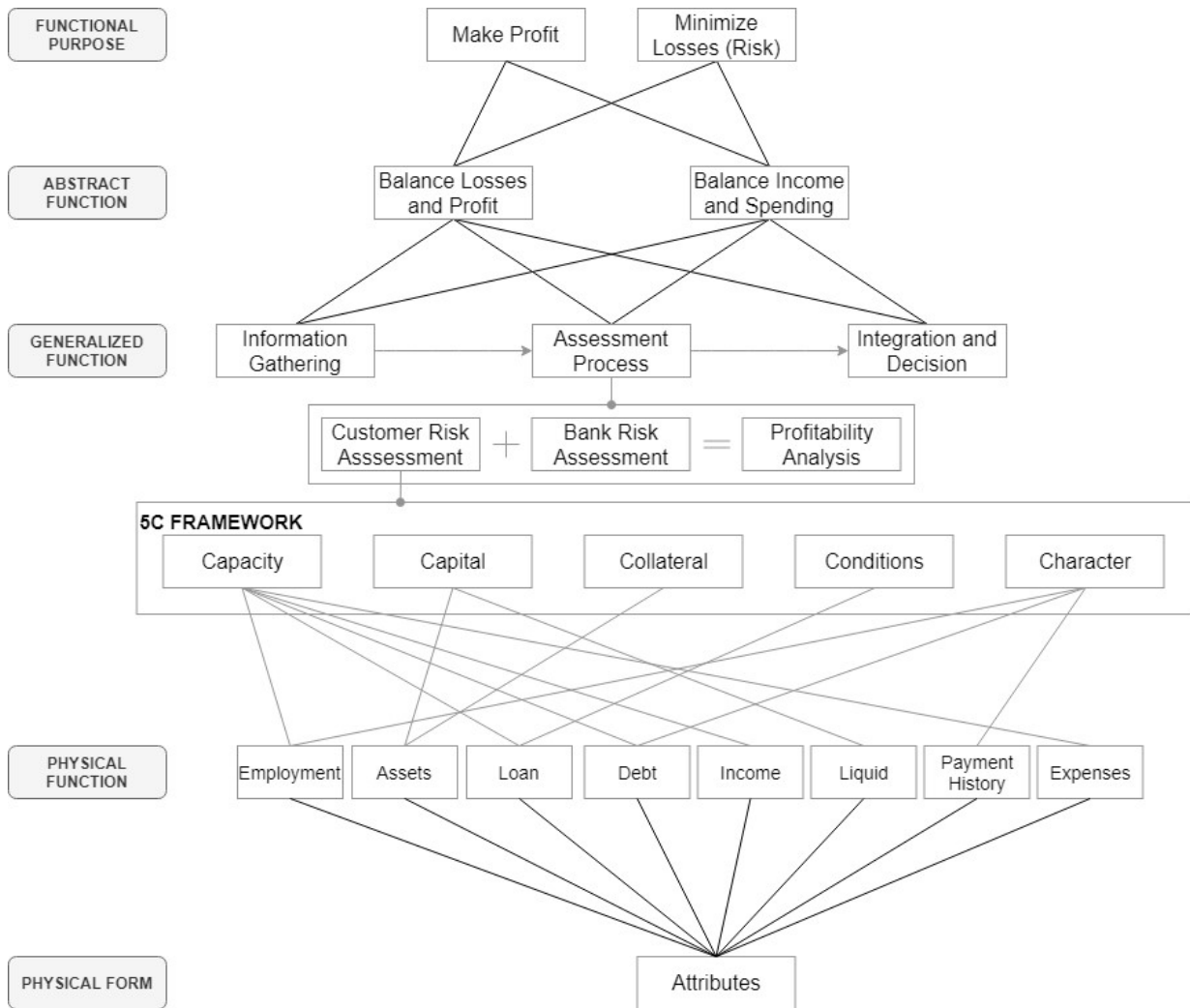


Figure 2.2: Abstraction hierarchy of loan evaluation system.

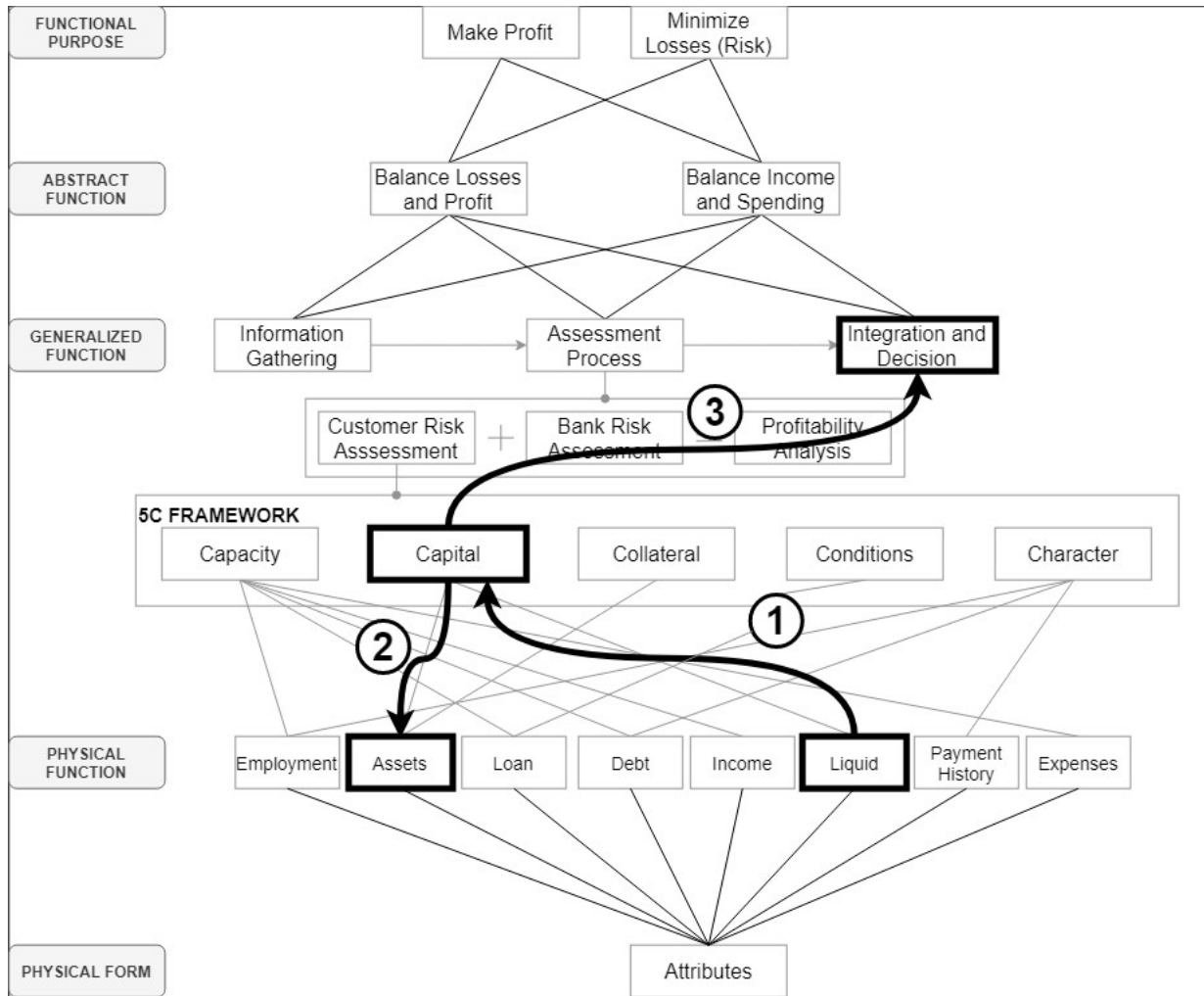



Figure 2.3: The flow used to generate explanations. In this example, liquid assets (e.g. account balances) achieve the generalized function CAPITAL (1), which can also be achieved using hard assets (2). CAPITAL influences decision (3) such that higher capital leads to better chances of approval.

## EXPLANATIONS

### MOST IMPORTANT FACTORS, ACCORDING TO THE AI PROGRAM

 Chequing account balance was **NO CHEQUING ACCOUNT**.

### IF THE FOLLOWING WERE TRUE, AI PROGRAM WOULD APPROVE.

 Chequing account balance is **LESS THAN \$2000**.

#### **CHEQUING ACCOUNT BALANCE**

**Chequing and Savings Account Balances** are good indicators of the **capital** the borrower can use to repay the loan if they are short on income. **Assets** such as properties can also be used for this purpose. In general, higher account balances and more valuable assets lead to more capital, and a higher chance of approval.

Figure 2.4: AH-based explanations. For each factor mentioned in the data-driven explanations, a description based on AH was presented (dotted box at the bottom).

## 2.2.3 Method

### Participants

We recruited 100 participants on Amazon Mechanical Turk platform. The mean age was 37.8 ( $SD = 10.5$ ). Gender distribution was 57% female, 43% male.

### Design

The experiment was a between-subjects design with two conditions: Baseline condition (data-driven only explanations) and AH-based explanations condition (data-driven + AH-based explanations). In the baseline condition, after presenting the predictions of the AI system, explanations generated using the LORE technique were presented, namely decision rules and counterfactuals (Figure 2.1). In AH-based explanations condition, decision rules and counterfactuals were presented similar to the baseline. Additionally, for each concept that was mentioned (e.g. “account balance”), a textual explanation about the concept’s domain significance (generated using AH) was presented (Figure 2.4).

## Procedure

The experiment was designed in a five-part survey format. A total of fifteen loan application cases (applicant’s financial profile information, AI model’s prediction, and explanations) were presented to participants in a static format. In the first part, demographics and background questions were asked and description of the AI system was provided. In the second part, five training cases were presented. For each case, after reading the applicant’s profile and explanations, participants had to answer simple questions (e.g. “What is the account balance?”). These questions were used as an attention test. In the third part, five case were presented, and participants were asked about their perceptions of the explanations. In both conditions, the cases included financial profile, AI system’s decision, and data-driven explanations. Additionally, in AH-based explanations condition, data-driven explanations were accompanied by AH-based descriptions. Of the five cases presented in this part, three were rejections and two were approvals. In the fourth part, five more cases were presented. This time, only financial profiles and decisions were presented, and explanations were omitted, therefore both conditions were similar. Participants were asked, for each case, to write what they would do to change AI system’s decision. The cases presented in this part were rejections, and participants had to figure out which property they would change and how much to get an “approval”. The final part of the study included open-ended questions about the AI system. Average completion time of the survey was 35 minutes.

## Measures

Table 1 shows the measures used in the experiment. Confidence in AI, human-likeness, adequate justification, and understandability questions were adapted from [53]. These four questions, along with the satisfaction question constituted user perception questions, and were asked after presenting explanations in the third part of the survey. Behavioral intention question was asked after presenting each case (without explanation) in the fourth part of the survey. Perceived learning and information amount questions were asked in the final part.

## Hypotheses

Based on the literature, we developed three hypotheses:

Measure	Response Form
<p><b>Confidence in AI:</b> The explanation makes me confident in the agent’s ability to perform its task.</p> <p><b>Human-likeness:</b> This explanation looks like it was made by a human.</p> <p><b>Adequate Justification:</b> This explanation adequately justifies the decision made by the agent.</p> <p><b>Understandability:</b> This explanation helped me understand why the agent decided as it did.</p> <p><b>Satisfaction:</b> This explanation was...</p>	<p>7-point scales, from “strongly disagree” to “strongly agree” for confidence, human-likeness, adequate justification, and understandability; from “unsatisfactory” to “satisfactory” for the satisfaction question.</p>
<p><b>Behavioral Intention:</b> If this were your loan application, what would you do to increase your chance of getting approval?</p>	<p>Open-ended response.</p>
<p><b>Perceived Learning:</b> How much did the explanations help you learn about loan decision process?</p>	<p>7-point scale, from “too little” to “too much”.</p>
<p><b>Information Amount:</b> The information provided when explaining the decisions was...</p>	<p>7-point scale, from “too little” to “too much”.</p>

Table 2.1: Measures used in the experiment.

*Hypothesis 1:* Participants in AH-based explanations condition will provide higher ratings in user perception questions than participants in the baseline condition. The literature on explanations [119] provide factors that make an explanation ‘good’, namely importance of causal reasoning, providing explanations that are suited to the user’s goals, and providing both explanations of local actions and global operations. AH-based explanations address several aspects of “good explanations”; they provide pathways users can choose based on their goals, global explanations through higher level abstractions, and convey more understandability by making the causal relationships in the domain explicit.

*Hypothesis 2:* Participants will report richer behavioral intentions in AH-based explanations condition compared to baseline condition. We defined richer behavioral intention as “mentioning more strategies in the behavioral intention questions”. The strategies could include actions such as increasing account balances, reducing debt, and so on. AH-based explanations provide information about both a feature’s significance in the domain and how features collectively achieve higher level functions. In other words, it provides a “map” where users can form multiple strategies to achieve the same outcome [56]. Participants in AH-based explanations condition will have a glimpse of this map through the richer descriptions and may consider alternative strategies better than participants in the baseline condition.

*Hypothesis 3:* Participants will report learning more about the loan evaluation process in AH-based explanations condition than participants in the baseline condition. While exposure to the data-driven explanations can help participants understand how the system works, with AH-based explanations, users have a better chance to form appropriate mental models regarding how the various elements relate to each other through the exposure to higher-level concepts and functional relationships identified in AH.

## 2.2.4 Results

Three participants could not complete the survey due to technical issues. Further, we remove one participant from the data as they failed in more than 50% of the attention questions. Therefore, the following analysis was conducting using data from 96 participants. Unless otherwise noted, comparisons were done using t-tests.

### Perceptions of Explanation Quality

Hypothesis 1 stated that AH-based explanations would be perceived superior than data-driven only explanations. To test this, we first collapsed all five cases and created average



scores for each of the five perception questions and conducted t-tests. Overall, the conditions did not differ significantly in user perception questions. Next, we examined the cases individually and observed significant differences in one of the cases (case 4). Specifically, there were significant differences in confidence in AI,  $t(95) = 2.01$ ,  $p = .047$ , and understandability,  $t(95) = 2.11$ ,  $p = .04$ . There was also a marginally significant difference in perceived human-likeness,  $t(95) = 1.78$ ,  $p = .079$ . Participants in AH-based condition reported more confidence in AI ( $M = 5.33$ ,  $SD = 1.32$ ), more understandability ( $M = 5.45$ ,  $SD = 1.16$ ), and more human-likeness ( $M = 4.63$ ,  $SD = 1.52$ ) than participants in the baseline condition ( $M = 4.75$ ,  $SD = 1.48$  for confidence in AI,  $M = 4.9$ ,  $SD = 1.39$  for understandability, and  $M = 4.01$ ,  $SD = 1.58$  for human-likeness). These results provide some evidence to support hypothesis 1, however it seems like there is a context effect as these differences were observed in only one of the cases.

## Behavioral Intention

The behavioral intention questions were coded in two ways. First, the number of concepts that were mentioned were counted. For example, if the answer suggested that loan amount should be lower and loan duration should be higher, the number of concepts mentioned would be two. Second, the answers were scored based on correctness. A t-test showed no significant difference in the number of concepts mentioned between conditions,  $t(95) = .02$ ,  $p = .98$ . On average, participants suggested 2.2 strategies in each case. There was a marginally significant difference in accuracy,  $t(95) = 1.73$ ,  $p = .086$ . The accuracy in the AH-based explanations condition was 60% ( $SD = .18$ ) and the accuracy in the baseline condition was 53% ( $SD = .2$ ). These results suggest that addition of AH-based explanations may improve the accuracy of predicting AI's behavior, however we should note that participants were incorrect in significant number of cases. Overall, hypothesis 2 was not supported as AH-based explanations did not improve the number of strategies mentioned by the participants.

## Perceived Learning

A t-test showed a significant difference between conditions in perceived learning,  $t(95) = 2.18$ ,  $p = .031$ . Participants indicated that they learned more about the loan decision process in AH-based explanations condition ( $M = 5.8$ ,  $SD = .97$ ) than participants in the baseline condition ( $M = 5.4$ ,  $SD = .81$ ). Hypothesis 3 was supported.

## Information Overload

A t-test showed that participants did not differ in their perceptions of the information amount between conditions,  $t(95) = .77, p = 0.44$ . Overall participants perceived that the amount of information was moderate,  $M = 3.9, SD = 1.03$  in the baseline condition, and  $M = 4.1, SD = 1.07$  in the AH-based explanations condition. This result suggests that addition of AH-based descriptions did not increase perceived information amount provided in explanations.

### 2.2.5 Discussion

In this experiment, we investigated the effects of adding AH-based explanations to the data-driven explanations on user perceptions of a loan evaluation AI system. The results showed that AH-based explanations improved perceptions of explanation quality in one of the five cases. Further, participants reported learning more about the loan evaluation process in the AH-based explanations condition. We did not observe a difference in behavioral intentions.

We found some evidence suggesting that adding an AH-based component to data-driven explanations can improve perceived confidence in AI, perceived humanness of the explanations and perceived helpfulness of explanations in learning about the AI system. However, it seems like the context matters as only one of the cases (case 4) revealed these effects. A comparison between case 4 and other cases showed that this case was one of the two cases where the AI system “approved” the loans, and the most important factor in approval was the assets the loan applicant had. A closer inspection of open-ended comments revealed that some participants were confused about the explanations involving assets. It is possible that providing domain knowledge-based information might be helpful in situations with particular characteristics (e.g. when familiarity with the situation is low), which is also consistent with what CWA aims to achieve [162]. It appears that AH-based explanations helped reduce the confusion some participants may have experienced with data-driven explanations. Still, these results call for a systematic investigation of contextual factors (e.g. ambiguity, complexity and familiarity) to identify the conditions where AH-based explanations are most useful.

We observed no differences in the number of strategies participants expressed to change the AI system’s decision. We believe two factors play a role here. First, the system was relatively simple, and some of the properties (e.g. loan amount and loan duration) were repeatedly mentioned in data-driven explanations. It is possible that, in this setup, providing data-driven explanations might be sufficient to obtain a good mental model of

how the system behaves. Second, the explanation generation process was kept simple to avoid lengthy descriptions. It is possible that providing more details may be needed in order to consider alternative strategies. For example, instead of limiting the explanations to only a few concepts, the entire AH could be revealed to participants and all of the pathways could be shown. Future work should systematically examine what kind of AH-based information is most useful in providing action support.

As hypothesized, participants in the AH-based explanations condition reported that they learned more about how the loan evaluation process works than participants in the baseline condition. These results support the idea that providing domain-relevant information can help users increase their knowledge and expertise. As black-box AI systems become part of the everyday life and work, non-expert users will increasingly have to deal with complex AI decisions and need to make sense of it. Providing domain knowledge along with data-driven explanations seems to be a promising approach in educating users about the problem space. Moreover, this approach can also be used in the training process in complex systems. Future domain experts who must work with black-box AI systems from day one can benefit from a tight integration of data-driven and domain knowledge-driven decision support systems.

Overall, these results are promising, and they invite further exploration of applying CWA in the explainable AI context and studying conditions under which a CWA approach would be most helpful in supporting users' understanding and decisions.

In this work, we explored a novel use case of CWA and utilized the AH to improve the quality of explanations of an AI system's predictions. CWA offers a strong toolkit to examine and analyze complex systems, and as AI systems start to tackle more complex problems, we believe CWA has an important role to play when it comes to supporting users of future AI systems.

## 2.3 Contributions and Conclusion

In this chapter, we presented an investigation of using AH to augment data-driven AI explanations. This study makes two contributions. First, we introduced a novel method to augment AI explanations with domain expertise. By taking advantage of CWA and the AH, we laid out the functional structure of the task domain (means-ends relationships) the AI is trained on. We then applied the relationships identified in the AH to a model-agnostic and local explanation technique to improve data-driven explanations. By combining data-driven and domain knowledge explanations, we presented a unified explanation interface

and identified how users' perceptions are affected by introducing domain knowledge into an XAI system. The second contribution was the way we utilized CWA. CWA has been traditionally used to gain a deeper understanding of the task domain, and the constraints identified in the analysis are then applied to graphical user interfaces (EID; [21]), training [121], and team design [16]. This work builds on the ideas presented in [19] and extends the current use cases of CWA and demonstrates that CWA, and in particular WDA can be beneficial in contextualizing the AI explanations by using the means-end relationships identified in the AH.

# Chapter 3

## Study 2: Investigating the Role of Domain Knowledge in Explainable AI

### 3.1 Introduction

In the previous chapter (Chapter 2), we explored how combining XAI with domain knowledge affects users' perceptions of explanations. However, the study did not adequately assess task performance in a realistic way due to the nature of the experimental setup. Understanding how domain knowledge influences perceptions of AI is valuable, however, it is not clear how it can help users to make better decisions and achieve better task performance. To address this gap, in this chapter, we present an experimental study that measured user behavior including task performance in a simulated environment. Similar to the previous study, we were interested in how integrating domain knowledge into an XAI interface affects user perceptions and user behavior, so we compared an AI-only situation with an AI + domain knowledge situation.

In this work, we explored how domain knowledge, identified by expert decision-makers, can be used to achieve a more human-centered approach to AI. We measured the effect of domain knowledge on trust in AI, reliance on AI, and task performance in an AI-assisted complex decision-making environment. In a peer-to-peer lending simulator, non-expert participants made financial investments using an AI assistant. The presence or absence of domain knowledge was manipulated. The results showed that participants who had access to domain knowledge relied less on the AI assistant when the AI assistant was incorrect and indicated less trust in the AI assistant. However, overall investing performance was not affected. These results suggest that providing domain knowledge can influence how non-

expert users use AI and could be a powerful tool to help these users develop appropriate levels of trust and reliance. This work was published as a journal article [45].

In the following sections, we present the study and discuss key findings. Additional information about the study, including additional analyses, can be found in the Appendix. Moreover, this study involved semi-structured interviews in addition to the experimental protocol. The analysis of the interviews will be discussed separately in Chapter 5.

### **3.2 The Effects of Domain Knowledge on Trust in Explainable AI and Task Performance: A Case of Peer-to-Peer Lending**

Today, Artificial Intelligence (AI) is being used to assist with complex decision-making in socio-technical systems such as finance and healthcare. Complex and capable AI systems utilize black-box algorithms [1], and there is an increasing concern about the negative consequences of using such systems. Not understanding how and why an AI system makes a decision may lead to distrust and under-reliance even if the AI advice is very accurate. Conversely, over-trusting a failing AI system can lead to devastating outcomes in high-stake decisions.

The field of Explainable AI (XAI) is trying to address these concerns by developing tools that help understand how an AI system makes a decision and what goes into it [69]. There are many different technical approaches to XAI [1], and the number of user studies on XAI is increasing. However, XAI approaches are driven more by revealing the workings of the AI, than by a human-centred approach.

This work aims to propose a human-centred approach to XAI by incorporating the domain knowledge of expert decision makers in an XAI interface. In particular, for non-expert users, gaps in domain knowledge can lead to misunderstandings, failure to interpret the explanations, and inappropriate levels of trust in and reliance on AI systems. Including the domain knowledge of expert users should allow less expert users to make better decisions.

To explore these issues, we present an experiment that examines the effects of providing domain knowledge to non-expert users on use, trust, and task performance when using AI in a complex financial decision-making situation. In the next sections, we will provide an overview of the related work, introduce research questions, describe our method, present the results and discuss the findings.

### 3.2.1 Background and Related Work

#### Explainable AI

It is crucial to provide information about why an AI system behaves in a particular way to establish trustworthiness and evaluation capability [117]. XAI has been proposed to achieve these objectives [1] by revealing information about how an AI system behaves and help users make sense of its predictions. XAI has multiple goals, including generating trust and understanding, satisfying legal and regulatory needs, ensuring social responsibility and fairness, among others [62].

XAI is a rapidly growing field [1, 89, 119]. In a review by [89], several explanation approaches have been identified, including using an interpretable method such as decision trees, model-agnostic methods, example-based explanations, among others. Instead of focusing only on explanations, recent approaches involve building AI systems that have interpretability and causability built into them by incorporating domain knowledge, allowing better reasoning about its output [86, 54]. Explaining the behavior of an AI model can take many forms. The explanations can be global (i.e., explaining the logic of the entire model) and local (i.e., explaining the logic of a particular prediction or decision). The explanations can also be model-specific, limited to certain models, or model-agnostic, applicable to many different models [1]. Each method has its distinct advantages and disadvantages, and recently comprehensive approaches to explainability have been proposed. For example, [105] lay out a question-driven framework focusing on understanding user needs and then selecting the appropriate explanation method. However, ensuring that XAI meets the needs of decision makers remains a significant challenge. Evaluating the quality of the explanations therefore becomes important to address this. There are many approaches that have been investigated to accurately evaluate the explanation quality (See [183], for a review on evaluating explanation quality). According to [48], explanations can be evaluated by experimenting with end-users in a real-world application setting (application-grounded), non-expert users in a simplified environment (human-grounded), or by a quantitative approach using formal definitions (functionality-grounded). Each approach requires defining a different set of measures to evaluate the explanation quality. For human-centered evaluations, one of the key metrics is trust towards AI [183].

In human-automation interaction, trust in automation has been a fundamental concept [77, 103, 131]. Failure to calibrate trust in an automated system properly may result in misuse (overreliance) and disuse (underreliance) of automation [131], decreased performance and less adoption. Trust in automation has been extensively studied (See [77], for a review; [147] for a meta-analysis on factors influencing trust). One of the most influential

approaches to studying trust in automation comes from [103]. According to Lee and See, three factors are critical in trusting an automated agent: performance, process, and purpose. Performance is defined as a human’s observation of results, the process is defined as a human’s assessment of how the system works, and purpose is defined as the intention of the system. To establish appropriate levels of trust, these three factors should match with each other in a human’s mind. Explanations in XAI can provide clues about the process, thereby helping users calibrate their trust towards AI appropriately.

User studies have shown some benefits of providing explanations on user perceptions and performance. For example, [128] showed that explainability features could reduce the time it takes to review summary case documents. [108] demonstrated that providing “why” (Why this prediction?) and “why not” (Why not the other prediction?) improved users’ understanding of AI, increased trust, and led to better task performance. Similarly, providing confidence metrics about an AI system’s predictions can lead to appropriate trust calibration [180], such as revealing the relationship between inputs and outputs [182]. On the other hand, the usefulness of explanations may depend on the cognitive load the user. [181] found that communicating the prediction uncertainty under high cognitive load led to lower trust in AI as participants did not have capacity to process the uncertainty information.

Explainability does not always work and can result in unintended outcomes. For example, it can lead to the illusion of explanatory depth [34], where users overestimate their understanding of AI by observing the explanations. Moreover, explanations can mislead users into believing that AI is more capable and correct than it is. For example, [96] showed that participants perceived an incorrect AI as more correct (or less incorrect) when explanations (post-hoc example-based) were present. Similarly, [23] demonstrated that healthcare practitioners over-relied on a clinical decision support tool when it was incorrect and when comprehensive explanations accompanied the diagnosis predictions. Conversely, participants also demonstrated under-reliance when the explanations were limited. Further evidence comes from [160], who showed that both rule-based and example-based explanations led to following an AI system’s advice more compared to no explanations even if the advice was incorrect. These findings highlight the complexity of explaining the predictions of an AI system for appropriate reliance. [90] demonstrated that clinicians performed worse when presented with an incorrect machine learning recommendation in a treatment selection study. Furthermore, in this study, providing feature-based explanations (i.e., contribution of features to a prediction) was associated with lower accuracy scores when the recommendation was incorrect.

These findings point to a potentially negative aspect of providing explanations: If the AI system fails to make a correct decision, explanations can lead to over-reliance and negative



task outcomes. This finding is a major concern, especially in high-stake decisions. We argue that to prevent these outcomes, accurate interpretation of explanations is important, which partly depends on having sufficient knowledge about a task domain. It is reasonable to assume that some users, when presented with elaborate explanations of a machine learning prediction, will be willing to perceive the explanations as a sign of AI’s capability, and this may lead to unwarranted trust in AI. This behavior is especially likely when the domain is complex and the decision makers are less expert. Therefore, providing domain knowledge to users may mitigate these explanation effects and help them make more accurate assessments of an AI system and its explanations.

## Domain Knowledge

No matter how good an AI system is at explaining its behavior, model, and process, interpreting these outputs becomes very difficult without necessary domain expertise. Paradoxically though, we often employ AI systems to help less expert users with complex decisions. Examples of AI for less expert users include healthcare AI assistants designed for patients, and financial investing AI assistants designed for retail investors.

Previous work suggests that domain expertise might be an essential factor in the context of XAI. For example, [169] found that when people lack domain knowledge, common explanation approaches such as feature importance and feature contribution did not satisfy the key goals of XAI, such as understanding, uncertainty awareness, and trust calibration. In another study, [150] found that participants with less task familiarity relied on AI more than participants with high task familiarity. Similarly, [129] showed that less expert users trusted an AI system than more expert users. Users with less expertise may also be prone to accepting AI advice when it is incorrect. For example, less experienced physicians were more likely to adhere to erroneous AI advice [116]. In another study, less experienced physicians considered the advice coming from AI as equally high quality as the advice coming from other physicians, while more experienced physicians perceived the quality of AI advice lower [60]. [180] concluded that task performance in the context of XAI may depend on factors beyond explanations, including “whether the human can bring in a unique set of knowledge that complements the AI’s errors” (p.296). These findings suggest that domain expertise in interpreting AI is important even in the case of domain experts.

One approach to address these challenges is to embed domain expertise in the user interface to improve the understandability and interpretability of AI recommendations. For example, [44] showed that adding knowledge-based explanations, derived from a cognitive work analysis, led to more positive perceptions of AI and better understanding of the situation. A similar approach is suggested by [109] in the form of “superimposition”,

which aims to map feature-based explanations to higher level concepts of a domain. [149] introduced the concept of “ambiguity-aware AI” which provides domain knowledge-based explanations about potential conflicting rules to facilitate medical reasoning. More comprehensive approaches utilizing domain knowledge have also been explored. For example, [43] used domain knowledge in the form of formal representations to build a hybrid AI with improved model accuracy that generated customized explanations.

While these studies propose promising approaches, they focus explicitly on providing model-driven explanations for the AI recommendations. This approach can lead to brittleness, as it supports only the user’s interpretation of the AI recommendation, rather than the user’s decision making process as a whole. In this work we propose taking a broader human-centred approach, starting with understanding how expert decision makers make their decisions, then embedding their domain knowledge in the interface. We hypothesize that this approach should help less expert users to identify failures of the AI and lead to appropriate trust and reliance.

## Decision-Making in Finance

The context of this work is Peer-to-Peer (P2P) lending. P2P lending platforms “provide a facility creating a marketplace where investors who wish to lend funds can find potential borrowers and provide credit through P2P Agreements” ([130], p.5). This objective is achieved by providing technology and infrastructure to facilitate interactions between investors (lenders) and borrowers. P2P lending platforms may offer a number of functions to assist borrowers and investors, such as identity verification, setting interest rates, money transfer from borrowers to investors, and ensuring legal compliance [130]. P2P lending platforms allow borrowers to obtain loans at lower cost and function as an alternative investing opportunity for investors as the risk-return trade-off and returns are close to other similar financial instruments [130]. In general, there are two categories of P2P lending: business and consumer. In business lending, investors lend money to (mostly small) businesses, whereas in consumer lending, loans are issued to individuals. The context of this study is consumer lending.

In its simplest form, P2P lending works the following way: Borrowers apply for loans by providing their information, and lenders can offer loans or fund the requested loans [7]. As the borrower pays back the loan, lenders earn interest. Depending on the website or the platform, the mechanisms through which the loans are funded may change. For example, a common approach is crowdfunding: Instead of each investor issuing a loan to one person, multiple investors can fund multiple loans by making smaller investments to reduce the

risk. Since the loans are unsecured in P2P lending in most cases, there is a considerable risk of losing the investments if borrowers default.

From a decision-making perspective, P2P lending provides a unique situation to study human-XAI interaction. Evaluating a borrower’s financial situation is key to ensure that the risk is manageable. While most platforms offer decision support tools such as assigning a grade to loan requests [65], at the end of the day, the decision to invest is left to investors. As in most lending environments, borrowers with a poorer credit history are subject to higher interest rates, which means higher returns for investors. Assessing the risk of a borrower requires an understanding of credit evaluation (i.e., domain expertise). However unlike banks, most investors don’t have access to advanced decision-support systems such as mathematical models financial institutions use. For experienced investors, assessing the risk of a borrower may be relatively easy; however, many P2P platforms are open to both experienced and inexperienced investors, creating a situation where the domain is complex (credit evaluation), but the users can be experts and non-experts. For non-expert users, decision-support tools such as AI may be beneficial, and AI and XAI have been explored in credit evaluation in the past [118, 66], including P2P lending [22]. However, the issues mentioned previously such as lack of domain knowledge, may be a barrier to make sense of XAI and use it appropriately. Therefore, in this study, we used a P2P lending platform as a testbed to explore the role of domain knowledge where non-expert investors interacted with an XAI system.

### 3.2.2 Research Questions and Hypotheses

The overall research questions we explored in this work are as follows:

- How does providing domain knowledge in the context of XAI affect task performance and reliance on AI?
- How does providing domain knowledge in the context of XAI affects perceptions of AI (trust in AI, perceived quality of explanations)?

Providing domain knowledge allows non-expert users to utilize domain expertise to evaluate the predictions of an AI system and interpret the explanations that support the prediction. Moreover, domain knowledge can help users make better assessments of the situation, affecting how the AI is being utilized in the decision-making process.

We argue that in situations where an AI system makes incorrect predictions, relying on domain knowledge can help users notice that the AI may be failing and help them avoid

relying on AI. In the absence of domain knowledge, users may be more inclined to trust and rely on AI as they have limited means to assess the performance of AI. Therefore, providing domain knowledge leads to more informed and accurate interpretation of AI output. From this perspective, we developed three hypotheses that reflect the potential outcomes of providing domain knowledge when interacting with an XAI system:

*Hypothesis 1a: Providing domain knowledge will lead to less reliance on AI when the AI makes incorrect predictions.* When an AI system makes a mistake, a meaningful way to identify the mistake and avoid relying on AI is by looking at the situation from a domain knowledge lens. For non-expert users, relying on domain knowledge is difficult. Providing domain knowledge may help them assess the situation more accurately and identify that the AI is not working correctly.

*Hypothesis 1b: Providing domain knowledge will lead to better task performance.* If participants can identify the situations where the AI is not working correctly and avoid relying on AI advice, this should lead to better task performance.

*Hypothesis 2: Providing domain knowledge will lead to less trust in AI.* Having access to domain knowledge can help users notice the discrepancies between the advice offered by the AI and domain-relevant guidance. They are observing that the AI is not in sync with their understanding of the task domain. This observation can affect the perceived capability of AI, leading to less trust, compared to having no access to such information.

*Hypothesis 3: Providing domain knowledge will lead to less positive perceptions of explanations.* Similar to *Hypothesis 2*, we expected that having access to domain knowledge will affect how well the explanations are received. Notably, we expected that users would have less confidence in AI explanations.

### 3.2.3 Method

#### Apparatus

To explore these research questions, we developed a website that mimics a P2P platform. We used a publicly available dataset to select borrower profiles and build a machine learning model. The dataset included over a million records of borrowers who obtained loans on a P2P lending platform. The records included loan details, financial background of the borrower, payments made and the loan status.

## Machine Learning Model

We used a subset of the dataset to build the machine learning model to predict the risk of default. We selected cases where the borrowers paid back the loan fully and cases that resulted in charge-offs or defaults. Whether or not the loan was paid back was the outcome variable. We trained a gradient boosting classifier (using the CatBoost library) on the dataset using an 80-20 split. This approach resulted in a model that had a 65% accuracy. Note that the model’s actual performance did not matter for the purpose of this study, as the cases that were shown to participants were handpicked from the dataset. This allowed us to present an AI model that had 75% accuracy among the cases shown to participants. Instead of using the predicted classes (i.e., default or not), we obtained the predicted probabilities (e.g., 68% chance of default) and used them throughout the study.

We consider the machine learning model as a black-box model for the purpose of this study. To explain the predictions of this model, we created the explainability component by applying the SHAP algorithm to identify each factor’s individual contribution to a particular prediction. SHAP is a state-of-the-art method for local explanations (i.e., explaining a specific prediction rather than overall model behavior, [110]). SHAP is an effective method that addresses multiple explanation needs [1], and it has also been used in credit evaluations in the past (e.g., [22]). Using SHAP allowed us to identify the contribution of each factor (direction and strength of contribution) to a default risk prediction (see Fig. 3.3, left panel for the outcome of SHAP). These feature contributions constituted the explanations.

## Website

We created a website as a testbed for this study. The website had three main components. On the loan listing page, participants could see available loans they could invest in (Fig. 3.1). These loans (5 in the training set, 20 in the main experiment) were selected semi-randomly from the dataset. On the borrower details page (Fig. 3.2), participants could see loan details (e.g., amount, term, interest rate), financial background of the applicant (e.g., income, debt, credit utilization). These factors are typically used to assess the creditworthiness of a borrower in the lending industry. On this page, participants could also invest in the loan fully or partially. Additionally, participants could hover over the question marks to read more about the financial terms.

We designed two visualizations that were key elements of the study: The machine learning assistant and a domain knowledge-based decision aid. The machine learning assistant (Fig. 3.3, left) included several pieces of information. First, it described the risk of default. The risk of default was the probability of default predicted by the machine learning

LW				Home	Loans	Portfolio	\$0	\$125,000	Logout	Review & Submit
							Currently Invested	Remaining Balance		
<b>Available Loans</b>										
#	Title	Asking For	Term	Interest Rate	Your Current Investment	ID	Actions			
1	Debt consolidation	\$6,400	36 months	14.52%	\$0	2b3b	<a href="#">DETAILS</a>			
2	Debt consolidation	\$6,000	36 months	14.08%	\$0	2b41	<a href="#">DETAILS</a>			
3	Debt consolidation	\$7,200	36 months	14.08%	\$0	2b45	<a href="#">DETAILS</a>			
4	Debt consolidation	\$16,000	36 months	15.99%	\$0	2b46	<a href="#">DETAILS</a>			
5	Credit card refinancing	\$3,600	36 months	15.31%	\$0	2b47	<a href="#">DETAILS</a>			
6	Credit card refinancing	\$15,000	36 months	14.99%	\$0	2b37	<a href="#">DETAILS</a>			
7	Debt consolidation	\$27,200	36 months	14.49%	\$0	2b42	<a href="#">DETAILS</a>			
8	Debt consolidation	\$18,000	36 months	15.99%	\$0	2b39	<a href="#">DETAILS</a>			
9	Debt consolidation	\$10,000	36 months	14.99%	\$0	2b38	<a href="#">DETAILS</a>			
10	Debt consolidation	\$12,800	36 months	13.67%	\$0	2b3e	<a href="#">DETAILS</a>			

Figure 3.1: List of available loans.

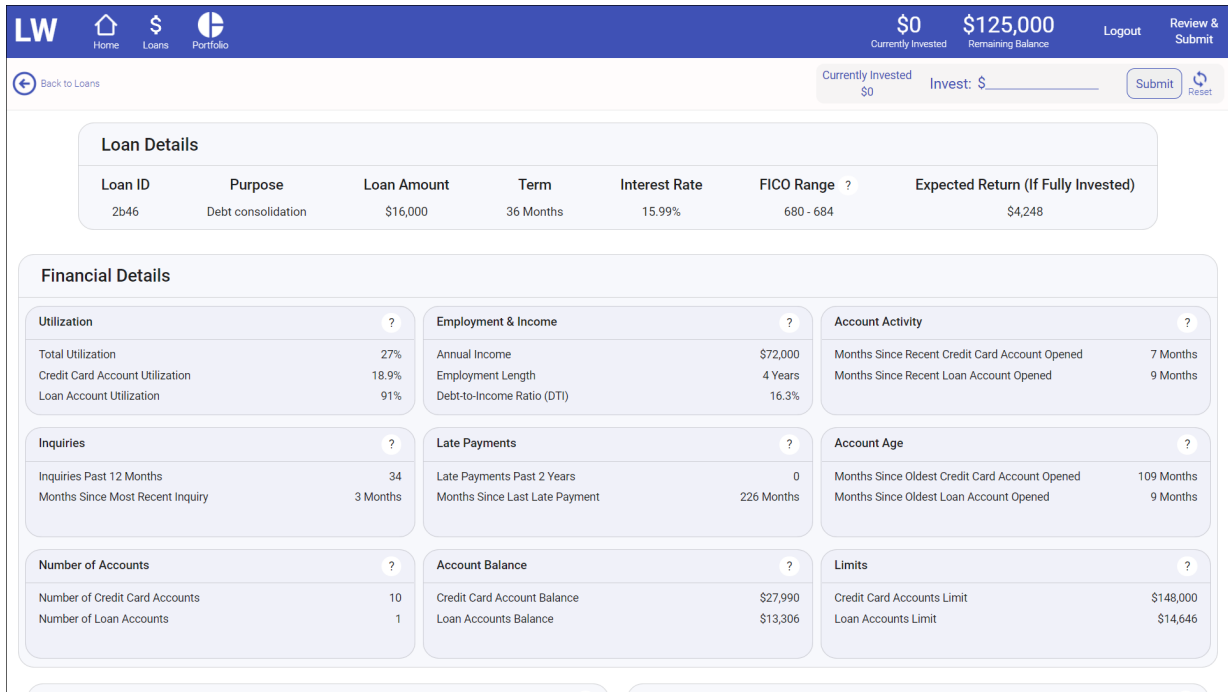


Figure 3.2: Financial profile of a borrower.

algorithm, and it could range from 0% to 100%. Next, the machine learning assistant provided the most important factor that contributed to this prediction. This information was obtained by applying the SHAP algorithm to extract the contribution of each factor to a prediction. Next, two graphs were shown to participants: Factors that decrease the risk and factors that increase the risk. Again, these factors and their contributions were provided by the SHAP method. We transformed the raw SHAP values into human-understandable percentages (e.g., Debt-to-Income ratio decreased the risk by 8% compared to an average applicant) to facilitate interpretation.

We designed a domain knowledge-based decision aid (key indicator panel; Fig. 3.3, right) to communicate domain knowledge regarding factors that affect the borrower's creditworthiness. Six key factors were identified through a domain analysis and talking to a professional credit expert. These six dimensions included credit utilization rate, debt-to-income ratio, number of late payments in the last 24 months, number of credit inquiries in the past year, employment length, and account age. For each of these factors, a threshold for making safe/risky decisions was identified. For example, generally, a credit utilization rate around 30% is considered a threshold. If the borrower has a credit utilization rate

higher than 30%, this might be regarded as a red flag and will decrease the chances of getting credit approval. Likewise, if the utilization rate is below 30%, the borrower can manage credit, which is considered a positive factor. The particular design we used aimed to show the ranges and where the borrower stands relative to minimums, maximums, and threshold values. The safe range (the range that most lenders prefer) is shown using green, and the risky range was shown using red. We used pins to show where the borrower stands. This information was not directly related to the risk prediction or AI explanations. Our goal with the key indicators panel was to give more context and provide a means through which participants could evaluate the explanations provided by the AI agent. For example, if a person's debt-to-income ratio was very high (shown in the red zone on the key indicators panel), and the AI cited debt-to-income ratio as one of the key factors increasing the default risk, participants were able to observe the match between AI's interpretation and expert interpretation of the borrower. In other instances though, AI failed to list debt-to-income ratio as one of the contributing factors, and participants might want to analyze the explanations further to understand the situation and interpret the AI more accurately.

Participants also had access to a portfolio page to see various charts to visualize the distribution of investments and expected returns.

## **Task**

The participants' task was to play the role of an investor, evaluate borrowers, and use the available virtual cash to make investments to maximize their profit. Evaluating the borrowers is a task that is very similar to a real world lending context: Given the credit history and financial background of the borrower, the default risk (or late payments) must be assessed. During this task, depending on the condition, participants had access to additional decision support tools: Machine learning assistant and key indicators panel. Participants could access additional information by hovering over the question marks on the website. Participants were allowed to spend their money however they wanted, and they could keep some (or all) of their money. They could use the machine learning assistant (and the key indicators panel in experimental condition);, however they didn't have to. In fact, during the training, they were told that the machine learning assistant may not always make the correct prediction, and it was ultimately their responsibility to take the risk. Participants could invest in any combination of loans out of the 20 loans listed on the website. There was no time limit, and participants were instructed to submit their portfolios when they were happy with their investments. Participants were not aware of the accuracy of the machine learning assistant as we wanted participants to form their opinions based on the information presented through the explanations and visualizations.





Figure 3.3: Machine learning assistant (on the left) and domain knowledge-based decision aid (on the right).

## Study Design

The study was a between-participants design with two conditions: baseline condition and experimental condition. The baseline condition involved AI-only scenarios, and the experimental condition involved AI + domain knowledge scenarios. In the baseline condition, participants saw borrower profiles including loan details, financial background of the borrower, and the machine learning assistant. In the experimental condition, participants saw everything present in the baseline condition. Additionally, participants in the experimental condition had access to the key indicators panel, and more information in the tooltips. In the baseline condition, the tooltips described what the term is, and provided the formula to calculate the value (e.g., “Credit utilization rate is the amount of credit divided by the credit limit”). In the experimental condition, the tooltips additionally stated why a factor matters or not, and how it is used by credit underwriters when assessing a borrower (e.g., “Having a high utilization rate (more than 30%) is a red flag that implies that the borrower is not managing debt well. If at any time this person’s income is gone, they will be left with lots of debt, which increases the risk of default.”). The tooltips in the experimental condition included substantially more domain knowledge about the factors, and we expected that reading the tooltips in the experimental condition will help participants gain a deeper understanding about borrower risk assessment. Note that the information presented on the key indicators panel was identical to the information presented in the tooltips, and key indicators panel served as a summary of the information that was available in the tooltips.

During the task, 20 available loans were presented. Half of the loans were safe (i.e., they were paid back fully), and the other half were defaulting loans (risky). Of the ten defaulting loans, five were correctly predicted by the machine learning model as high risk (the default chance > 50%), and the other five were incorrectly predicted as low risk (the default chance < 50%). As a result, the AI assistant had 75% accuracy.

## Sample

The sample consisted of 40 university students. The average age of the participants was 20.7 ( $SD = 3.1$ ). Nineteen participants identified themselves as male, and 21 participants identified themselves as female.

## Data Collection and Procedure

The study started with a pre-study questionnaire, followed by a training session. Participants watched a six-minute video during the training, introducing P2P lending, the

website, and the task. Then, participants completed a short training session where five loans were shown on the website. The composition of loans was similar to the loan composition in the main task. After the training session, participants completed the main task. The average task completion time was 22.4 minutes ( $SD = 12.7$ ). After the main task, participants filled out a post-study questionnaire.

## Measures

Several behavioral and subjective measures were collected. Except for background measures, all other measures were collected during the study or in the post-study questionnaire. Background questions were asked in the pre-study questionnaire.

### Behavioral Measures

We collected the following behavioral measures in the study:

**Portfolio Size:** The amount of money participants had at the end of the study after calculating losses and profits.

**The number of Safe Investments:** The number of investments made in safe loans.

**The number of Risky Investments:** The number of investments made in defaulting loans.

**The number of Incorrect AI Investments:** The number of investments made in defaulting loans where the AI incorrectly under-predicted the default risk.

**Safe Investment Amount:** The amount of money invested in safe loans.

**Risky Investment Amount:** The amount of money invested in defaulting loans.

**Incorrect AI Investment Amount:** The amount of money invested in defaulting loans where the AI incorrectly under-predicted the default risk.

**Time Spent on Tooltips:** The time participants spent reading the tooltips.

### Subjective Measures

We collected the following subjective measures in this study:

**Trust in AI:** A checklist for Trust between People and Automation [92] was used to measure trust in the machine learning assistant.

**The Perceived Accuracy of AI:** Perceived accuracy of the machine learning assistant was asked on a slider (0-100).

**The Explanation Quality:** Four 7-point Likert scale items were used to measure several dimensions of explanation quality: Confidence in AI (The explanation makes me confident in the agent’s ability to perform its task), human-likeness (This explanation looks like it was made by a human), adequate justification (This explanation adequately justifies the decision made by the agent), and understanding (This explanation helped me understand why the agent decided as it did). These items were adapted from [53] who developed these metrics based on the Technology Acceptance Model [41] and Unified Theory of Acceptance and Use of Technology Model [161].

**Information Amount:** Perceived information amount (the amount of information presented on the screen) was measured using a 5-point Likert scale.

**Task Difficulty:** Perceived task difficulty was measured using a 5-point Likert scale.

Additionally, we collected the following background measures:

**Financial Literacy:** Financial literacy was assessed using items from five instruments that measure various aspects of financial literacy [178, 57, 28, 80, 40]. We only used items that could be applicable to the credit and lending context (as opposed to specific financial instruments such as stocks and bonds). This approach resulted in a 24-item questionnaire.

**Risk Attitudes:** To measure risk attitudes, we used two instruments: DOSPERT (Domain-Specific Risk-Taking) scale [170] and SIRI (Stimulating-Instrumental Risk Inventory; [179]). From DOSPERT, we only used investment and gambling sub-scales. SIRI also provided two sub-scales: SRT (Stimulating Risk Taking) and IRT (Instrumental Risk Taking).

### 3.2.4 Results

We used Mann-Whitney U tests for ordinal variables. For continuous variables, we used independent samples t-tests if the assumptions were met. Otherwise, we used Mann-Whitney U tests.

We first looked at the time spent reading tooltips, as reading tooltips would mean that participants were exposed to domain knowledge. There was a significant difference between conditions in time spent reading the tooltips,  $t(38) = -4.7, p < .001$ . Participants in the experimental condition spent more time ( $M = 3.02$  minutes,  $SD = 1.61$ ) on reading the tooltips than participants in the baseline condition ( $M = 1.1$  minutes,  $SD = .87$ ). This

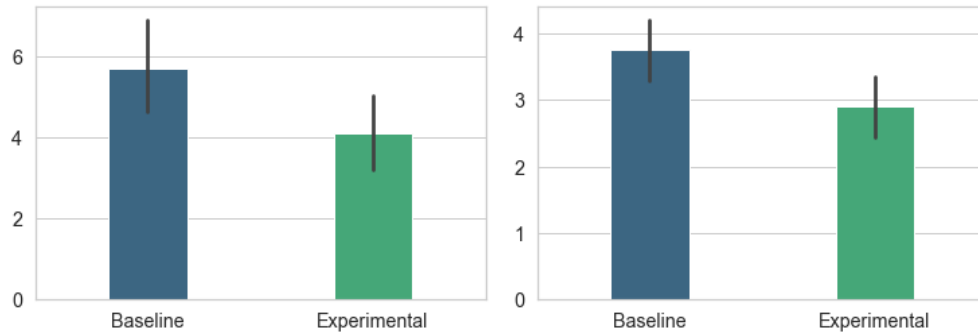


Figure 3.4: The number of risky investments (on the left) and the number of investments among loans where the AI was incorrect (on the right).

result indicates that participants in the experimental condition were exposed to domain-knowledge more than participants in the baseline condition.

### Task Performance

In this study, we considered two categories of task performance: The number of investments and the amount invested.

In terms of the overall portfolio (after calculating profits and losses), there was no significant difference between groups,  $t(38) = .23$ ,  $p = .82$ . The average portfolio size was \$98,011 ( $SD = \$9,280$ ). This result also means that participants lost around \$27,000 on average.

There was no difference between groups in the number of safe investments (Mann–Whitney  $U = 221.5$ ,  $p = .56$ ). However, participants in the experimental condition ( $M = 4.1$ ,  $SD = 2.2$ ) made significantly fewer risky investments than participants in the baseline condition ( $M = 5.7$ ,  $SD = 2.7$ ), Mann–Whitney  $U = 271.5$ ,  $p = .05$ . Moreover, among loans where the AI made incorrect predictions, participants in the experimental group made fewer investments ( $M = 2.9$ ,  $SD = 1.1$ ) than participants in the baseline condition ( $M = 3.8$ ,  $SD = 1.1$ ), Mann–Whitney  $U = 283.5$ ,  $p = .02$  (Fig. 3.4). These results support *Hypothesis 1a*.

In terms of the invested amount, there was no difference between conditions in the safe investment amount ( $t(38) = -.35$ ,  $p = .72$ ), risky investment amount ( $t(38) = .31$ ,  $p = .75$ ), and investment amount when the AI was incorrect ( $t(38) = .61$ ,  $p = .55$ ). *Hypothesis 1b* was not supported.

## Subjective Ratings

Overall trust in AI was not different between conditions,  $t(38) = 1.77, p = .09$ . However, when we explored trust and distrust dimensions of the trust scale separately, we found that participants in the experimental condition indicated lower ratings in the “trust” dimension ( $M = 4.53, SD = .95$ ) than participants in the baseline condition ( $M = 5.11, SD = .8$ ),  $t(38) = 2.1, p = .04$ . There was no difference between conditions in the distrust dimension,  $t(38) = .81, p = .42$ . *Hypothesis 2* was partially supported.

Although there was a 5% difference in perceived accuracy of AI between conditions, there was no statistical difference, Mann–Whitney  $U = 239.5, p = .29$ . Overall, participants perceived the AI as 77.1% accurate ( $SD = 10.73$ ), which is close to the actual accuracy of AI (75%).

There were no differences between conditions in explanation quality measures, all  $p$ 's  $> .05$ . Average ratings were:  $M = 5.4, SD = 1.13$  for confidence in explanations;  $M = 4.8, SD = 1.32$  for human-likeness of explanations;  $M = 5.6, SD = .78$  for adequate justification;  $M = 6, SD = .95$  for understanding. *Hypothesis 3* was not supported.

There was a significant difference between conditions in perceived information amount (Mann–Whitney  $U = 261, p = .05$ ). Participants in the experimental condition ( $M = 2.95, SD = .69$ ) indicated that there was less information presented on the website than participants in the baseline condition ( $M = 3.4, SD = .6$ ). There was no difference between conditions in perceived task difficulty, Mann–Whitney  $U = 177.5, p = .53$ .

## Correlations

Trust in AI was significantly positively correlated with most of the explanation quality measures: Confidence in explanations ( $r(38) = .55, p < .001$ ), adequate justification ( $r(38) = .46, p = .003$ ), and understanding ( $r(38) = .53, p < .001$ ). Trust in AI was also positively correlated with perceived AI accuracy ( $r(38) = .43, p = .006$ ). In terms of performance metrics, trust was positively correlated with overall number of investments ( $r(38) = .35, p = .03$ ). Trust was also positively correlated with the number of risky investments in loans where the AI was incorrect ( $r(38) = .53, p = .02$ ) and investment amount ( $r(38) = .47, p = .04$ ), however this relationship was observed only in the experimental condition.

Financial literacy was negatively correlated with risky investment amount in the experimental condition ( $r(38) = -.48, p = .03$ ) but not in the baseline condition ( $r(38) = .17, p = .48$ ). No other correlations between financial literacy and task performance metrics were significant. Financial literacy was also positively correlated with understanding

explanation quality only in the experimental condition,  $r(38) = .60$ ,  $p = .005$ . (baseline condition:  $r(38) = -.17$ ,  $p = .49$ ). No other correlations between financial literacy and other subjective measures were significant. Finally, there were no meaningful correlations between task performance metrics and risk-taking attitudes (DOSPERT and SIRI scores).

### 3.2.5 Discussion

In this study, we explored the idea of the domain knowledge gap when interacting with an XAI system. We examined the effect of providing domain knowledge on task performance, perceptions and use of AI. The results showed that participants relied on AI less when domain-relevant guidance was present, and they indicated less trust in AI. However, risk-taking behavior in terms of investment amount was not affected.

#### The Effect of Domain Knowledge

Participants made fewer risky investments in the experimental condition. This effect seems to be driven by fewer investments when the AI was under-predicting the default risk. We believe that providing domain-relevant guidance was the key factor here. First, the cases where the AI under-predicted the default risk all had one or more characteristics in the “risky zone” on the key indicators panel. This was intentional, as we wanted to create a situation where there was a discrepancy between the AI assistant and the domain-relevant guidance. For example, in one case, the borrower had over \$400,000 annual income and a low debt-to-income ratio (8.4%) and the default chance was predicted as 32% by the AI assistant. However, this borrower also had a high credit utilization rate (56%) and had eight credit inquiries in the past year. These factors were considered as red flags and were presented in the “red zone” on the key indicators panel. Looking at the key indicators panel, it was clear that there were red flags. Here, participants had to make a judgement: Follow the AI advice (as 32% risk was relatively low among the loan set we had) and invest or take the red flags indicated by the key indicators graph into account and avoid investing. It appears participants chose the latter option, as participants in the experimental group made fewer investments in such loans. Note that these red flags were mentioned in the AI explanations; however, the AI did not consider these as top factors.

We believe that there could be two explanations. First, looking at the domain-relevant guidance, participants may have noticed the red flags. This might have led them to consider the case carefully, or ignore it altogether (i.e., avoid investing), despite the low default risk prediction by the AI. In this case, the behavior would be primarily driven by the

presence of additional domain knowledge. However, informal interviews after the study session showed that almost all participants took the AI assistant into account on risky loans. Second, participants may have noticed the discrepancies between the AI assistant and key indicators graph (e.g., seeing red flags on the key indicators graph yet getting a low default risk prediction from the AI assistant), and question the AI guidance, resulting in less reliance on AI and more reliance on domain-relevant guidance. This explanation is also in line with the differences in trust in AI between conditions. Note that other than the five loans where the AI under-predicted the default risk, in all other cases, the information presented on the key indicators graph and the AI assistant were congruent, and following the domain-relevant guidance or relying on AI would not make a difference. These results suggest that conveying domain knowledge effectively reduced the over-reliance on AI when it was incorrect.

While we focused on the potential benefits of integrating domain-relevant information into an AI-based system, the findings also suggest that limited understanding of the domain (baseline condition) can lead to unwarranted trust and reliance. This finding is in line with previous research that showed that less expert users are more willing to accept AI advice than expert users [116, 60], and trust the AI more [129].

These results also highlight the complexity of closing the domain knowledge gap. Suppose an AI system provides domain knowledge to help users to make better sense of the explanations. In that case, it can also create situations where the AI decision or explanations may not match with domain knowledge. In our case, this led to a situation where participants could avoid relying on AI when it was incorrect. However, there could be unintended consequences in other contexts. For example, if the domain knowledge is misunderstood or is not complete, it can lead to incorrect interpretation of the AI system. This possibility challenges us to consider how much domain expertise people will need to make sense of AI and AI explanations, and how this domain knowledge will be obtained and communicated to the users or embedded into the AI system.

Unexpectedly, financial literacy was barely related to performance metrics. It is likely that the measures we used were not explicitly addressing P2P lending and credit evaluation, or were not effective in capturing the level of literacy that was needed in this task. While we observed some relationships between financial literacy and other metrics, it is difficult to draw conclusions based on the observed relationships.



## Investments and Risk Taking

We found that participants made fewer risky investments in the experimental condition, but the amount of money invested was not different between conditions. While people's risk-taking behaviors were not the study's main focus, we need to acknowledge several factors that may have contributed to these results. First, there was no real consequence of losing money in the experiment. In the informal interviews after the study, we noticed that most participants tried to make the right decisions in choosing creditworthy borrowers. Still, the amount of money they invested in these borrowers was not of concern. Second, most participants said that they would invest significantly less if they were using their own money. It appears that risk-taking attitudes were not at play in this study. Therefore, in future studies, it would be valuable to create an environment that involves real monetary risks and incentive mechanisms to accurately draw conclusions about risk-taking behavior.

## Subjective Measures

Participants in the experimental condition indicated less trust towards AI. We believe that these differences come from the presence of domain knowledge in the experimental condition. This further strengthens the argument that participants were able to identify situations where they should not rely on AI. The presence of domain knowledge may have resulted in perceptions that AI is not the only source of truth as opposed to the baseline condition where the only decision support tool was the AI assistant. This might have led to questioning the AI assistant more in the experimental condition, resulting in less trust and reliance.

Looking at the correlations, we believe that trust seems to be a central factor. Trust was both associated with behavioral measures such as investing in risky loans when the AI was incorrect and perceived explanation quality. These findings are in line with previous research on the importance of trust in automated systems and AI. [152, 103]. The observed correlation between trust and the number of investments made in loans that the AI assistant in the experiment condition incorrectly predicted suggests that trust might be the determining factor especially when there is conflicting information (i.e., the AI assistant and key indicators panel). In other words, situations where the AI might be failing may be the true test of trust in AI.

Perceived explanation quality was not different between conditions. These results may suggest that explanations were strongly tied to the prediction regardless of the accuracy of the prediction. It is possible that participants were looking for the congruence between the AI prediction and the explanations. Even if the prediction of AI was incorrect, as long as

the explanations were in line with the prediction, participants may have developed similar perceptions in both conditions. Most of the AI predictions in this study were of this type. It is also possible that perceived explanation quality may depend on and vary between cases; therefore an aggregate measure may not be the most effective metric. We believe that it is valuable to ask explanation quality questions separately for each prediction or a group of predictions (e.g., correct and incorrect predictions) in future studies. Finally, we only used a limited set of metrics to evaluate the explanation quality. Measuring the quality and effectiveness of explanations is an emerging area [72], and numerous approaches have been proposed [79, 183]. In particular, measures related to user mental models [79] including understanding the causability [83] can help better understand the nuances of explanation quality.

Surprisingly, participants in the experimental condition indicated that there was less information content on the website than participants in the baseline condition. Informal interviews after the study session revealed that some participants were overwhelmed with the amount of information presented on the website. It is possible that participants in the experimental condition were able to make more sense of the information presented as the domain knowledge was more complete and action-driven, therefore did not feel as overwhelmed as participants in the baseline condition. Domain knowledge and the additional visualization did not increase task difficulty, so in this case, providing these additional tools was both beneficial, and was not detrimental in terms of difficulty, task completion time, or information overload.

## Implications

These findings have several implications for designing future AI systems and future research in human-centered AI. First, we demonstrated that providing domain knowledge could be potentially useful to help users avoid some of the mistakes AI systems will make, especially if those mistakes are identifiable by relying on domain knowledge. This information becomes especially important if the AI is not very reliable.

Second, our findings suggest that domain expertise should be an important consideration when designing or testing AI systems. Considering that one of the areas where AI will be very useful is to support people with less expertise in a task domain, it is important to take domain expertise into account in future studies and build explanation facilities that don't lead to over-trust and over-reliance. In an ideal world, we would expect an AI system to implicitly model the domain, and make predictions that align with what we know about a domain. However, numerous examples and incidents have shown that this is not always

the case. Therefore, we see value in exploring communicating domain knowledge and AI to help users make the more informed decisions and develop appropriate trust in AI.

In human factors literature, information displays and visualizations based on extensive analysis of a task domain have been studied extensively [163, 164, 21, 12]. Usually, these information displays are built to allow reasoning at multiple levels and allow users to access the information they need at the right time to make complex decisions or monitor the status of a system and deal with uncertainties. We argue that a similar approach is worthwhile to explore in the context of XAI. Understanding what information users will need to make sense of AI explanations and allowing them to deal with unexpected situations (e.g., AI failures) can benefit future AI-involved systems.

## Limitations

This study had several limitations. First, we selected the loans such that relying on domain knowledge would always lead to a correct decision, whereas relying on AI would not. In reality, the situation would be more complex as there would be situations where a borrower with no apparent red flags (according to the lending industry best practices) may still fail to pay back the loan. This study did not account for these cases. This study also did not examine situations where AI over-predicted the risk (false positives) as these were not deemed critical cases in a lending context.

The accuracy of the AI assistant might be another limitation. While the accuracy of the AI assistant (75%) closely matched participants' perceived accuracy (77%), it is possible the benefits of domain knowledge may depend on the AI accuracy, therefore future studies should consider using models with high vs. low accuracy to better understand when domain knowledge is most beneficial.

The choice of the explanation approach is also a limitation, as we used only one of the many explanation techniques that have been developed. In our model, explanations provided by the SHAP method did not always make sense. However, since we handpicked the cases to present to participants, we were able to eliminate such explanations. Still, the type of explanation that are obtained from this technique (the relative contribution of each factor) may not be the most appropriate approach to explain the AI in our case. Future studies should consider other approaches, such as including causability in explanations [84].

Another limitation was that the investment volume was not realistic. In real-world P2P lending, investors usually partially fund loans and in small amounts. Rather than each investor fully funding a loan, a group of investors partially fund many loans, which reduces the risk for an individual investor. Since we didn't have the crowdfunding aspect

in our app and we wanted to give the participants the option to fully fund loans, we had to increase the amount of capital significantly. However, most participants mentioned that capital they had access to was not unrealistic, perhaps because they were non-expert investors.

We believe that having no real incentives or the risk of monetary loss affected participants' investment decisions. Future studies should consider creating a situation where participants would be faced with the risk of losing their real money, as this would help avoid situations where the participant was aware of the risk but did not care enough because it was virtual money.

Finally, this study focused on non-expert users. It is likely that non-expert users' expectations from AI are different than expert users who have significant domain expertise. Moreover, the way explanations are utilized might also be affected by expertise. Therefore, future studies should involve both expert and non-expert users.

## Future Directions

We believe several future directions are valuable to explore based on the findings from this study. First, an interesting question is how to synthesize domain knowledge and the outputs of AI and integrate domain-relevant information more tightly to the AI system. Several approaches have been proposed in the past that leverage knowledge graphs (e.g., [26, 27, 88, 54]). Such approaches are promising as they provide inherent interpretability and integration of domain knowledge into AI systems, which is critical to make sure users of future AI systems are better equipped to deal with AI failures. Second, this study has shown that perceptions of AI can be influenced by the additional information presented on the system's interface. This finding opens up the question of whether or not AI systems can be evaluated in isolation (i.e., decoupled from the actual system) as we expect additional information to be present for the foreseeable future in complex systems. Finally, some of the results we obtained might be influenced by domain-specific factors, therefore the role of domain knowledge when interacting with AI systems should be studied across a range of domains.

## 3.3 Contributions and Conclusion

In this work, we presented an experimental study where we examined the role of domain knowledge when interacting with an XAI system in a realistic decision-making scenario.

The findings highlighted the importance of domain knowledge in trust and reliance in the context of XAI, and demonstrated that providing users appropriate domain knowledge helps them avoid erroneous AI advice..

This chapter made several contributions. First, we presented a study of XAI in a realistic task context. While user behavior in XAI has been studied in the past, relatively fewer studies have taken place in a realistic context with a complex decision-making problem. Second, we demonstrated that having access to domain knowledge improved task performance, especially when the AI is incorrect. Third, we presented a method to leverage domain knowledge in the form of visualizations to help users make sense of the situation and evaluate AI from a domain knowledge perspective.

# Chapter 4

## Embedding Domain Knowledge in Explainable AI

### 4.1 Introduction

In the previous chapter (Chapter 3), we explored the effects of providing domain knowledge on task performance in an XAI context. The results revealed the beneficial effects of domain knowledge in dealing with imperfect AI, however we also identified opportunities to further improve the design of the AI. Based on the insights gathered from the interviews conducted as part of the study (presented in Chapter 5), we conducted a follow-up session with another group of participants and tested a third condition where the domain knowledge was embedded into the AI explanations. In this chapter, we present the extension of the previous study in the form of a new design, and present data from 20 participants and analyze how this condition compares to the previous two conditions tested in Chapter 3.

### 4.2 Background

We had several motivations for conducting this extension study. First, while we observed differences between conditions in the previous study, the experimental group was not different in terms of the amount of money invested in safe and risky loans, and overall portfolio. This suggests that perhaps providing domain knowledge, while useful, was not convincing enough to have an influence on the amount of risk participants were willing to take in terms of investment amount. During the interviews conducted after each session, we noticed a

common pattern: Most participants had significant respect for AI and in the capabilities of the AI system. While participants were aware that the AI might be incorrect, they stated that the AI likely knows more than them when it comes to this task. We believe that this “AI superiority” effect can be leveraged to communicate domain knowledge more directly, especially when the AI may be incorrect or there is a conflicting interpretation between the AI and the domain knowledge.

Instead of providing domain knowledge on a user interface component that is external to the AI and expecting users to compare AI and expert knowledge, perhaps a more effective method would be to integrate domain knowledge in AI such that the AI is capable of evaluating its prediction against expert knowledge and communicate possible mismatches to the user. Such an approach would be akin to communicating the limitations of the AI, and informing users about the possibility that the AI might not be working well. Previous research on automation transparency suggests that communicating information about the uncertainties of an automated system improved human-automation task performance [10, 30]. Moreover, explicitly communicating the situations where the automation has limited capability improved task performance and trust [115, 30]. This approach is also suggested by the *Situation Awareness-based Agent Transparency* (SAT) model [31] which emphasizes the importance of communicating the limitations to the user to help users make better sense of the automated agent and its capabilities.

Integrating domain knowledge into an AI system has been studied in the past from multiple angles. Some researchers proposed that by using ontologies and knowledge graphs that describe the domain(s) of interest, explainable AI systems can be more aware of the task domain, and the explanation process can be facilitated by contextualizing and providing explanations that are more understandable [26, 27, 58]. For example, [135] proposed a methodology to use semantic web to build knowledge-based explainable AI systems in the biomedical domain. Similarly, Lecue et al. [100] laid out several opportunities for a knowledge graph approach in AI, including improving AI systems’ reasoning, embedding causality and semantic connections. Other forms of embedding domain knowledge have also been suggested. For example, Balayan et al, [9] explored integrating a semantic layer into a neural network that allows embedding domain knowledge into the model building process. Similarly, Islam et al. [88, 87] explored the concept of infusing domain knowledge into a black-box machine learning model to enable domain-relevant and interpretable explanations. While these approaches are promising from a technical perspective, for the purpose of this study, we imagined what the end result of such an integration would be, and decided to focus on a situation where the AI assistant’s interpretation did not match with an expert knowledge interpretation. For example, if the AI predicted a low default risk but there were clear red flags according to expert knowledge, this might constitute a

limitation of the AI for the current prediction, or be an indication that the AI may not be performing well. In this case, communicating this information to the users may help them to question the capability of AI further, and may have significant influence on the decision-making process.

In this study, the embedded domain knowledge was used to highlight the discrepancies between the output of the AI (prediction and explanations) and the expert domain knowledge (decision rules and guidelines) that was obtained through CWA as discussed in previous chapters.

### 4.3 Research Questions and Hypotheses

In this work, we explored the following research questions:

- How does embedding domain knowledge into AI affect user perceptions?
- How does embedding domain knowledge into AI affect task performance?

Similar to the previous study, we developed several hypotheses about the proposed benefit of embedding domain knowledge into AI. These hypothesis are developed in relation to the previous conditions used described in Chapter 3. We hypothesized that if the domain knowledge is embedded into AI:

- *Hypothesis 1:* Participants will make fewer investments in loans where the AI made incorrect predictions than participants in the previous two conditions. If the information coming from the AI itself would be perceived as more valuable and important, it is possible that participants might be more hesitant to make investments in loans where the AI communicates the limitations and the uncertainty. Therefore, we expected that participants would make even fewer investments than participants in the previous study.
- *Hypothesis 2:* Participants will have better portfolios (more money after calculating profits and losses) than participants in the previous two conditions. Similar to the first hypothesis, we expected that if receiving communication from the AI itself will be perceived as a warning sign, then we would expect lower investment amounts in incorrectly predicted loans, and as a result, participants should have a higher overall portfolio than participants in the previous two conditions.



- *Hypothesis 3*: Participants will provide higher trust ratings for the AI than in the previous two conditions. In line with previous work [115, 30, 167], we expected that the AI that discloses its weaknesses would be perceived as more trustworthy than the AI that does not.

## 4.4 Method

The loan set, task, procedure, and measures were identical to the study presented in Chapter 3. For the sake of clarity, the condition names we adopted for this analysis was different than in the main study. Here, we will refer to *experimental* condition from the previous study (Chapter 3 as the *external* condition, signifying that the presentation of domain knowledge on a panel was external to the AI. The condition we report here is called the *embedded* condition as the domain knowledge was integrated into the AI interface.

### 4.4.1 Apparatus

To create the *embedded* condition, we designed the AI interface such that it would show a red box if there is a discrepancy between domain knowledge based interpretation and AI's own interpretation, and this applied to only situations where the predicted default risk was below 50% yet the actual risk was higher (the borrower defaulted). Figure 4.1 shows the user interface for the *embedded* condition. Note that in the *embedded* condition, the domain knowledge based decision aid that was used before, the key indicators panel (Figure 3.3), was not present. However, the domain knowledge presented in the tooltips was the same as in the *external* condition. For each case where there were red flags from a domain knowledge perspective, the red box provided stated that the prediction may not be accurate, and provided information about the red flags, Further information was available upon hovering over "Learn More", which provided domain knowledge information similar similar to the information presented in other tooltips.

The red flags indicated in the red boxes were the same red flags that could be obtained by observing the key indicators panel in the *external* condition (Pins falling in the red zone in Figure 3.3). The only difference was that the interpretation of these red flags were made explicit and baked into the AI interface to create the perception that it was part of the machine learning assistant, and not a separate module. These red boxes appeared in five cases where the AI made incorrect predictions by under-predicting the risk.

## Machine Learning Assistant

?

### Summary

The Machine Learning Assistant predicts that this loan has a **32%** default risk. The most important factor in this prediction was **Debt-to-Income Ratio** which decreased the default risk. Detailed explanations of the factors contributing to this decision are provided below.

**! This prediction may not be accurate.**

**Credit Utilization Rate is more than 30%. Most credit experts would consider this as high risk. [Learn More](#)**

**Number of Inquiries is more than 2. Most credit experts would consider this as high risk. [Learn More](#)**

### Explanations and Contribution of Each Factor

#### Factors that Decrease the Risk

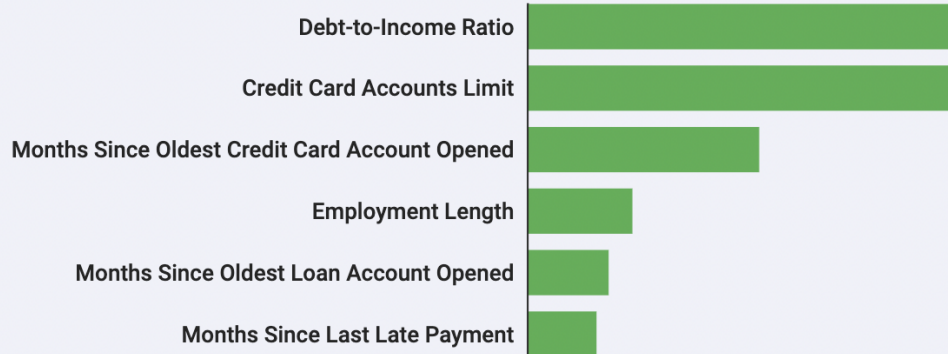


Figure 4.1: The machine learning assistant in the embedded condition. The red alert box was shown when the risk predicted by the AI was lower than 50% yet the borrower had red flags according to expert domain knowledge

## 4.4.2 Sample

The sample consisted of 20 university students. The average age of the participants was 22.5 ( $SD = 2$ ). Twelve participants identified themselves as male, and eight participants identified themselves as female.

## 4.5 Results

Since the primary purpose of this study is to understand how embedding domain knowledge into XAI affects task performance, the analysis was a comparison of the task performance observed in this study with the task performance observed in previous conditions. While we present a three-way comparison, we only highlighted the comparisons that were relevant to the *embedded* condition. The comparison of the *baseline* and *external* group was discussed in Chapter 3 and will not be further discussed. Instead, we present the analysis that pertains to the research questions and hypotheses introduced in this chapter.

For the following analysis, we report ANOVAs if the assumptions were met (with Tukey’s HSD post hoc tests). If the normality assumption was not met, we used Kruskal-Wallis tests for the main analysis and Dunn’s tests with Bonferroni’s correction procedure for post hoc comparisons. If the normality assumption was met but the samples had unequal variances, we used Welch’s ANOVA, followed by Games-Howell post hoc tests.

### 4.5.1 Task Performance

Similar to Chapter 3, we considered two categories of task performance: The number of investments and the amount invested. All means and standard deviations for task performance are shown in Table 4.1.

In terms of overall portfolio size (after calculating profits and losses), there was a significant difference between groups,  $F(2, 57) = 11.81, p < .001$ . A post hoc Tukey test showed that participants in the *embedded* group had significantly larger portfolios than participants in the *external* and *baseline* groups (both  $p$ ’s = .001), as shown in Figure

There were no significant differences between groups in the number of safe investments,  $F(2, 57) = .47, p = .62$ , and in the number of risky investments,  $H(2) = 4.53, p = .10$ . There was a significant difference in the number of investments among loans where the AI made incorrect predictions,  $H(2) = 10.88, p = .004$ . Participants in the *embedded* condition made

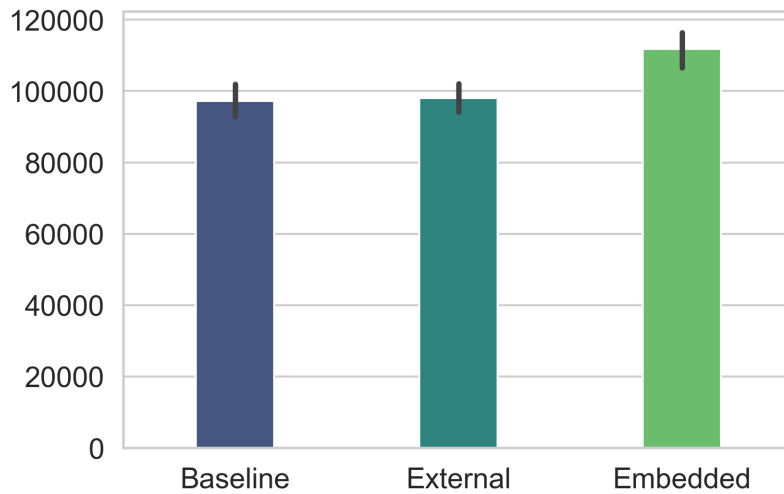


Figure 4.2: Differences in overall portfolio.

significantly fewer investments in these loans than participants in the *baseline* condition,  $p = .003$  (Figure 4.3).

In terms of the amount of money invested, there was no difference between groups in the safe investment amount,  $F(2, 57) = 1.82$ ,  $p = .17$ . However, there was a significant difference in the risky investment amount,  $F(2, 57) = 9.44$ ,  $p < .001$ . A Tukey post hoc test showed that participants in the *embedded* condition invested significantly less money in risky loans than participants in the *baseline* ( $p = .001$ ) and *external* ( $p = .002$ ) conditions (Figure 4.4). Furthermore, this difference seems to come from the differences in the investments made in loans where the AI made incorrect predictions,  $H(2) = 20.23$ ,  $p < .001$ . Participants in the *embedded* condition invested less in such loans than participants in the *baseline* ( $p = .001$ ) and *external* ( $p < .001$ ) conditions (Figure 4.4).

## 4.5.2 Subjective Ratings

Overall trust in AI was not different between groups,  $F(2, 57) = 2.32$ ,  $p = .11$ . Similar to the previous analysis (Chapter 3 Section 3.2.4), when we analyzed trust and distrust dimensions, we observed a significant difference in trust,  $F(2, 57) = 3.07$ ,  $p = .05$  (Figure 4.5). However, post hoc Tukey tests were not significant, all  $p$ 's  $> .05$ . The distrust

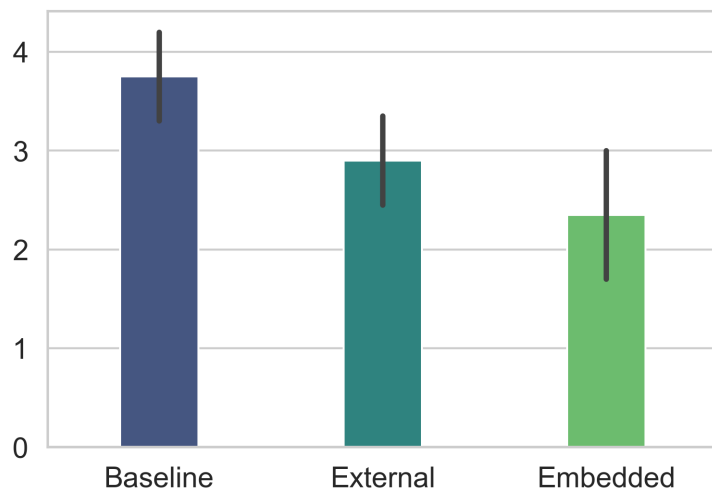


Figure 4.3: Differences in the number of investments made in loans where the AI made incorrect predictions.

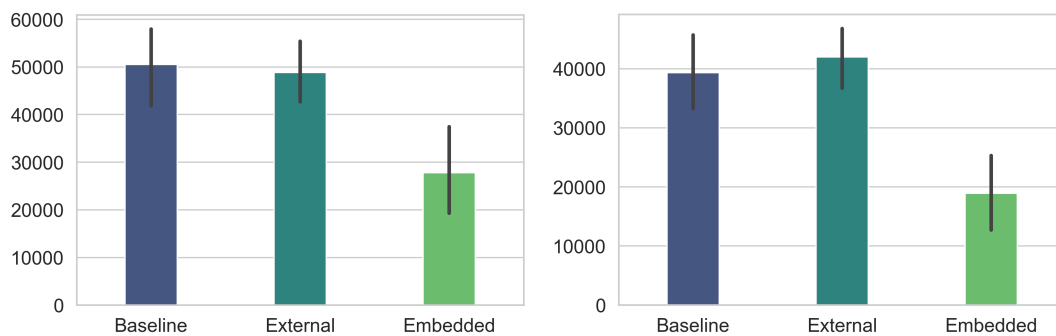


Figure 4.4: Differences in the amount of money invested in risky loans (left), and in loans where the AI made incorrect predictions (right).

Table 4.1: Descriptive statistics.

Measure	Baseline		External		Embedded	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Portfolio	\$97,269	\$10,535	\$98,011	\$9,281	\$111,847	\$12,054
Number of Safe Investments	7.5	2.44	7.1	2	6.75	2.83
Number of Risky Investments	5.7	2.72	4.1	2.17	4.1	2.85
Number of Incorrect AI Investments	3.75	1.07	2.9	1.07	2.35	1.5
Safe Investment Amount	\$52,471	\$17,336	\$54,250	\$15,212	\$43,834	\$22,187
Risky Investment Amount	\$50,483	\$17,969	\$48,842	\$14,848	\$27,784	\$21,784
Incorrect AI Investment Amount	\$39,348	\$14,979	\$41,973	\$12,199	\$18,942	\$15,085
Trust in AI	5.38	.70	4.94	.86	5.42	.77
Trust in AI - Trust Dimension	5.11	.80	4.53	.95	5.14	.88
Trust in AI - Distrust Dimension	2.26	.69	2.49	1.07	2.18	1.02
Perceived Accuracy of AI	79.8%	7.6%	74.37%	12.89%	72.35%	19.29%
Perceived Information Amount	3.4	.60	2.95	.69	3.5	.69

dimension was not different,  $F(2, 57) = .57, p = .57$ . The perceived accuracy of the AI was also not different between groups,  $H(2) = 2.59, p = .27$ .

The perceived information amount was different between conditions,  $H(2) = 6.73, p = .04$ . As shown in Figure 4.6, participants in the *embedded* condition indicated higher ratings than participants in the *external* condition,  $p = .04$ . Other subjective measures, including perceptions of explanation quality, were not significantly different, all  $p$ 's  $> .05$ .

Similar to the results reported in Chapter 3 there were no meaningful relationships between risk-taking attitudes (DOSPERT and SIRI scores) and task performance metrics. There was also no relationship between gender and task performance metrics.

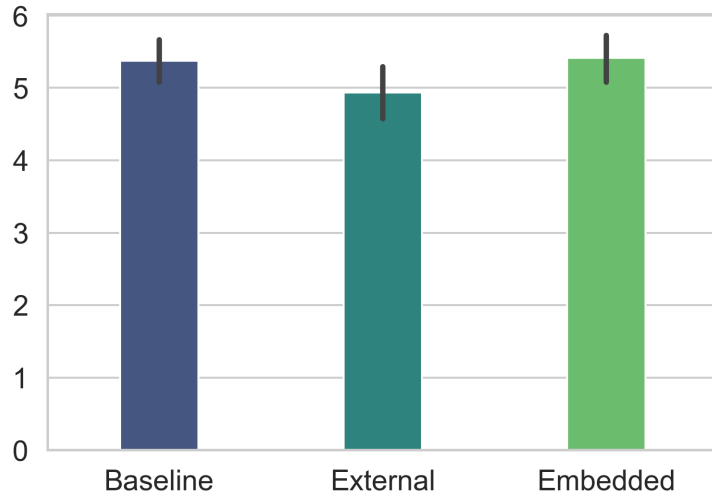


Figure 4.5: Trust in AI (trust dimension).

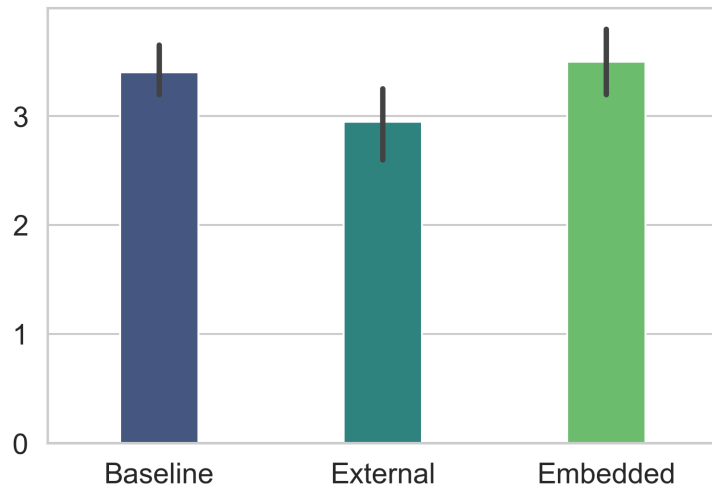


Figure 4.6: Perceived information amount on the user interface.

## 4.6 Discussion

In this work, we explored the effects of embedding domain knowledge into AI on perceptions and task performance, and compared the results with the results obtained from the previous study presented in Chapter 3. We found that in most metrics, the performance resembled the *external* condition, however the main difference was that participants in the *embedded* condition invested less in risky loans, and ultimately had better portfolio performance at the end of the study.

### 4.6.1 The Effect of Embedding Domain Knowledge

We found that participants in embedded condition outperformed participants in the baseline and experimental conditions in terms of the overall portfolio. These results are in a stark contrast to the previous results where providing domain knowledge on an external panel in the *external* condition did not make a difference in the amount of money participants invested compared to the *baseline* condition. In the *embedded* condition, it appears that the warnings coming from the AI were extremely powerful in preventing people from investing in those risky loans. These results support the idea that the communicating an automation’s limitations and indications of failures can improve task performance [31]. Moreover, the fact that the information was coming from AI instead of being displayed elsewhere may also have played a role. Compared to presenting the information on a separate panel, in this case, the information was presented directly on the AI interface, and this approach seems to be much more powerful. We also observed that, during the interviews after the study sessions, most participants indicated that they took the warnings seriously, and adopted a more cautious approach.

Note that except the “interpretation” that the AI prediction might not be accurate, the reasoning provided as to why it may not be accurate was similar to the interpretation depicted on the key indicators panel in the *external* condition. A main difference between the *embedded* and *external* conditions was that the the AI assistant in the *embedded* condition only presented explanations for red flags (factors where the borrower stood in the red zone on the key indicators panel; Figure 3.3). Another difference was that, in the *external* condition, the key indicators panel showed richer information with respect to how close or far away the borrower is from the acceptable thresholds.

Communicating the limitations of the AI provided a clear indication that the AI assistant should not be trusted easily. This in turn could have led to not investing in the borrower or carefully re-evaluating the situation. This is, to some extent, what we tried



to achieve with the key indicators panel in the *external* condition, where we expected the presence of domain knowledge to increase the likelihood of questioning the AI assistant. However, it looks like simply providing domain knowledge to less expert users may not be sufficient to reduce over-reliance on AI completely. Instead, explicitly pointing out the possible mismatch between domain knowledge and AI prediction seems to be much more effective.

### 4.6.2 Subjective Measures

Positive trust results were significant, however the post-hoc tests were not significant. These results could be due to lack of power of post hoc tests, or the significant omnibus ANOVA might be a false alarm [158]. It is difficult to draw conclusions based on these results, however we don't think there is convincing evidence to support *Hypothesis 3*. Based on the automation transparency research [32, 29, 30, 115] we expected higher trust ratings in the *embedded* condition as the AI assistant communicated situations when it was not reliable. One explanation might be that although the AI was communicating when it was not reliable, the justification for the unreliability involved how credit experts would assess the borrower. It is possible that participants may have attributed the unreliability information to the website as a whole rather than to the AI system, which may have affected the trust ratings. Perhaps more direct measures related to the AI model could be used, such as uncertainty of predictions [13]. Nevertheless, understanding how trust is shaped in the XAI context requires more research, as there are conflicting findings in the literature, e.g. [140, 180].

Surprisingly, perceived information amount in the *embedded* condition was higher than *external* condition, but was not different from *baseline* condition. These results were somewhat surprising, as in terms of user interface elements, *external* condition had the highest density, yet received the lowest ratings. One explanation is that the key indicators panel, through visualizations, made it easier to process the available information. As some participants indicated, it provided the opportunity to quickly glance and understand the situation of the borrower.

### 4.6.3 Implications

This work demonstrated an effective method to integrate domain knowledge into AI. This approach involved explicitly warning the users when there were red flags about a borrower yet the AI assistant outputted lower default risk predictions. We believe this approach

will be useful to deal with an imperfect AI. As we will discuss in Chapter 5, participants relied on a limited understanding of credit and lending to evaluate the AI predictions and explanations. Leveraging domain knowledge seems to improve this evaluation process. In the foreseeable future, we expect AI systems to be used along with existing decision-support systems that were built based on domain expertise, and this study presented a method to integrate these two types of systems. Moreover, the findings also suggest that using AI as the primary decision-support system that is able to integrate domain knowledge might be an effective design approach. The findings also suggest that communicating the reliability information might be an effective method to help users rely on AI appropriately. While our goal was to warn the users based on the domain knowledge and not AI related factors such as model uncertainty [13], communicating the reliability, regardless of the source, might be effective. This is in line with previous work on disclosing the reliability of an automation system [142].

#### 4.6.4 Limitations

Looking at the findings, it can be argued that the communicating that the prediction might be inaccurate might have played an important role. However, both before and after the training session, the fact that the AI may be inaccurate was emphasized multiple times. In fact, interviews conducted after the session revealed that most participants were aware that the AI might be inaccurate in not only the cases where there were explicit warnings but in other cases as well. Furthermore, some participants indicated that upon seeing a warning, they assessed the importance of it by looking at the justification for the warning (embedded domain knowledge). These suggest that perhaps both the presence of the warning itself and the domain knowledge played a role in this study, and the data is not conclusive as to which factor was more influential. Therefore, we suggest that future studies should examine these factors in isolation.

Moreover, the specific language, iconography, and visual design elements may have played a role as well. The design decisions were made to establish consistency between the *external* and *embedded* conditions, such as using the red color to indicate the red flags about a borrower, however we acknowledge that this is a limitation, and future studies should explore multiple designs and language to embed domain knowledge in XAI.

### 4.6.5 Future Directions

As discussed before, future studies should systematically examine the language and visual presentation of reliability warnings. Furthermore, type of reliability information should also be studied. In this work, we utilized domain knowledge to communicate these warnings, however it is not clear how well other types of reliability information would be perceived, or if additional information is needed at all. Perhaps simply warning the users without giving context could achieve a similar effect. These issues should be addressed in future studies.

## 4.7 Contributions and Conclusion

In this chapter, we described our approach to embed domain knowledge in an XAI system, and presented a comparison of task performance between this condition and the previous two conditions that are tested in 3. This study extends the findings from the previous study. The primary contribution of this study was demonstrating how embedding domain knowledge in XAI can help users to reduce relying on AI when it is incorrect, especially when domain knowledge is used to “audit” the AI explanations. We showed that highlighting the mismatch between the AI assistant’s assessment and expert knowledge can help users to critically evaluate the AI assistant and avoid following incorrect advice.

# Chapter 5

## Using Cognitive Work Analysis to Understand Decision-Making with AI

### 5.1 Introduction

In this chapter, we situate the findings presented in earlier chapters in the context of CWA, and provide a discussion how CWA can be used and extended in the context of AI and XAI. Our goal in this chapter is to synthesize the findings from a CWA perspective and discuss the implications for using CWA in future AI studies, and identify areas where CWA can be used in designing AI and XAI systems.

We also support the analysis by sharing insights from the 30-minute long semi-structured interviews that were conducted after each session in studies presented in Chapters 3 and 4. These interviews mainly focused on the lending task and aimed to understand how participants used different pieces of information that were available to them.

### 5.2 Method

To understand how investments were made in our experimental study, we conducted ConTA and SA based on the qualitative and quantitative analyses presented in earlier chapters. The goal of ConTA was to understand the workflow of the lending task, and the goal of SA was to identify strategies used by the participants. We only used these two stages in this chapter as these were the most relevant in our context. We already discussed

the application of AH to this context in Chapter 2. Our studies did not involve a social structure, therefore SOCA was left out. Similarly, we didn't conduct WCA as design was not within the scope of this analysis. ConTA and SA were used descriptively to better understand the findings from the experimental studies.

### 5.3 Modeling Workflow with AI: Control Task Analysis

We used decision ladders to model the cognitive processes involved in the credit risk assessment task. The decision ladder is a diagram-based tool developed by Rasmussen [136] to map the mental activity in a sequential way. These activities include data processing steps (boxes) and the state of knowledge that emerges from those activities (circles).

A general workflow is shown in Figure 5.1. The flow starts with an ALERT which represents opening up a borrower's file. Next, the details about the loan, financial background of the borrower need to be reviewed, which is represented in the OBSERVE step. This step also involved reviewing the AI assistant and any other information such as the key indicators panel. Next, positive and negative factors about the borrower need to be identified, and their risk of default needs to be assessed. This is depicted in the IDENTIFY state. Once the risk is identified (SYSTEM STATE), options about the possible actions need to be evaluated. In this case, the evaluation may entail investment strategies. For example, full investment can be made where the borrower receives all the money they were asking for. However, based on the participant's goals (represented as GOALS), they may choose to diversify their portfolio and make a partial investment. Of course, they need to take into account the interest rate, as well as other opportunities (i.e., other loans which may have higher yields). Once a decision is made (GOAL STATE), the rest involves deciding on the task (TASK) and actions (PROCEDURE) to complete the investment. In our app, this process was straightforward, therefore we will not discuss the right leg of the decision ladder (from TASK to EXECUTE) in this analysis.

Based on our observations and analysis, we identified that the IDENTIFY part of the decision ladder was the most complex and the core part of the task. Identifying risks involved assessing how financial factors come together and creating a picture of the borrower. To achieve this, participants had access to the financial background of the borrower, the AI assistant, tooltips that help with the financial factors (domain knowledge), and depending on the condition, the key indicators graph (*external* condition) or the domain knowledge-based warnings (*embedded* condition).

# P2P Lending

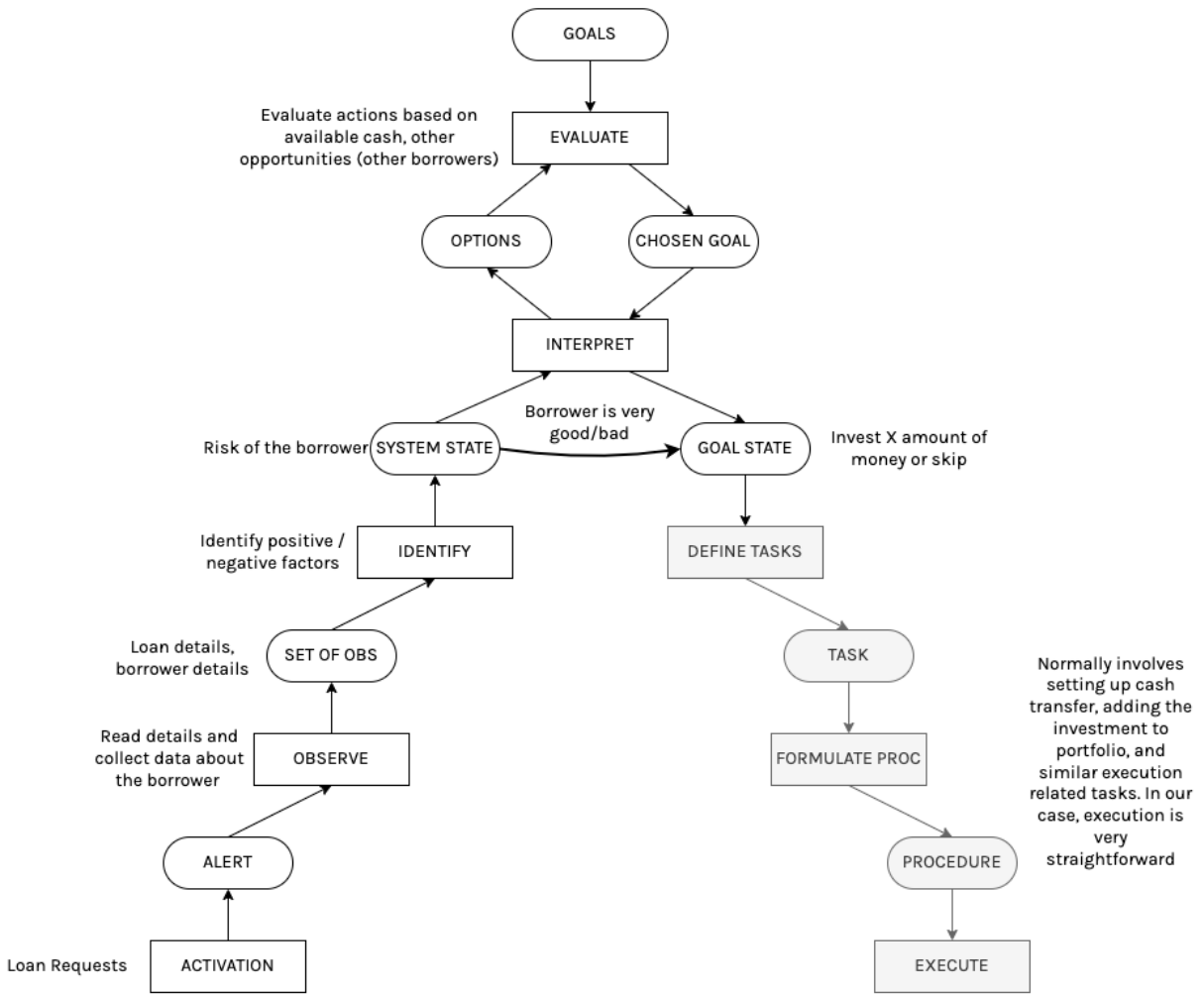


Figure 5.1: Decision ladder depicting how investment decisions can be made on the P2P platform

Note that the process described above was a generic description of the workflow regardless of the presence of AI or domain knowledge on the user interface.

Rasmussen [136] observed that skilled operators can enter this process from any entry point, and don't necessarily follow the steps linearly. Instead, they take two types of shortcuts: As a result of an activity, they can reach a knowledge state that is further down in the sequence. For example, in a troubleshooting scenario, after observing the signs of the faults (OBSERVE), expert users can immediately gain an understanding of the problem (SYSTEM STATE). The second type of shortcut involves achieving a knowledge state from another knowledge state through association. For example, once the fault is identified (SYSTEM STATE), the associated actions (TASK) becomes immediately accessible to the user as they have gained significant experience in the past which leads to an associative leap from SYSTEM STATE to TASK. In our analysis, we also observed actions that could be considered as shortcuts. For example, a shortcut from SYSTEM STATE to GOAL STATE is depicted in Figure 5.1. This shortcut took place in two forms. First, if the borrower was identified as a very high risk borrower, the decision would be to not invest and move to another borrower. Second, if the borrower risk is very low, a full investment without much consideration could be made. We observed this behavior mostly in cases where the borrower had a very good outlook and were asking for a very low amount (e.g., \$4,000) compared to the cash available (\$125,000).

Two common shortcuts are depicted in Figure 5.2. One shortcut, shortcut **B**, involved a scenario where the AI assistant predicted a high default risk for a borrower. What constitutes "high" depended on the participant, but most participants who have seen that the AI's prediction was higher than the risk they wanted to tolerate (OBSERVE) did not consider the borrower further and did not make any further assessment (GOAL STATE). This suggests that most participants were not concerned with false negatives, i.e., the AI predicts high risk but the borrower did well. Instead, participants didn't want to take the chance and skipped the borrower in most cases. Participants also used this shortcut, regardless of the AI assistant's prediction, if they observed signs of risks based on their prior beliefs. For example, some participants were very strict about late payments, and they immediately decided to not invest if the borrower had late payments in the past two years.

Another shortcut that we observed included premature conclusions when the AI predicted a low risk (Figure 5.2). In this scenario, the AI predicted a low default risk (according to the participant's definition of low risk), and this led to a premature conclusion that the borrower is safe. At this point, the only decision that needed to be made is how much money to allocate, and whether to invest now or later. This pattern was observed less than the previous shortcut, however, we noticed that a low default risk percentage was a key

P2P Lending with AI

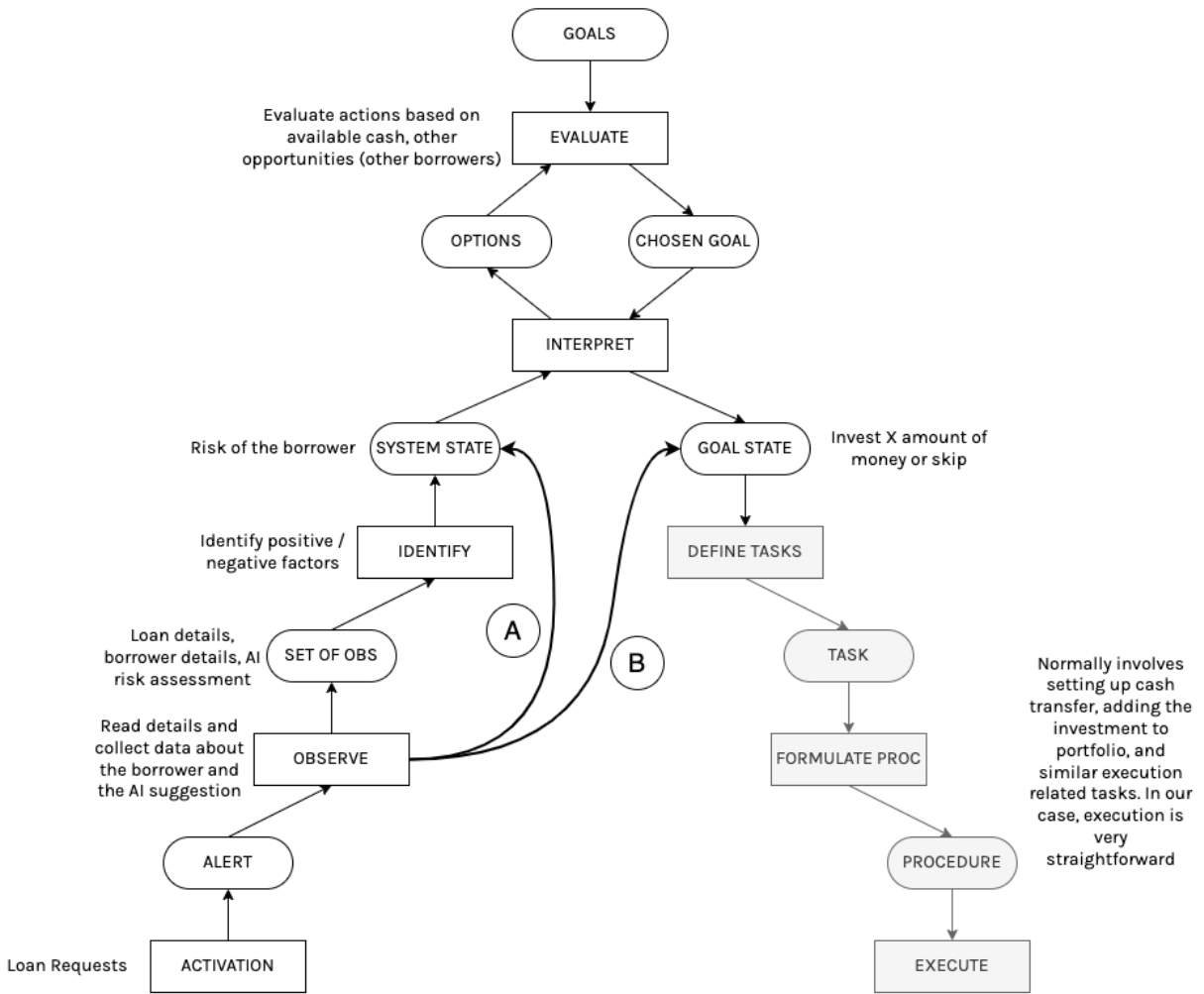


Figure 5.2: Decision ladder depicting how investment decisions with the AI assistant can be made on the P2P platform. In this case, two common shortcuts taken by the participants were shown. The shortcut **A** refers to premature conclusions (mostly low-risk assessments) based on the signs provided by the AI. The shortcut **B** refers to making a decision when the AI assistant outputted a high default risk percentage



factor in achieving a fast conclusion. Most participants, when asked about their preferred threshold, provided three ranges: a low percentage below which they would consider very safe, a high percentage above which they would consider very risky, and the range between the two where they would analyze the case further. For most participants, default risk percentages below the “low” threshold led to quick decisions about the borrower’s creditworthiness. This shortcut was also used in situations where the participant was focusing only on one or two parameters (based on their prior knowledge), and upon seeing that these factors were within their risk threshold, they immediately started thinking about how much to invest.

In both shortcuts, we observed that what constitutes a “low” or “high” risk depended on the participant. The percentage that was perceived as “high” or “low” risk differed among participants, ranging from 20% to 70%. Below are some of the comments illustrating the differences between participants preferred risk thresholds:

- *“When I clicked [on the loan] I scrolled down to the percentage that the AI would give me. And then if it’s about 50, I won’t do it. If it’s under 50, I’ll look closer. [P50]”*
- *“I think when it said like 40 I was still OK but then when it went to like 70s I was like maybe not.[P48]”*
- *“So I first looked at the percent of default risk, and just in terms of my comfort level, if it was like over 45, I didn’t really go for it.” [P52]*

## 5.4 Modeling Strategies using Strategies Analysis

While control task analysis is mainly concerned with “what” needs to be done, strategies analysis is mostly concerned with “how” it can be done. The primary motivation to conduct a strategies analysis is to identify effective processes that can be used to achieve the goal, depending on the changing situation [163]. For example, in the control task analysis phase, we outlined the fact that a creditworthiness evaluation has to be done after observing and processing the financial profile of the borrower and the loan conditions (i.e., amount, term, interest rate). However, there are multiple ways in which this evaluation can be achieved, and this is our goal with the strategies analysis in this section.

We used Vicente’s approach to model the strategies [163]. Vicente discussed that certain procedures (strategies) can be more economical (i.e., require less cognitive resources) than others. The goal of identifying these strategies is to design for them by including means

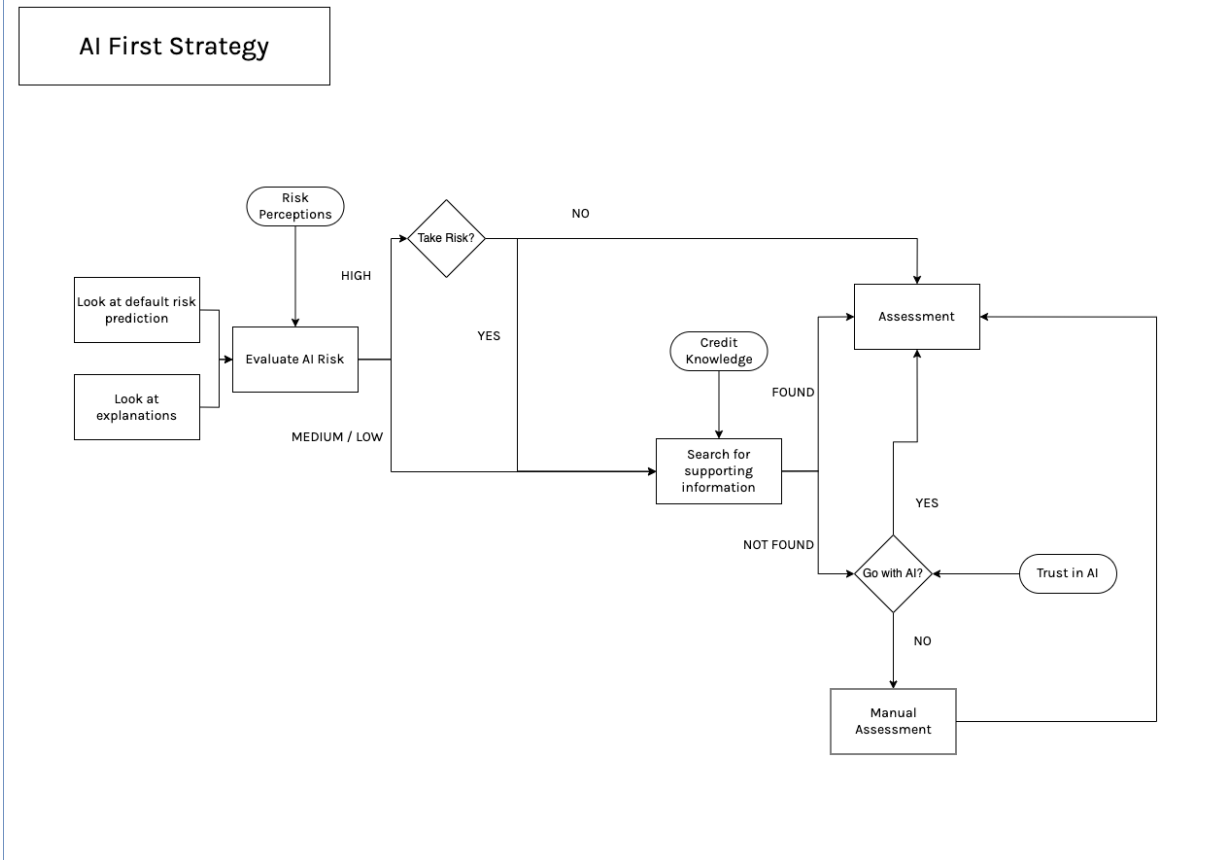


Figure 5.3: The AI-first strategy employed by participants.

to effectively use these strategies such as providing adequate information support. Vicente also emphasized the importance of defining strategies as categories as opposed to instances of those categories, and we adopted this approach. To identify the strategies, we relied on the interviews that were conducted as part of the experimental study presented in Chapters 3 and 4. We used information flow maps [163] to describe the strategies. Note that our approach was descriptive, however formative approaches to SA have also been proposed in the past [75].

In the SA, we focused on the left leg of the decision ladder that was described in the previous section. Specifically, we analyzed how participants went from OBSERVE to SYSTEM STATE, namely the risk assessment phase. We used information flow maps as described by Vicente [163]. The SA revealed a number of strategies that are used by the participants to assess the risk of the borrower.

During the interviews, we selected a few loans and asked participants to walk us through their decision-making process and explain how they made their risk assessments. Below are a few examples that show the diversity of the strategies mentioned during the interviews:

- *“First, I do my own check with the variables I just told you and then I go and check the graph and then look back and forth a bit.” (P13)*
- *“I’ll start with looking on my own and making my decision then going to the assistant for more help to see if I should go with it or not.” (P45)*
- *“I didn’t really pay as much attention to the information about their credit that was shown. I would scroll down to the chart that was provided by the machine learning assistant, which helped me evaluate the risks. And honestly, I did rely on that quite a bit.” [P38]*
- *“The machine learning assistant is the thing I would look at at first just to give me a general overview of what the computer thought it was a good idea or not. And then I had read over all the little question marks, so then I would go through each one and make sure they fall within low risk.” [P52]*

The interviews and the experimental data revealed three strategies. One strategy we observed was the *AI-first* strategy. In this strategy (Figure 5.3), participants started by looking at the AI assistant and the explanations and made sense of the evaluation. Participants’ preferred risk thresholds and risk perceptions influenced what they made out of the AI. If the AI assistant suggested that the risk is high, participants would consider

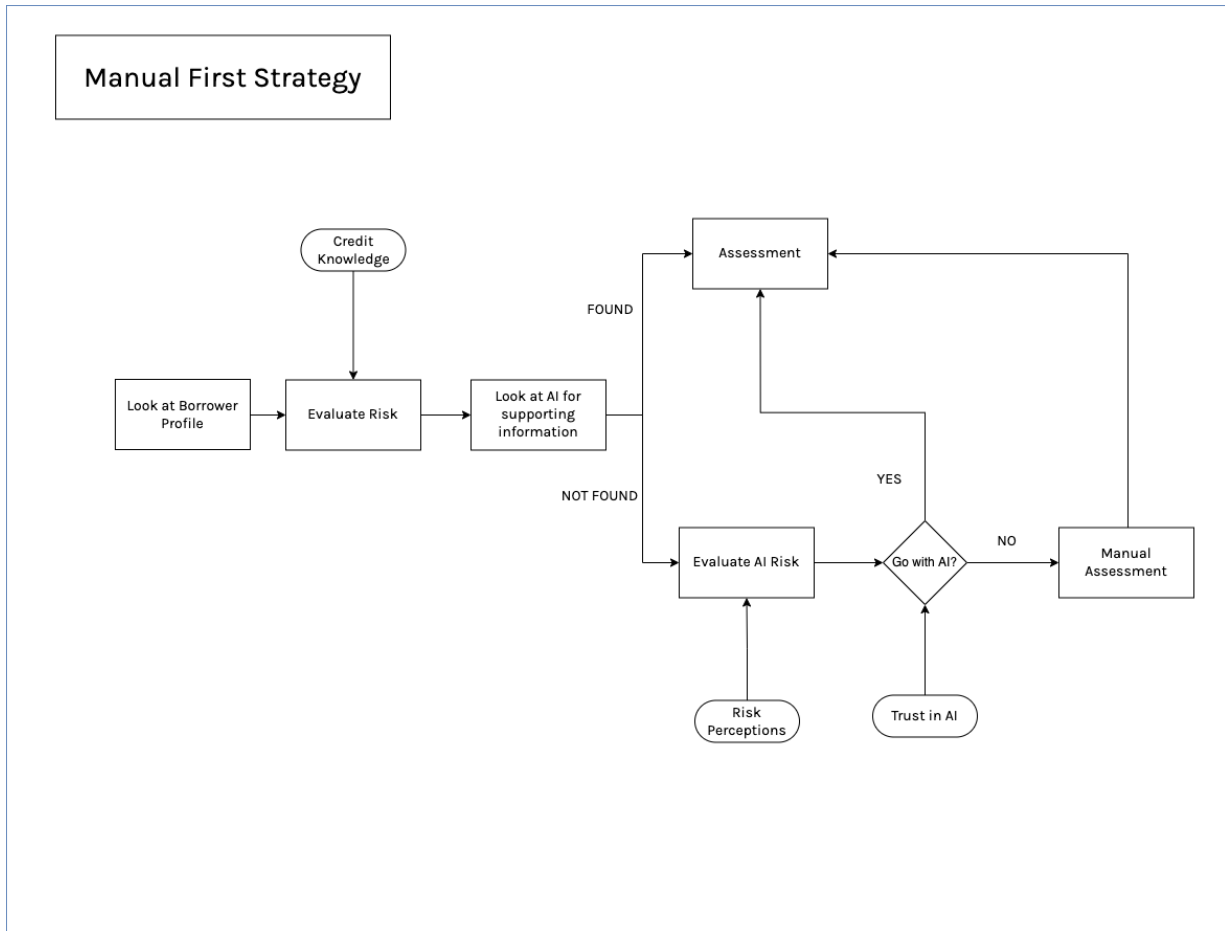


Figure 5.4: Manual-first strategy employed by participants.

if they are willing to take the risk. If they didn't want to take the risk, they finalized the assessment (i.e., skip the borrower). If they were willing to take the risk or if the AI suggested a low to medium risk, they would look at the borrower's profile, and search for information that supports the AI assistant's assessment. Finding support led to the assessment. If they couldn't find supporting information, they had to make a decision about whether to follow the AI or not. This decision was influenced by how trustworthy participants thought the AI was. Some participants indicated that in such cases, they would think that the AI has much more knowledge than themselves, and would trust the AI. In other cases, participants indicated that the factors AI considered were not reasonable and opted to ignore the AI and did their own assessment.

Another common strategy was *Manual-first* where participants would start with a manual assessment of the borrower by reviewing the loan-related and financial background-related factors, forming a preliminary opinion, then checking with the AI to see if they match. As shown in Figure 5.4, the strategy starts with looking at the borrower profile and evaluating the risk. Participants' prior beliefs and priorities affected this evaluation process. For example, if they believed that income and debt are important but late payments are not important, they would form their opinions based on income and debt alone. Next, they would check the AI assistant to see if it supports their evaluation. If they could find support, this led to a quick assessment. If not, they would evaluate the AI assistant's assessment. Similar to the *AI-first* strategy, participants' risk perceptions and preferred risk thresholds played a role in this process. After checking the AI assistant, if they were convinced, they relied on the AI instead of their own assessment. If not, the only option is to invest based on their own assessment. Similar to the *AI-first* strategy, trust in the AI assistant played a role here. If participants believed that the AI is not trustworthy in this case, they ignored it. One of the common signs that led to distrust was the mismatch between what participants thought is important and what the AI thought is important. At this point, participants had to reconcile the two different evaluations. As a result, sometimes they either changed their initial opinions and trust the AI assistant, or ignored it (or put less weight into AI), and mostly relied on their own assessment. The following comment is an example of this thought process:

*"I looked at why it [AI] made its decisions, so I looked at the weights and the explanations from the model to what they were putting a lot of weight into and if I hadn't thought about that then I'd consider changing my evaluation. Otherwise, if I think that the explanations by the model seem not convincing, then I would go with my own instinct."* [P47].

A unique strategy that was used by some participants was leveraging the AI and the explanation interface to guide the manual decision-making process (*AI-guided* strategy; Figure 5.5). This strategy was similar to the *Manual-first* strategy, however, before the manual assessment, participants checked the AI explanations to get an idea of what they should be looking for and built a mental model (Model of Risk). This model is then used to guide the manual assessment process. Participants who adopted this strategy often indicated that the default risk percentage was not informative enough, and they wanted to understand the factors that are important for a given borrower profile, so that they could focus on those factors to make their assessment.

Finally, Figure 5.6 shows how embedding domain knowledge in the AI interface influenced the *AI-first* strategy according to participants in the *embedded* condition. Here, before or during evaluating the AI assistant's assessment, participants could see a warning

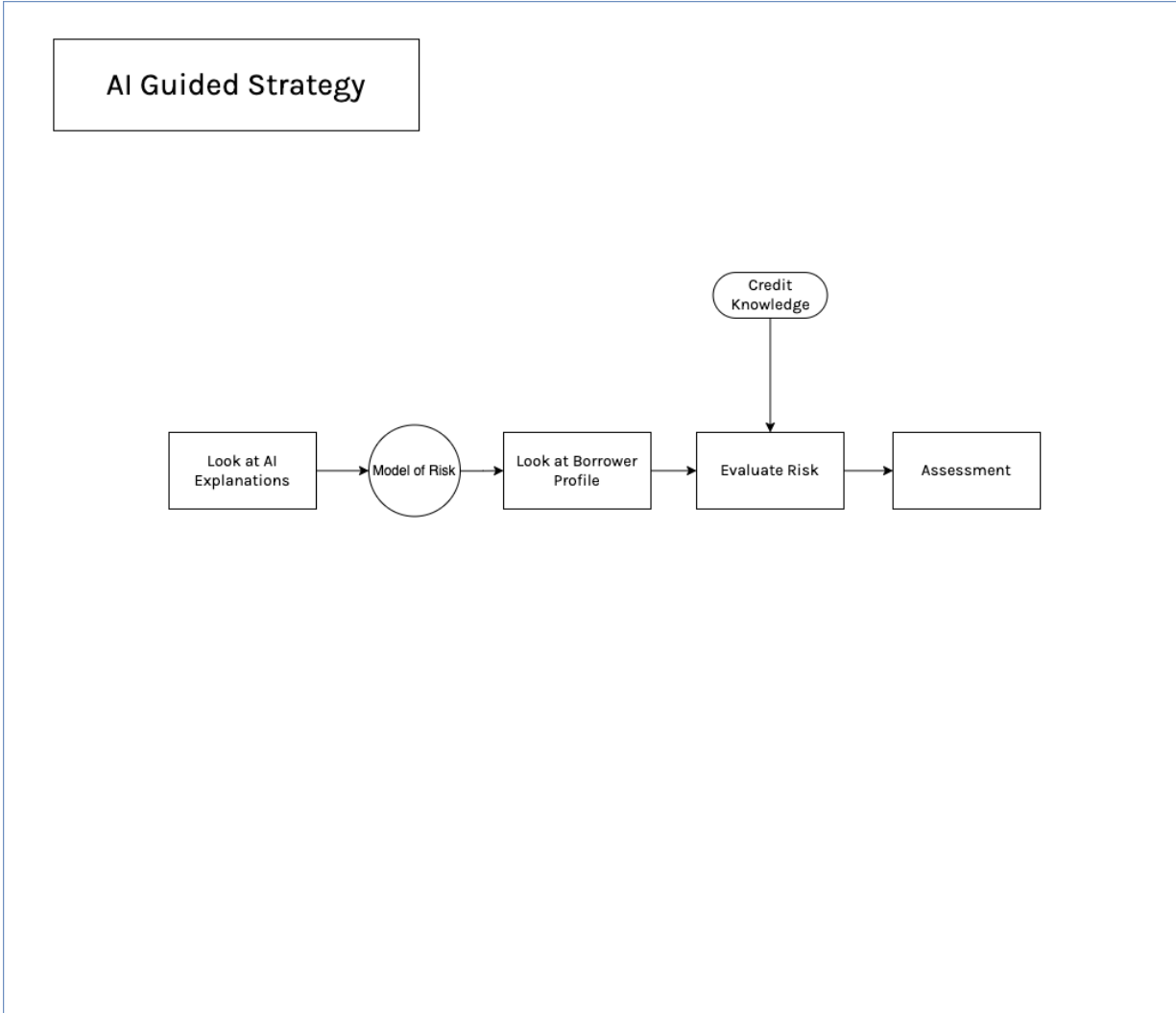


Figure 5.5: The AI-guided strategy employed by participants.

and relevant domain knowledge explanations about the borrower’s situation. If the warning and the explanation made sense to them, this would alter their perceptions of the AI assistant’s trustworthiness, leading to ignoring the AI and relying on manual assessment. In a way, domain knowledge and associated warnings could act as a sanity check, add friction, and delay the decision to rely on the AI assistant immediately.

This was evident in some comments: “*Sometimes the machine learning had the red pop-up where it’s like this prediction may not be accurate, so those ones I looked at more in depth into. [P48]*”. The presence of such information led to suspicion of the AI: “*I saw that several times the machine learning model said a low prediction or low risk, but there were tips [embedded domain knowledge] saying it was higher risk. Yeah, and during that point I would have lower faith, and I guess it’s more useful to follow the tips.*”. It appears that the presence of the warnings led to a decision point about whether to rely on the AI or not, as expressed by one participant: “*There were the alerts that said that this affects the default risk [embedded domain knowledge]. That’s when I would probably go against what the machine learning assistant advised and went through with the data on my own and accordingly made my own decision.*” [P57]. Moreover, the the warnings indeed seem to have affected participants’ trust, as shown in the orange “YES” path in Figure 5.6 and in this comment: “*I wouldn’t follow its [AI] advice directly, because sometimes, when you see those red boxes, it kind of deteriorates your trust in the assistant.*” [P49].

Note that taking the warnings seriously was still influenced by prior knowledge and beliefs. For example, some participants, after seeing warnings about the number of inquiries, indicated that they would not take it seriously as they didn’t believe the number of inquiries was a big deal.

The strategies participants used were not static, and from the interviews, our understanding is that participants frequently switched between strategies, depending on the context. For example, participants could start with the *Manual-first* strategy, however, if the borrower’s financial profile was challenging (e.g., involving conflicting signals), then they might switch to the *AI-first* strategy. Similarly, if the AI assistant’s evaluation did not make sense in the *AI-first* strategy (e.g., highlighting factors that don’t seem to be relevant), participants could switch to a *Manual-first* strategy. The *AI-guided* strategy was either adopted completely or was used in conjunction with other strategies. For example, in the *Manual-first* strategy, examination of AI and explanations sometimes led to discovering factors that participants did not consider before. In such cases, participants would go back to the borrower’s profile and include those factors in their assessment as well, e.g., If a participant did not consider account age-related factors, upon seeing that the AI listed one of the account age factors as the top contributor, the participant would go back and re-evaluate the borrower based on the new information.

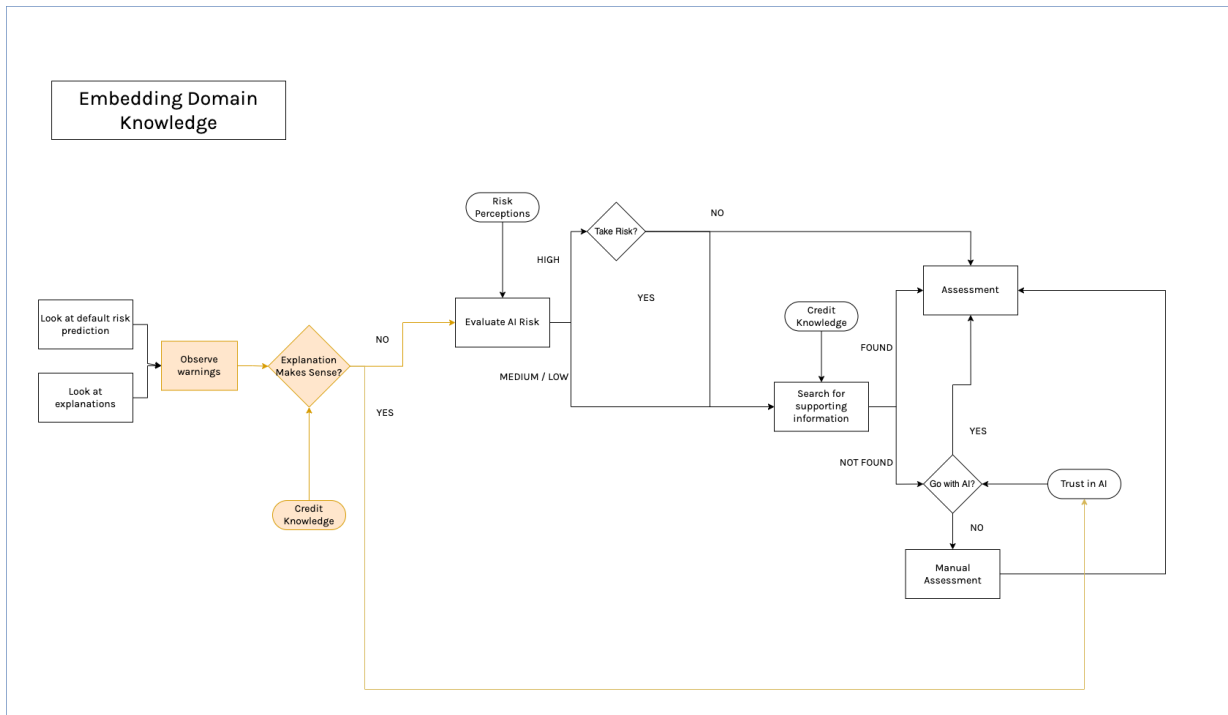


Figure 5.6: Information flow map showing the effect of embedding domain knowledge in AI interface on changes in the AI-first strategy.



One of the common elements in the *AI-first* and *Manual-first* strategy was the decision about relying on AI after evaluating its output. As illustrated in Figures 5.3 and 5.4, trust in the capabilities of AI was an influential factor. The interviews revealed that uncertainty about the capabilities of the AI assistant played a role: *“I tried to use the machine learning, but then at the back of my head I was just like oh well you never know what’s going to happen. It’s just it could be wrong. I don’t know. (P23)”*. Part of the uncertainty come from the lack of participant’s own knowledge: *“I’d think that this [borrower] would have a lower chance of defaulting, but then I would scroll down and see the percentage was relatively high. In that case, I think that I would trust the machine learning system more than my own intuition. Since again, I don’t have that much experience with something like this.” [P7]*.

## 5.5 Discussion

### 5.5.1 Control Task Analysis

We applied control task analysis to understand participants’ workflow in the lending task, and to describe the processes that we observed during data collection. The decision ladders were built to reflect the flow and to identify the shortcuts that were used by the participants. Strategies analysis was conducted to identify how participants approached the decision-making task with AI.

We found that participants took several shortcuts during the task. According to Rasmussen [136], the shortcuts on a decision ladder represent the expert behavior, and the motivation for building a complete decision ladder is to present the steps the novice users need to take. However, contrary to this notion, we found that non-expert participants in our study took shortcuts (e.g., skipping a borrower if the AI output a default risk percentage higher than the participant’s threshold). A difference between the shortcuts that are discussed in a traditional decision ladder and our context is that the shortcuts our participants took can be described as naive thinking as opposed to expert thinking. Note that this does not necessarily mean that the shortcuts yielded negative results. On the contrary, some of the shortcuts, e.g. skipping a borrower if the risk percentage is high and possibly making the trade-off about the false positives can be considered a decent solution. However, these shortcuts were not a result of domain expertise in the traditional decision ladder sense. These shortcuts, in fact, represented a naive way of thinking about the task, relying on past experiences that may or may not be related to task experience in this domain, or due to poor understanding of the situation. This difference partly comes

from the fact that traditionally, decision ladders are constructed in a formative sense, i.e., to identify what can be done, instead of describing what is being done by observing how experts work and gaining insights. In our case, insights gathered from non-expert users could only reveal what was done and didn't reveal the processes that experts might exhibit.

As mentioned before, some of these shortcuts depended on participants' risk perceptions. It appears that probabilistic outputs are perceived differently by different individuals, and this has some implications for future research. First, determining thresholds becomes problematic in any research design. For example, creating conditions where the AI has high vs. low confidence in its prediction (presented with probabilities) would not create the same effect for each participant, and it might be difficult to draw conclusions based on such an experimental setup. Consider this statement from one participant: *"I did not invest very much and most of the credit for that goes to the machine learning assistant because it seemed like in a lot of the cases that I would have invested in normally the machine learning assistant said, oh, this has a 44% default risk and I said, no, I'm not going there."* [P58]. While this statement shows how powerful an AI system can be in shaping users' decisions, it also highlights the challenges associated with designing highly controlled studies.

One solution to address this issue might be using class labels instead of percentages (e.g., risky vs. not-risky), however, this may not always be practical. Another approach might be employing some form of calibration, similar to eye-tracking studies. In any case, we argue that it is important to identify issues pertaining to the perception of probabilities in future studies.

In a traditional ConTA, identifying the shortcuts suggest that the design of technology should enable and/or support this level of expert thinking. However, our findings with non-experts suggest that the design implications can be more complicated. For one, certain shortcuts, e.g., jumping from a state of knowledge or action to another state of knowledge (jumping to conclusions) may not always be something we want the users to engage in. In that case, perhaps the design should prevent or make it difficult to adopt these shortcuts. Ultimately, the shortcuts provide a cognitive advantage wherein the users don't have to engage in higher-level, mostly complex forms of thinking. However, automation complacency and other negative effects of introducing AI into the decision-making can inadvertently create shortcuts that may reduce cognitive complexity yet lead to more errors (e.g., seeing a low default risk percentage and prematurely concluding that the borrower is safe). Therefore, it appears that for non-experts, identifying what kind of heuristics and shortcuts people use is very valuable, and can lead to important design decisions to prevent such naive shortcuts. In fact, the interventions that were explored in this work (i.e., domain knowledge) can be thought of as design practice that aim to prevent this level

of thinking and force participants to reason about the problem using domain knowledge, rather than leaving the decision-making to the AI system.

### 5.5.2 Modeling AI on Decision Ladders

While we evaluated ways to model the AI in the decision ladders, ultimately, the best solution was to not model it explicitly. The primary reason was that there was no role allocation between humans and AI. The AI and the human did the exact thing: Look at the status of the world (borrower profiles), and make a risk assessment. In this sense, the AI was not very different than having a financial analyst team member who may share their opinion about an investment opportunity or act as an advisor. The user ultimately is in charge of making a decision (stepping through the decision ladder), but has access to the AI assistant’s interpretation of the world and can take advice from their AI teammate. The way our participants described their workflow suggested that the AI was another piece of information content they checked to make risk assessments.

While the control task analysis helped us identify the processes that were relevant in the lending context and describe the shortcuts participants have used, the presence of AI in the model was limited. As stated before, in our context, describing the AI assistant as another information source that can be utilized to make the risk assessments, led to dropping it from the decision ladders. The best example of modeling automated agents on a decision ladder comes from [104]. Li et al. proposed a Degree of Automation (DOA) layering approach to understand how different degrees of automation affect human work, and presented a comparison of low vs. high DOA that can be conceptualized on a decision ladder. They identified four main parts of the decision ladder corresponding to four stages of automation: Information Acquisition (ACTIVATION and OBSERVE), Information Analysis (IDENTIFY), Decision Selection (INTERPRET and EVALUATE), Action Implementation (TASK, PROCEDURE, and EXECUTION). According to this conceptualization, the AI used in this work is considered an Information Analysis automation and can be represented in the IDENTIFY part of the decision ladder. However, in our analysis, we didn’t find any advantage of using this approach. It appears that the level and stages of the automated system of interest are key here. Li et al. [104] suggested that their approach is well suited for modeling automated systems that have (a) variable DOA, and (b) works at multiple stages. Our AI had a fixed degree of automation and was responsible for only one stage, therefore we were not able to take advantage of the DOA layering approach. However, we believe that such an approach will be useful if the AI is expected to handle certain functions on its own in the future, e.g., automatically allocate funds based on the predicted risk of a borrower.

### 5.5.3 Strategies Analysis

While ConTA helped us better understand the effect of AI in this context, the type of decision-making scenario we used in our experimental study seems to be better represented in a strategies analysis. Conceptualizing the AI as an information source had some advantages for the purpose of strategies analysis. From the strategies analysis, it is very clear that the presence of AI did have an effect on the strategies participants employed (compared to a situation where there is no AI). The presence of AI increased the variety of strategies, as well as created new categories of strategies (for example, AI-guided manual assessment).

Note that the SA we presented in this chapter does not have the same complexity as in typical complex systems. The system of interest did not have physical components and complex physical processes that are valuable to model (e.g., changing modes and settings of the system, monitoring a dynamic response from the system to these changes) as our interface was fairly simple to use. Still, we were able to identify a few different ways of evaluating the creditworthiness of the borrower. In a more complex (and dynamic) system such as real-time investment (e.g. stock market trading) the strategies would involve much richer procedures (cognitive and physical). Nevertheless, the strategies identified here have implications for designing user interfaces for AI systems.

As pointed by [163], the goal of strategies analysis is to inform design that allows users to easily transition between different strategies. In our context, this might look in the form of filters (e.g., filtering out borrowers with “high risk” percentages) or better UI components that allow seamless transition between AI-based and manual assessment (e.g. displaying side-by-side comparison of various information on the interface). However, in this analysis, some of the strategies employed by our participants were questionable, and should perhaps be not supported. For example, the *AI-first* strategy, while cognitively economic, can yield biased decisions such as anchoring [126, 172] and confirmation bias [127, 36] as it can result in searching for supporting information rather than making an assessment free from the influence of AI. In this case, for example, a design implication for those who employ this strategy might be providing a number of different metrics such as confidence ratings or otherwise force users to engage in deliberate thinking.

The need for better evaluation tools was mentioned by some participants. For example, one participant said that showing the historical performance of the AI might help: *“I would like to have some sort of an indicator of how accurate this AI might be. Like in similar decisions has the AI been correct or not.”* [P40]. Other participants wanted to understand what an “average” applicant looked like. Since we converted the contribution of each factor from arbitrary numbers (provided by the XAI technique) to human-understandable

percentages (e.g., Credit card utilization of 27.9% decreased the default risk by 3.6% compared to an average applicant), it was not obvious what an average applicant is according to the AI.

However, we should note that most participants felt overwhelmed with the number of information presented on the XAI interface, as described by one participant: *“Yeah, so ideally it would just tell me like yes or no and I wouldn’t have to think anymore, but because it gave you the information [explanations] I’m like, OK, I should probably look at this so I don’t know how much it helped ... It just felt like, yeah, you’re helping me but you’re also giving me this information so I feel compelled to look at it.”*. From a usability perspective, limiting the visible information to the essentials, while giving the users the ability to access more information is a reasonable suggestion, for example using a progressive disclosure approach [156]. While it is tempting to present “data” to back up a prediction, we observed that some participants were quite confused with how to interpret that data, especially when it comes to numbers and percentages. Instead of providing all the explanation facilities that are available, a design for non-experts should include a few key pieces of information that are easy to digest. For example, instead of using excessive charts and visualizations, the explanations should focus more on information that is easier to explain and understand. At the same time, including more contextual elements can be beneficial for some users. These could include a historical context (historical performance), performance in a comparative context (i.e., how the AI performs in other similar cases). Perhaps an interesting opportunity for future research is using strategies analysis to guide an iterative design process. For example, design choices can be evaluated and prioritized based on how well they address the brittleness identified in strategies analysis, and the design can be improved in an iterative fashion.

Not all strategies were questionable. In fact, it appears that relying on AI when the default risk is high may be a convenient and cognitively economic strategy. In this case, providing appropriate support tools to help users quickly disregard those cases might be appropriate. For the **AI-guided** strategy, it looks like explainable AI can be used as a tool to help users understand the decision context better, in addition to helping them to understand and evaluate the AI, however if the explanations draw an incomplete picture of the situation, this might lead to incorrect assumptions. Another aspect of the problem comes from the fact that the models of the world that feed certain actions (ellipses in the Figures 5.3 and 5.4) are highly fragile and variable. In the original SA [163], Vicente discusses that their SA presents an idealized process, which suggests that the models of the world (e.g., knowledge about how an equipment functions) that affect or are affected by actions are implicitly accurate. In our case, inaccurate models can result in breakdown and incorrect actions, such as unnecessary strategy switching. Therefore, we argue that

in our case, the design implications should support the users with the cognitively complex parts of the strategy, but also prevent them acting with inaccurate representation of the situation.

The findings on how participants evaluated the machine learning explanations illustrate this problem, support the idea that no matter how good the explanations are, they will still be evaluated through a domain expertise lens, and in the case of non-expert users, this evaluation will likely involve naive assumptions and limited understanding of the task domain. For example, one participant said “*Whenever there is a mismatch, I would understand why the machine learning algorithm is considering something is very risky ... If I feel it can be ignored then I would just not go with the machine learning algorithm.*” [P54].

Most participants indicated that if they don’t know or care about a particular factor (in most cases, the number of past credit inquiries was one of them), they would not consider that as an important factor to take into account. For example, if the AI assistant provided a high default risk percentage, but one of the most important factors was something the participant didn’t think is important, they would undermine the AI assistant’s prediction and treat it as lower or higher than it actually is. Previous work showed that providing explanations can lead to overtrust and overreliance [96, 23, 160]. Our findings reveal another perspective: Explanations can lead to undermining the capability of the AI system if explanations are perceived as incorrect or irrelevant. Of course, these perceptions are heavily influenced by the prior knowledge and experience of the user. In the case of non-expert users, this can lead to under-reliance on AI. As we argued and explored in this work, equipping users with sufficient domain knowledge might be key to addressing these issues and make the strategies more resilient.

Vicente [163] also discussed the reasons for having multiple strategies and concluded that different strategies are adopted usually due to the changing circumstances and changes in the mental workload or mental demands of the task. From the interviews, we also observed that participants tended to switch to a more “cognitively economic” strategy when the case was not clear (e.g., the risk is not clearly high or clearly low), or the circumstances were different (e.g., a borrower asking for \$6,000 might have resulted in a different strategy than a borrower who is asking for \$32,000). Another form of strategy switching was observed during the assessment process. For example, In the *Manual First* strategy, after reviewing the borrower’s profile and making a preliminary assessment, participants could look at the AI assistant’s risk prediction and explanations, realizing that they haven’t considered a factor that was important for the AI, and switch to the *AI-guided* strategy. In other instances, we observed that participants took a mixed approach involving both *AI-first* and *Manual-first* strategies. For example, participants could evaluate the risk manually by looking at the factors that are important to them and looking at the AI

assistant for supporting information (*Manual-first*). For the remaining factors, they may follow an *AI-first* strategy, where they would look at how the AI assistant assessed risk and check manually to confirm the AI assistant’s predictions. These examples show that task switching was taking place frequently, and was influenced by both contextual factors and prior knowledge and experience.

Finally, explanations can be a double-edged sword. They can lead to more positive experiences and more trust in AI, however they can also result in undermining the AI if there is a mismatch between the explanations and the user’s understanding of the situation. One approach to address this issue is focusing on communicating the limitations and potential risks over justifications. As we discussed before, assessing how well the explanations justify the AI behavior requires a level of domain expertise, and in the case of non-expert users, reading the explanations without sufficient domain expertise can have unintended consequences.

One of the interventions we explored in this work to address this issue (embedding domain knowledge in AI) resulted in modified strategies (Figure 5.6). When there was a warning, participants processed the warning before evaluating in the *AI-first* strategy. If the justification for these warnings (domain experts’ opinions) made sense to them, their trust in the AI assistant’s prediction would decrease for the current case, and they may switch to manual assessment. Compared to the absence of such information, embedding domain knowledge added friction to the decision-making process and increased the likelihood of deliberate thinking. It made the strategy slightly more complex (added additional actions and decisions), however resulted in better utilization of the strategy. It appears that this approach might be useful in improving existing designs. However, it also illustrates the complexity of designing for strategies in our context with non-expert users.

Overall, we argue that SA has been extremely useful in gaining a better understanding of the P2P context, and we believe that it is very suitable for the type of decision-making scenarios that we examined. We were able to identify how AI affected the approaches taken, and potential issues that may decrease the effectiveness of strategies. While we also identified opportunities for design, it appears that translating these insights into design is not straightforward, and requires further analysis. We believe that a WCA, especially the SRK framework, can help close this gap by identifying the requirements associated with cognitive support.

### 5.5.4 Future Work

While we opted to not model the AI assistant on the decision ladder, we are interested to see how the AI will be conceptualized in future studies. In safety-critical systems such as healthcare or defence, machine learning-based AI systems are not expected to replace the human role in the foreseeable future, and modeling the function the AI plays and describing it in CWA models will likely require more attempts to better understand how to model the AI that is most beneficial.

Going forward, we believe that a CWA approach to human-AI interaction can help understand and identify opportunities for future research and design. Each system has unique features, and capturing these across different domains and systems has the potential to reveal common characteristics that may start the foundation for developing guidelines and principles.

Finally, our analysis was limited to non-expert users. We believe that there is value in conducting a similar analysis with expert users and identifying the differences between non-expert and expert approaches. Perhaps this can help close the gap and answer the questions such as “should design support or prevent this shortcut or strategy” more accurately.

## 5.6 Contributions and Conclusion

In this chapter, we presented a CWA-oriented discussion based on the findings obtained from previous studies. Our main contribution was providing recommendations on how CWA can be extended to adapt to the AI/XAI context. We laid out several ways of conceptualizing AI systems using the existing CWA tools. We also identified opportunities to apply CWA in future AI context and made the case for future use of CWA in the AI/XAI context.

We believe that the main benefit of conducting ConTA and SA was to identify how AI can influence the workflow and strategies people may adopt in a decision-making situation like this. We found that especially SA has been very useful in understanding how AI can shape strategic decisions participants made. In this work, we demonstrated the usefulness of SA in the context of AI, and we recommend future AI work to consider conducting SA in addition to more typical stages such as WDA and ConTA.



# Chapter 6

## General Discussion and Conclusion

### 6.1 Introduction

In this chapter, we summarize the findings, and present a broader discussion about the implications of this work, state our contributions, acknowledge the limitations of the work, and discuss future work.

### 6.2 Summary of Key Findings

Below is a summary of key findings from the experimental studies and additional analyses discussed in this work.

- In the study presented in Chapter 2, we found that augmenting explanations about an AI system's predictions had an observable, but small effect on perceptions of explanations, however, the behavioral intentions for subsequent actions were not affected. This study also demonstrated how AH can be used to provide reasoning mechanisms to make sense of XAI. Finally, adding domain knowledge helped participants to gain a better understanding of the lending domain.
- In the study presented in Chapter 3, we found that providing domain knowledge along with XAI led to better task performance, and this was primarily caused by participants avoiding the situations where the AI was incorrect. Perceptions of explanations were not affected by the presence of domain knowledge. Adding domain knowledge also resulted in lower levels of trust in AI.

- In the follow-up study presented in Chapter 4, we found that embedding domain knowledge into the XAI user interface and communicating information about possible AI failures improved task performance further, and resulted in higher profits. The results suggested that participants were much more sensitive to the discrepancies between domain knowledge and AI advice when these information are presented together.
- In Chapter 5, we synthesized our findings using CWA and presented control task analysis and strategies analysis to describe our context. We identified the ways in which the presence of AI influenced strategies used in the decision-making task. We also identified the shortcuts participants have used and discussed the design implications. This chapter demonstrated the usefulness of CWA in understanding how non-expert users interact with AI systems.

## 6.3 Research Questions

In this work, we posed two broad research questions, and in this section, we discuss how our findings address these.

- What is the role and importance of domain knowledge in human-AI interaction?
  - In what ways can domain knowledge be used to support users to make better decisions when working with an AI system?
  - How does domain knowledge affect perceptions and use of an AI system?
- How can we leverage Cognitive Work Analysis in understanding and designing AI systems?
  - How can CWA be used to gain a deeper understanding of a human-AI interaction?
  - How can CWA be used to design better AI systems?

### 6.3.1 What is the role and importance of domain knowledge in human-AI interaction?

Our findings showed that domain knowledge helps users make sense of XAI explanations (Chapter 2) and identify situations where AI makes a mistake (Chapters 3 and 4) and

integrate AI advice better into decision-making process, as demonstrated by increased task performance metrics.

### **In what ways can domain knowledge be used to support users to make better decisions when working with an AI system?**

We demonstrated a number of ways in which domain knowledge can be leveraged to support users. Using AH as a knowledge representation framework [19], we showed that AI explanations can be improved by communicating the relevance of the factors that AI considered in its decision (Chapter 2).

In Chapter 3, we showed that adding domain knowledge to XAI (in a complementary display) increases task performance, primarily due to ignoring AI advice when it is incorrect and reduced the likelihood of overtrust in AI. In Chapter 4, we showed that embedding domain knowledge into XAI (in the form of warnings) increased task performance even further, and allowed participants to avoid erroneous AI advice.

### **How does domain knowledge affect perceptions and use of an AI system?**

We found some evidence that adding domain knowledge to XAI can lead to more positive perceptions of AI (Chapter 2), however failed to demonstrate it in the experimental studies (Chapters 3 and 4). As noted in the corresponding chapters, evaluating subjective opinions towards XAI is challenging [72], and our aggregate approach (measuring user perceptions at the end of the study instead of a case by case basis) limited our ability to draw conclusions further. However, trust in AI (positive trust) was affected by the presence of domain knowledge (Chapter 3) which suggests that expert knowledge, if integrated into XAI, can be useful to help users calibrate their trust.

As demonstrated in Chapters 3 and 4, adding domain knowledge led to less reliance on XAI in high risk cases and more questioning the AI advice when it was incorrect. In our analysis in Chapter 5, we showed that lack of domain knowledge resulted in non-optimal, mostly naive strategies when working with the XAI.

### **6.3.2 How can we leverage Cognitive Work Analysis in understanding and designing AI systems?**

Our findings showed that CWA can be used to identify opportunities to improve data-driven XAI explanations (Chapter 2), design decision-support tools that can be used to

assess XAI (Chapter 3), and gain a deeper understanding of how the system is being used (Chapter 5).

### **How can CWA be used to gain a deeper understanding of a human-AI interaction?**

In Chapter 5, we showed that ConTA and SA are particularly useful to understand how users use XAI. Used in a descriptive fashion, this analysis revealed how non-expert users approach decision-making with AI, and how presence of explanations lead to naive assumptions and strategies. This led to a better understanding of how users worked with our XAI, and revealed opportunities to improve the design to prevent less effective ways of reasoning about and using the AI system.

### **How can CWA be used to design better AI systems?**

We explored how AH can be used to map domain knowledge to XAI explanations to help users make sense of explanations and predict future AI behavior (Chapter 2). In experimental studies in Chapters 3 and 4, we showed that expert knowledge, obtained by CWA, can be used to improve task performance. Finally, we identified naive shortcuts and strategies participants have used (Chapter 5) and make the case that future design of AI systems can use these to promote more effective strategies and eliminate strategies that can lead to errors.

## **6.4 Discussion and Implications**

In this work, we addressed the issue of the domain knowledge gap in a context where the users lack domain knowledge but dealing with a complex domain and supported by AI. We believe that as the AI becomes better at reducing the complexity of the existing complex fields and make them accessible to users with less domain experience, the issues surrounding the domain knowledge gap will become more important. In this work, we explored a number of different approaches to close this gap and help users of AI systems make better decisions.

### 6.4.1 Domain Knowledge and AI

We identified opportunities to contextualize AI explanations by integrating domain knowledge (Chapter 2) and using AH as a tool for reasoning about explanations and AI advice. This approach can be used to improve trustworthiness of future AI systems by helping users make sense of the predictions and explanations. Additionally, it can be used to convey causality in explanations, which can lead to more accurate interpretations of AI. This approach is also complementary to the recent work in knowledge-based XAI [100, 84, 84, 59] which aims to integrate domain knowledge tightly into AI systems by representing conceptual knowledge (in the form of knowledge graphs and ontologies) during the training process. As AH is a form of knowledge representation framework [19], there is potential to combine both approaches to produce AI systems that encompass expert domain knowledge and have the capability of providing explanations that support reasoning. In this work, we focused on the communication of explanations and found some evidence that it can be beneficial to users, however it appears that how and when this approach can be best utilized requires more research that investigates contextual and user-related factors.

When an AI system makes mistakes, it is up to the user to identify the situation, and rely on AI appropriately. However, as we (Chapters 3 and 4) and others have shown [116, 60, 91, 150, 129], lack of domain knowledge can make it difficult to make an accurate assessment of the AI advice, even if explanations are present. While this is not an AI-only problem, AI systems built without taking into account the domain knowledge gap will have limited success in augmenting human decision-making. In this work, we showed several ways of integrating domain knowledge into human-decision making in XAI, and demonstrated that this approach is potent in identifying AI failures and improving task performance. Future AI systems should consider integrating domain knowledge into the models and provide facilities to help users become better decision-makers.

XAI was born out of concerns regarding lack of tools to assess how well an AI system performs, and to develop appropriate levels of trust and reliance. We believe that this work contributes to these efforts by showing the benefits of adding domain knowledge into this process. One on hand, this work showed that XAI approaches that don't consider users' level of domain expertise will have limited capacity to support the users. On the other hand, XAI can be a great opportunity to tackle the domain knowledge gap as we demonstrated in this work. The goal of explanations of an AI system could be, in addition to providing justification for a model's behavior, to equip the users with domain knowledge that is needed to make sense of the AI output, as well as help them increase their expertise. Such an approach can reveal interesting opportunities for XAI in becoming a resource to learn from. Indeed, some studies have shown that XAI has potential to help users widen

their domain knowledge [37, 67] and to take role in teaching people [72]. In our case, such an approach can help create a financial investing AI system that not only helps users increase the profit they make, but also help them become better investors.

Regarding XAI challenges, Gunning et al. [72] described the challenge of starting with computers (i.e., focusing on explanation generation) vs. starting with humans (i.e., focusing on user needs and capabilities, [117]). Upon reflecting on our work, we believe a third approach, a domain analysis approach, can be valuable to address this challenge. As demonstrated in this work, understanding the task domain can help improve communication (Chapter 2), identify key information needed to work effectively with XAI (Chapter 3), improve task performance (Chapter 4), and understand pitfalls and reveal opportunities to support users of XAI (Chapter 5). This approach can add unique value that is more difficult to obtain by focusing on technology or users alone.

## 6.4.2 Cognitive Work Analysis and AI

This work also presented an exploration of CWA in the XAI context. In Chapter 2, we presented an approach of using CWA and AH to augment the explanations of the model. In Chapter 5, we conducted ConTA and SA to understand the findings from the studies. As we demonstrated, CWA was very useful in gaining a deeper understanding of the XAI context but also had significant design implications. We believe, that CWA, and in particular WDA, has a lot more to offer when it comes to building AI systems, and can have a significant impact on the entire AI life cycle. For example, a full AH about lending and credit could reveal how much coverage an existing AI system has. In an ideal situation, we would want AI to represent the real world fairly accurately, and be comprehensive. By building a full AH and comparing it to an existing AI system, factors that were left out could be identified and further analysis can be done on "how" the AI system should be built. In fact, this can be a particular advantage during the initial stages of a model development. For example, a comprehensive AH can identify "what data to collect" and help evaluate the existing models based on the domain analysis. Furthermore, AH can also contribute to model building and feature engineering stages. Important relationships and derived variables that are identified in AH can guide which features need to be included in the model, and how important they are. The importance of debt-to-income ratio is a good example of a derived variable that renders the debt and income less useful. A model built on debt or income alone therefore could be improved by exploring and utilizing debt-to-income ratio. While finance or credit is a well established field and has involved modeling ever since, in other, less well-known domains, we can expect significant advantages of such an approach. Other stages of CWA have significant

design implications. Once a model is built, the information design around the AI and XAI significantly impacts how users utilize them, as demonstrated in this work. ConTA and SA were particularly useful in identifying design challenges (i.e., naive heuristics and sub-optimal strategies) but also design opportunities (i.e., What functions should the interface support for better performance?). In our work, we focused on non-expert users, however, we believe that identifying how expert users would tackle the same tasks, and analyzing their approach using ConTA and SA can be very beneficial in designing future AI user interfaces that help users gain a better understanding of the task domain. For example, ConTA can identify how experts associate different knowledge states, and SA can identify how experts achieve these knowledge states through adopting efficient strategies. The insights acquired from this analysis can then be used to design interfaces that “guide” non-expert users to adopt a more expert approach.

### 6.4.3 Generalizability

In this work, we focused on a set of specific problems. These included a decision-making task, an AI that is trained to solve a tabular data problem (as opposed to image recognition or natural language processing), and credit and lending domain. These constraints limit the generalizability of the findings to other tasks, AI systems, and domains. For example, a task that involves visual perception (e.g., analysis of X-ray scans) using an image recognition AI in the healthcare domain might have unique challenges that don’t apply to our context, therefore the applicability of our findings are limited to similar task environment. However, decision-making is fundamental to many complex fields, and our findings on how AI influences human decision-making and our approach to leveraging domain knowledge to support users should be applicable to similar problems across domains.

## 6.5 Contributions

The work has a number of contributions to research and design. The main contributions included the following:

- We demonstrated that adding domain knowledge to XAI has a number of benefits to users. Through experimental studies, we explored how domain knowledge affects perceptions and use of AI in a financial decision-making task. We showed that non-expert users can benefit from having access to domain knowledge when interacting

with an AI system. While the role of domain expertise in the context of AI and XAI have been studied in the past, to our knowledge, this is one of the first attempts at supporting users to overcome the limitations of the lack of domain expertise. We also demonstrated that embedding domain knowledge into XAI is valuable in helping users avoid relying on AI when it makes a mistake. Taken together, these findings present opportunities to improve future AI systems (through integration of domain knowledge), as well as contribute to the growing body of research on domain expertise in XAI.

- We presented a novel use case of Cognitive Work Analysis (CWA). We leveraged Work Domain Analysis (WDA) to improve the existing explainable AI techniques. Furthermore, we applied Control Task Analysis and Strategies Analysis to understand how AI influences non-expert users' decision-making process. This work builds on and extends the CWA research on reasoning about automation [19] and modeling automation [104]. This work made a unique contribution to CWA research by investigating an understudied area (XAI) and by demonstrating the usefulness an under-utilized component of CWA (strategies analysis) while providing unique insights that can be leveraged to design future AI systems.
- Through user interviews, we identified perceptions of AI, and how AI is being integrated into decision-making by non-expert users. We identified a number of issues pertaining to XAI, and opportunities to improve design of XAI. Furthermore, our analysis revealed the complexity of designing for non-expert users, and recommendations for future research and design were developed.
- We introduced a testbed that simulates a realistic financial investing decision-making situation that can be used to explore future research questions in the investing space or human-AI interaction. We demonstrated different use cases for the testbed and explored multiple aspects of the human-AI interaction.
- This work also provides insights about how non-expert users engage in investing, and has implications for integrating AI systems to Peer-to-Peer lending platforms and similar investing contexts. We demonstrated how information presented on the user interface influences investors' decision-making processes and their investment decisions.



## 6.6 Limitations

Individual limitations of each study was presented in earlier chapters. Overall, there are several limitations that are important to discuss. First, this work explored a particular task domain, namely credit and lending, which may limit the generalizability of the conclusions to other domains and decision-making situations. Similar work in other domains and task environments will be needed to establish a deeper understanding of this problem space. Second, throughout this work, we framed our participants as non-experts, however, expertise is not a binary construct [157], and users of AI systems will undoubtedly have varying levels of expertise. Defining expertise can be quite challenging [133], and this limits our ability to generalize the findings to populations where expertise levels vary significantly. However, we will be able to situate our findings better as more research is conducted with groups who have varying levels of domain expertise.

Another limitation was that the focus of this work was on a particular AI technology and prediction situation with a limited number of explanation tools (machine learning models with tabular data). The world of AI is vast, and there are numerous prediction problems that require different technical approaches (e.g., image classification problems). Moreover, there is a growing number of explanations techniques, each providing a unique output and language to convey the justification of the AI prediction. While we tried to keep our approach model-agnostic, different machine learning problems will likely have unique properties that we were not able to encounter due to our model and problem choices. Nevertheless, we believe that our general approach should be applicable to other machine learning problems, although the specific findings may differ.

Finally, the studies presented in this work were single-session experiments where participants had limited opportunity to observe and interact with the AI system. Therefore, the findings have limited applicability to long-term use of such systems. It is likely that the way participants utilize AI and form trust will evolve after multiple interactions (Learned trust, [114]), which calls for longitudinal studies to better understand how the relationship between explanations, domain knowledge, and trust evolves over time.

## 6.7 Future Work

In addition to the recommendations for future work described in various chapters, we should mention that there are at least two major lines of research that would have significant benefits. First, the role of domain knowledge should be further studied. This

work presents an initial exploration of some of the ways in which domain knowledge might make a difference, however there are many more questions remained unanswered. For example, in the first study (Chapter 2), we observed that augmenting explanations using domain knowledge have some effect on user perceptions, but it was unclear whether this effect depended on the contextual factors such as ambiguity of the situation. Systematically examining the contextual factors where such an approach is most beneficial remains a challenge. In the second and the subsequent study, we demonstrated that providing domain knowledge led to less reliance on AI when it was incorrect and better task performance. Future studies should examine the “right” level of domain knowledge for different problems, and systematically study how it should be conveyed to the users. Factors such as information amount, visual/presentation factors, and communication language are candidates for future research.

Another opportunity involves leveraging CWA throughout the entire machine learning life cycle. This work focused mainly on the model outputs and real world usage of an AI system, and explored how CWA can be used to support this process. However, we argue that CWA could add significant value to other stages of the machine learning life cycle. Some of these suggestions are discussed earlier in this chapter. These include better integration of domain knowledge into the AI system, and using various stages of CWA to support processes related to problem definition, data collection, feature engineering and model building, and model interpretation in the AI life cycle. While the benefits to end users, as demonstrated in this work, were noteworthy, we firmly believe that CWA can have even more impact on building human-centered AI systems if integrated in earlier stages of the machine learning life cycle. We envision that implementing CWA during the initial stages of a machine learning project, exploring ways to support engineers and developers in key decisions, and creating an AI system that best represents human expertise would be a significant contribution, and remains an important challenge to address going forward.

## 6.8 Conclusion

This work aimed at addressing the limitations of existing XAI approaches by investigating how domain knowledge can be leveraged. Through qualitative and quantitative approaches, we showed potential benefits of communicating domain knowledge to users, and identified opportunities for research and design. We hope that the insights and findings presented in this thesis inform future research on human-centered AI and future design of AI applications.

# References

- [1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.
- [2] Patricia A Alexander. Domain knowledge: Evolving themes and emerging concerns. *Educational psychologist*, 27(1):33–51, 1992.
- [3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.
- [4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.
- [5] Jan Auernhammer. Human-centered ai: The role of human-centered design research in the development of ai. 2020.
- [6] D Bacham and J Zhao. Machine learning: Challenges, lessons, and opportunities in credit risk modeling. *Moody's Analytics Risk Perspectives*, 9, 2017.
- [7] Alexander Bachmann, Alexander Becker, Daniel Buerckner, Michel Hilker, Frank Kock, Mark Lehmann, Phillip Tiburtius, and Burkhardt Funk. Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, 16(2):1, 2011.
- [8] John E Baiden. The 5 c's of credit in the lending industry. *Available at SSRN 1872804*, 2011.

- [9] Vladimir Balayan, Pedro Saleiro, Catarina Belém, Ludwig Krippahl, and Pedro Bizarro. Teaching the machine to explain itself using domain knowledge. *arXiv preprint arXiv:2012.01932*, 2020.
- [10] Ellen J Bass, Leigh A Baumgart, and Kathryn Klein Shepley. The effect of information analysis automation display content on human judgment performance in noisy environments. *Journal of cognitive engineering and decision making*, 7(1):49–65, 2013.
- [11] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. The role of domain expertise in trusting and following explainable ai decision support systems. *Journal of Decision Systems*, pages 1–29, 2021.
- [12] Kevin B Bennett and John M Flach. *Display and interface design: Subtle science, exact art*. CRC Press, 2011.
- [13] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.
- [14] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, page 1, 2017.
- [15] Stewart A Birrell, Mark S Young, Daniel P Jenkins, and Neville A Stanton. Cognitive work analysis for safe and efficient driving. *Theoretical Issues in Ergonomics Science*, 13(4):430–449, 2012.
- [16] Ann M Bisantz and Catherine M Burns. *Applications of cognitive work analysis*. CRC Press, 2008.
- [17] Ann M Bisantz, John D Lee, Jonathan Pfautz, Catherine Burns, William C Elm, and Priyadarshini R Pennathur. Bridging the gap between cognitive systems engineering analysis, design and practice. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 57, pages 334–338. SAGE Publications Sage CA: Los Angeles, CA, 2013.
- [18] Ann M Bisantz and Younho Seong. Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics*, 28(2):85–97, 2001.

- [19] Ann M Bisantz and Kim J Vicente. Making the abstraction hierarchy concrete. *International Journal of human-computer studies*, 40(1):83–117, 1994.
- [20] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.
- [21] Catherine M Burns and John Hajdukiewicz. *Ecological interface design*. CRC Press, 2004.
- [22] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, 3:26, 2020.
- [23] Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.
- [24] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14, 2019.
- [25] Longbing Cao. Ai in finance: A review. *Available at SSRN 3647625*, 2020.
- [26] Irene Celino. Who is this explanation for? human intelligence and knowledge graphs for explainable AI., 2020.
- [27] Shruthi Chari, Daniel M Gruen, Oshani Seneviratne, and Deborah L McGuinness. Directions for explainable knowledge-enabled systems. *arXiv preprint arXiv:2003.07523*, 2020.
- [28] Haiyang Chen and Ronald P Volpe. An analysis of personal financial literacy among college students. *Financial services review*, 7(2):107–128, 1998.
- [29] Jessie YC Chen and Michael J Barnes. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, 44(1):13–29, 2014.

- [30] Jessie YC Chen, Michael J Barnes, Anthony R Selkowitz, Kimberly Stowers, Shan G Lakhmani, and Nicholas Kasdaglis. Human-autonomy teaming and agent transparency. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*, pages 28–31, 2016.
- [31] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science*, 19(3):259–282, 2018.
- [32] Li Chen and Pearl Pu. Trust building in recommender agents. In *Proceedings of the Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces at the 2nd International Conference on E-Business and Telecommunication Networks*, pages 135–145. Citeseer, 2005.
- [33] Michelene TH Chi, Robert Glaser, and Marshall J Farr. *The nature of expertise*. Psychology Press, 2014.
- [34] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*, pages 307–317, 2021.
- [35] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. Toward personalized xai: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298:103503, 2021.
- [36] Maia B Cook and Harvey S Smallman. Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors*, 50(5):745–754, 2008.
- [37] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna Van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. Intellingo: An intelligible translation environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.
- [38] Miranda Cornelissen, Paul M Salmon, Daniel P Jenkins, and Michael G Lenné. How can they do it? a structured approach to the strategies analysis phase of cognitive work analysis. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 55, pages 340–344. SAGE Publications Sage CA: Los Angeles, CA, 2011.

- [39] Méлина Côté and Benoît Lamarche. Artificial intelligence in nutrition research: perspectives on current and future applications. *Applied Physiology, Nutrition, and Metabolism*, (ja), 2021.
- [40] Marsha Courchane and Peter Zorn. Consumer literacy and creditworthiness. *Proceedings, Federal Reserve Bank of Chicago*, 2005.
- [41] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.
- [42] Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. User acceptance of computer technology: a comparison of two theoretical models. *Management science*, 35(8):982–1003, 1989.
- [43] Joachim de Greeff, Maaïke HT de Boer, Fieke HJ Hillerström, Freek Bomhof, Wiard Jorritsma, and Mark A Neerincx. The fate system: Fair, transparent and explainable decision making. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, 2021.
- [44] Murat Dikmen and Catherine Burns. Abstraction hierarchy based explainable artificial intelligence. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 64, pages 319–323. SAGE Publications Sage CA: Los Angeles, CA, 2020.
- [45] Murat Dikmen and Catherine Burns. The effects of domain knowledge on trust in explainable ai and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, page 102792, 2022.
- [46] Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*, 2017.
- [47] S Dorton and Samantha Harper. Trustable ai: A critical challenge for naval intelligence. *Center for International Maritime Security (CIMSEC)*. Retrieved from: <https://cimsec.org/trustable-ai-a-critical-challenge-for-naval-intelligence>, 2021.
- [48] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [49] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

- [50] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
- [51] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.
- [52] Simon D’Alfonso. Ai in mental health. *Current Opinion in Psychology*, 36:112–117, 2020.
- [53] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274. ACM, 2019.
- [54] Alberto Fernandez, Francisco Herrera, Oscar Cordon, Maria Jose del Jesus, and Francesco Marcelloni. Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational intelligence magazine*, 14(1):69–81, 2019.
- [55] FICO. Fico score 8 and why there are multiple versions of fico scores.
- [56] Raya Fidel and Annelise Mark Pejtersen. From information behaviour research to the design of information systems: The cognitive work analysis framework. *Information Research: an international electronic journal*, 10(1):n1, 2004.
- [57] James Forbes and S Murat Kara. Confidence mediates how investment knowledge influences investing self-efficacy. *Journal of economic psychology*, 31(3):435–443, 2010.
- [58] Masaru Fuji, Katsuhito Nakazawa, and Hiroaki Yoshida. “trustworthy and explainable AI” achieved through knowledge graphs and social implementation. *FUJITSU SCIENTIFIC & TECHNICAL JOURNAL*, 56(1):39–45, 2020.
- [59] Giuseppe Futia and Antonio Vetrò. On the integration of knowledge graphs into deep learning models for a more comprehensible ai—three challenges for future research. *Information*, 11(2):122, 2020.
- [60] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lerner, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):1–8, 2021.



- [61] Julie Gerlings, Millie Søndergaard Jensen, and Arisa Shollo. Explainable ai, but explainable to whom? *arXiv preprint arXiv:2106.05568*, 2021.
- [62] Julie Gerlings, Arisa Shollo, and Ioanna Constantiou. Reviewing the need for explainable artificial intelligence (xAI). *arXiv preprint arXiv:2012.01007*, 2020.
- [63] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction*, pages 295–303. Springer, 2018.
- [64] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, 38(3):50–57, 2017.
- [65] goPeer. Rates and fees. <https://gopeer.ca/fees/>, n.d. Accessed: 2021-06-22.
- [66] Alex Gramegna and Paolo Giudici. Why to buy insurance? an explainable artificial intelligence approach. *Risks*, 8(4):137, 2020.
- [67] Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, pages 497–530, 1999.
- [68] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.
- [69] David Gunning. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2(2), 2017.
- [70] David Gunning and David Aha. Darpa’s explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, 2019.
- [71] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, 2019.
- [72] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. Darpa’s explainable ai (xai) program: A retrospective, 2021.
- [73] Jessica Hamzelou. Ai system is better than human doctors at predicting breast cancer, Jan 2020.

- [74] Peter A Hancock, Richard J Jagacinski, Raja Parasuraman, Christopher D Wickens, Glenn F Wilson, and David B Kaber. Human-automation interaction research: Past, present, and future. *Ergonomics in Design*, 21(2):9–14, 2013.
- [75] Maureen E Hassall and Penelope M Sanderson. A formative approach to the strategies analysis phase of cognitive work analysis. *Theoretical Issues in Ergonomics Science*, 15(3):215–261, 2014.
- [76] Antony Hilliard, Laura Thompson, and Cam Ngo. Demonstrating cwa strategies analysis: a case study of municipal winter maintenance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52, pages 262–266. SAGE Publications Sage CA: Los Angeles, CA, 2008.
- [77] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.
- [78] Robert Hoffman, Shane Mueller, Gary Klein, and Jordan Litman. Measuring trust in the xai context. 2021.
- [79] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.
- [80] Jeanne M Hogarth and Marianne A Hilgert. Financial knowledge, experience and learning preferences: Preliminary results from a new survey on financial literacy. *Consumer Interest Annual*, 48(1):1–7, 2002.
- [81] Andreas Holzinger. From machine learning to explainable ai. In *2018 world symposium on digital intelligence for systems and machines (DISA)*, pages 55–66. IEEE, 2018.
- [82] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.
- [83] Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the quality of explanations: the system causability scale (scs). *KI-Künstliche Intelligenz*, 34(2):193–198, 2020.
- [84] Andreas Holzinger, Bernd Malle, Anna Saranti, and Bastian Pfeifer. Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. *Information Fusion*, 71:28–37, 2021.

- [85] Andreas T Holzinger and Heimo Muller. Toward human–ai interfaces to support explainability and causability in medical ai. *Computer*, 54(10):78–86, 2021.
- [86] Miroslav Hudec, Erika Mináriková, Radko Mesiar, Anna Saranti, and Andreas Holzinger. Classification by ordinal sums of conjunctive and disjunctive functions for explainable ai and interpretable machine learning solutions. *Knowledge-Based Systems*, 220:106916, 2021.
- [87] Sheikh Rabiul Islam, William Eberle, Sid Bundy, and Sheikh Khaled Ghafoor. Infusing domain knowledge in AI-based” black box” models for better explainability with application in bankruptcy prediction. *arXiv preprint arXiv:1905.11474*, 2019.
- [88] Sheikh Rabiul Islam, William Eberle, Sheikh K Ghafoor, Ambareen Siraj, and Mike Rogers. Domain knowledge aided explainable artificial intelligence for intrusion detection and response. *arXiv preprint arXiv:1911.09853*, 2019.
- [89] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*, 2021.
- [90] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational psychiatry*, 11(1):1–9, 2021.
- [91] Marijn Janssen, Martijn Hartog, Ricardo Matheus, Aaron Yi Ding, and George Kuk. Will algorithms blind people? the effect of explainable ai and decision-makers’ experience on ai-supported decision-making in government. *Social Science Computer Review*, page 0894439320980118, 2020.
- [92] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71, 2000.
- [93] Helen Jiang and Erwen Senge. On two xai cultures: A case study of non-technical explanations in deployed ai system. *arXiv preprint arXiv:2112.01016*, 2021.
- [94] Brian Kalis, Matt Collier, and Richard Fu. 10 promising ai applications in health care. *Harvard Business Review*, 2018.

- [95] Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. *arXiv preprint arXiv:2103.01035*, 2021.
- [96] Eoin M Kenny, Courtney Ford, Molly Quinn, and Mark T Keane. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, 294:103459, 2021.
- [97] Gary Klein, Robert Hoffman, Shane Mueller, and Emily Newsome. Modeling the process by which people try to explain complex things to others. *Journal of Cognitive Engineering and Decision Making*, page 15553434211045154, 2021.
- [98] Ilker Koksall. How ai determines the diet plans, Mar 2020.
- [99] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: A survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.
- [100] Freddy Lecue. On the role of knowledge graphs in explainable AI. *Semantic Web*, 11(1):41–51, 2020.
- [101] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [102] John Lee and Neville Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992.
- [103] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.
- [104] Yeti Li and Catherine M Burns. Modeling automation with cognitive work analysis to support human-automation coordination. *Journal of cognitive engineering and decision making*, 11(4):299–322, 2017.
- [105] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.
- [106] Q Vera Liao and Kush R Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.

- [107] Shu-Hsien Liao. Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert systems with applications*, 28(1):93–103, 2005.
- [108] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128, 2009.
- [109] Roman Lukyanenko, Arturo Castellanos, Veda C Storey, Alfred Castillo, Monica Chiarini Tremblay, and Jeffrey Parsons. Superimposition: Augmenting machine learning outputs with conceptual models for explainable AI. In *International Conference on Conceptual Modeling*, pages 26–34. Springer, 2020.
- [110] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- [111] Maya S Luster and Brandon J Pitts. Trust in automation: The effects of system certainty on decision-making. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 65, pages 32–36. SAGE Publications Sage CA: Los Angeles, CA, 2021.
- [112] Jan Maarten Schraagen, Sabin Kerwien Lopez, Carolin Schneider, Vivien Schneider, Stephanie Tönjes, and Emma Wiechmann. The role of transparency and explainability in automated systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 65, pages 27–31. SAGE Publications Sage CA: Los Angeles, CA, 2021.
- [113] Tauseef Ibne Mamun, Kenzie Baker, Hunter Malinowski, Rober R Hoffman, and Shane T Mueller. Assessing collaborative explanations of ai using explanation goodness criteria. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 65, pages 988–993. SAGE Publications Sage CA: Los Angeles, CA, 2021.
- [114] Stephen Marsh and Mark R Dibben. The role of trust in information science and technology. *Annual Review of Information Science and Technology (ARIST)*, 37:465–98, 2003.

- [115] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human factors*, 58(3):401–415, 2016.
- [116] Massimo Micocci, Simone Borsci, Viral Thakerar, Simon Walne, Yasmine Manshadi, Finlay Edridge, Daniel Mullarkey, Peter Buckle, and George B Hanna. Do gps trust artificial intelligence insights and what could this mean for patient care? a case study on gps skin cancer diagnosis in the uk. 2021.
- [117] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.
- [118] Branka Hadji Misheva, Joerg Osterrieder, Ali Hirsra, Onkar Kulkarni, and Stephen Fung Lin. Explainable AI in credit risk management. *arXiv preprint arXiv:2103.00949*, 2021.
- [119] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1902.01876*, 2019.
- [120] Bonnie M Muir and Neville Moray. Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460, 1996.
- [121] Neelam Naikar. Beyond interface design: Further applications of cognitive work analysis. *International journal of industrial ergonomics*, 36(5):423–438, 2006.
- [122] Neelam Naikar. An examination of the key concepts of the five phases of cognitive work analysis with examples from a familiar system. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 447–451. SAGE Publications Sage CA: Los Angeles, CA, 2006.
- [123] Neelam Naikar. Cognitive work analysis: An influential legacy extending beyond human factors and engineering. *Applied ergonomics*, 59:528–540, 2017.
- [124] Neelam Naikar, Anna Moylan, and Brett Pearce. Analysing activity in complex systems with cognitive work analysis: concepts, guidelines and case study for control task analysis. *Theoretical Issues in Ergonomics Science*, 7(4):371–394, 2006.

- [125] Mohammad Naiseh, Reem S Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. Nudging through friction: an approach for calibrating trust in explainable ai.
- [126] Feng Ni, David Arnott, and Shijia Gao. The anchoring effect in business intelligence supported decision-making. *Journal of Decision Systems*, 28(2):67–81, 2019.
- [127] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.
- [128] Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. Towards explainable AI: Assessing the usefulness and impact of added explainability features in legal document summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.
- [129] Mahsan Nourani, Joanie King, and Eric Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 112–121, 2020.
- [130] Oxera. The economics of peer-to-peer lending. Technical report, Sep 2016.
- [131] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.
- [132] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297, 2000.
- [133] Vimla L Patel, José F Arocha, and Jiajie Zhang. Thinking and reasoning in medicine. *The Cambridge handbook of thinking and reasoning*, 14:727–750, 2005.
- [134] Vanessa Gail Perry. Is ignorance bliss? consumer accuracy in judgments about credit ratings. *Journal of Consumer Affairs*, 42(2):189–205, 2008.
- [135] Catia Pesquita. Towards semantic integration for explainable artificial intelligence in the biomedical domain. In *HEALTHINF*, pages 747–753, 2021.
- [136] Jens Rasmussen. Outlines of a hybrid model of the process plant operator. In *Monitoring behavior and supervisory control*, pages 371–383. Springer, 1976.
- [137] Jens Rasmussen, Annelise Mark Pejtersen, and Len P Goodstein. Cognitive systems engineering. 1994.

- [138] Jens Rasmussen, Annelise Mark Pejtersen, and Kjeld Schmidt. *Taxonomy for cognitive work analysis*. 1990.
- [139] Gemma JM Read, Paul M Salmon, and Michael G Lenné. From work analysis to work design: A review of cognitive work analysis design applications. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 56, pages 368–372. SAGE Publications Sage CA: Los Angeles, CA, 2012.
- [140] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.
- [141] Mark O Riedl. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36, 2019.
- [142] Jennifer M Ross, James L Szalma, Peter A Hancock, John S Barnett, and Grant Taylor. The effect of automation reliability on user automation trust and reliance in a search-and-rescue scenario. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52, pages 1340–1344. Sage Publications Sage CA: Los Angeles, CA, 2008.
- [143] Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *stat*, 1050:26, 2018.
- [144] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *arXiv preprint arXiv:2111.06420*, 2021.
- [145] Julian Sanchez, Wendy A Rogers, Arthur D Fisk, and Ericka Rovira. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, 15(2):134–160, 2014.
- [146] Melissa Saragih and Ben W Morrison. The effect of past algorithmic performance and decision significance on algorithmic advice acceptance. *International Journal of Human-Computer Interaction*, pages 1–10, 2021.
- [147] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3):377–400, 2016.



- [148] Mike Schaekermann. *Human-AI Interaction in the Presence of Ambiguity: From Deliberation-based Labeling to Ambiguity-aware AI*. PhD thesis, 2020.
- [149] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. Ambiguity-aware AI assistants for medical data analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.
- [150] James Schaffer, John O’Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 240–251, 2019.
- [151] Thomas B Sheridan and William L Verplank. Human and computer control of undersea teleoperators. Technical report, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab, 1978.
- [152] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146:102551, 2021.
- [153] Donghoon Shin, Sachin Grover, Kenneth Holstein, and Adam Perer. Characterizing human explanation strategies to inform the design of explainable ai for building damage assessment. *arXiv preprint arXiv:2111.02626*, 2021.
- [154] Ben Shneiderman. Human-centered artificial intelligence: three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3):109–124, 2020.
- [155] Keng Siau and Weiyu Wang. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, 31(2):47–53, 2018.
- [156] Aaron Springer and Steve Whittaker. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 107–120, 2019.
- [157] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

- [158] CHEN Tian, XU Manfei, TU Justin, WANG Hongyue, and NIU Xiaohui. Relationship between omnibus and post-hoc tests: An investigation of performance of the f test in anova. *Shanghai archives of psychiatry*, 30(1):60, 2018.
- [159] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374, 2019.
- [160] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404, 2021.
- [161] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003.
- [162] Kim J Vicente. Task analysis, cognitive task analysis, cognitive work analysis: what’s the difference? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 39, pages 534–537. SAGE Publications Sage CA: Los Angeles, CA, 1995.
- [163] Kim J Vicente. *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. CRC Press, 1999.
- [164] Kim J Vicente and Jens Rasmussen. Ecological interface design: Theoretical foundations. *IEEE Transactions on systems, man, and cybernetics*, 22(4):589–606, 1992.
- [165] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr.(2017). *Harvard Journal of Law & Technology*, 31:841, 2017.
- [166] Fei-Yue Wang. Toward a revolution in transportation operations: Ai for complex systems. *IEEE Intelligent Systems*, 23(6):8–13, 2008.
- [167] Lu Wang, Greg A Jamieson, and Justin G Hollands. Trust and reliance on an automated combat identification system. *Human factors*, 51(3):281–291, 2009.
- [168] Ning Wang, David V Pynadath, and Susan G Hill. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 109–116. IEEE Press, 2016.

- [169] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in AI-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.
- [170] Elke U Weber, Ann-Renee Blais, and Nancy E Betz. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of behavioral decision making*, 15(4):263–290, 2002.
- [171] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. ” do you trust me?” increasing user-trust by integrating virtual agents in explainable ai interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 7–9, 2019.
- [172] Christopher D Wickens, Shaw L Ketels, Alice F Healy, Carolyn J Buck-Gengler, and Lyle E Bourne Jr. The anchoring heuristic in intelligence integration: A bias in need of de-biasing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54, pages 2324–2328. SAGE Publications Sage CA: Los Angeles, CA, 2010.
- [173] Christopher D Wickens, Huiyang Li, Amy Santamaria, Angelia Sebok, and Nadine B Sarter. Stages and levels of automation: An integrated meta-analysis. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 54, pages 389–393. Sage Publications Sage CA: Los Angeles, CA, 2010.
- [174] Roger Wohlner. How ai is shaping the advisory landscape, Sep 2021.
- [175] George Wright and Peter Ayton. Eliciting and modelling expert knowledge. *Decision Support Systems*, 3(1):13–26, 1987.
- [176] Wei Xu. Toward human-centered ai: a perspective from human-computer interaction. *interactions*, 26(4):42–46, 2019.
- [177] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020.
- [178] Hui-Nee Au Yong and Kock-Lim Tan. The influence of financial literacy towards risk tolerance. *International journal of business and society*, 18(3):469–484, 2017.

- [179] Tomasz Zaleskiewicz. Beyond risk seeking and risk aversion: Personality and the dual nature of economic risk taking. *European journal of Personality*, 15(S1):S105–S122, 2001.
- [180] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.
- [181] Jianlong Zhou, Syed Z Arshad, Simon Luo, and Fang Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP conference on human-computer interaction*, pages 23–39. Springer, 2017.
- [182] Jianlong Zhou, Syed Z Arshad, Kun Yu, and Fang Chen. Correlation for user confidence in predictive decision making. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, pages 252–256, 2016.
- [183] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

# APPENDICES

# Appendix A

## Study Materials and Additional Data Analysis

### A.1 Study 1

#### A.1.1 Study Materials

##### Loan Knowledge Quiz

In this section, we present the questions we used to assess participant's knowledge about loan decision process. We developed our own questions in addition to compiling questions from two financial literacy related questionnaires [134, 80]. **Bold** items are correct answers.

Q1: How much do you know about loan decision process?

- Nothing
- Very little
- Some
- A fair amount
- A lot

Q2: Loans are approved only based on the applicant's credit score

- True
- **False**

Q3: What is the threshold for credit card utilization (the ratio of the money you spent and the credit card limit) for an applicant to be considered safe?

- 10%
- 20%
- **30%**
- 40%

Q4: Which of the following is not included when assessing an applicant's credit score?

- Credit card utilization (the ratio of the money you spent and the credit card limit)
- **Income / Debt Ratio**
- Payment history
- Amounts Owed

Q5: For a loan application to be approved, Debt-Income Ratio (DTI) is considered.

- **True**
- False

Q6: Which of the following is considered as part of 5C's of Credit

- Capacity
- Capital
- Character
- **All of the above are considered important**

Q7: When applying for loans, the loan purpose is not important. The monthly installment rate - income ratio is enough for a lender to make a decision.

- True
- **False**

Q8: Your net worth is

- The difference between your expenditures and income.
- **The difference between your liabilities and assets.**
- The difference between your cash inflow and outflow.
- The difference between your bank borrowings and savings.

Q9: You will improve your creditworthiness by

- Visiting your local commercial bank.
- **Showing no record of personal bankruptcies in recent years.**
- Paying cash for all goods and services.
- Borrowing large amounts of money from your friends.
- Donating money to charity.

Q10: If you co-sign a loan for a friend, then

- **You become responsible for the loan payments if your friend defaults.**
- It means that your friend cannot receive the loan by himself.
- You are entitled to receive part of the loan.
- Both A and B.
- Both A and C.



## Questions From [80]

Q11: Making payments late on your bills can make it more difficult to take out a loan.

- **True**
- False

Q12: Your credit rating is not affected by how much you charge on your credit cards.

- True
- **False**

Q13: Your credit report includes employment data, your payment history, any inquiries made by creditors, and any public record information.

- **True**
- False

Q14: When you use your home as collateral for a loan, there is no chance of losing your home

- True
- **False**

## Questions From [134]

*What will be the impact on the interest rate people pay on a loan if they...?*

Q15: ...always eventually pay off their debts, but are sometimes late on monthly bills

- **Rate would be higher**
- No impact
- Rate would be lower

Q16: ...get someone else to co-sign the loan with them

- Rate would be higher
- No impact
- **Rate would be lower**

Q17: ...have never borrowed money before

- **Rate would be higher**
- No impact
- Rate would be lower

Q18: ...offer the lender some collateral for the loan

- Rate would be higher
- No impact
- **Rate would be lower**

*What will be the impact on a person's credit rating if they...?*

Q19: ...charge lots of money on several credit cards, and make the minimum payments each month

- **Rating would be hurt**
- No impact
- Rating would improve

Q20: ...have a good payment record and apply for many new credit cards

- **Rating would be hurt**
- No impact
- Rating would improve

Q21: ...skip a student loan payment

- **Rating would be hurt**
- No impact
- Rating would improve

Q22: ...never borrow money or use a credit card for anything

- **Rating would be hurt**
- No impact
- Rating would improve

Q23: ...miss a couple of loan payments but make them up, plus interest, the next month

- **Rating would be hurt**
- No impact
- Rating would improve

### AH-Based Explanations

In this section, we present the explanation texts that we created by using AH. We tried to keep the length similar across explanations, although there was some variance.

Table A.1: AH-based explanations

Concept	Explanation	Word Count
Loan Amount	<b>Loan Amount</b> along with other loan attributes such as <b>Loan Duration</b> and <b>Loan Purpose</b> are used to assess overall <b>loan conditions</b> such interest rates and monthly installments, and are indicators of the burden on the borrower. In general, lower loan amount and longer loan duration lead to better loan conditions and a higher chance of approval.	56

Loan Duration	<b>Loan Duration</b> along with other loan attributes such as <b>Loan Amount</b> and <b>Loan Purpose</b> are used to assess overall <b>loan conditions</b> such interest rates and monthly installments, and are indicators of the burden on the borrower. In general, longer loan duration and lower loan amount leads to better loan conditions and a higher chance of approval.	56
Credit History	<b>Credit History</b> of the applicant is used to assess <b>credit score</b> , which determines how <b>trustworthy</b> the applicant is. <b>Unpaid debt</b> also affects credit score. In general, better credit history (e.g. all credits paid back timely) and less unpaid debts lead to better credit score, more trustworthiness and a higher chance of approval	52
Other Debtors	Having <b>Other Debtors</b> can be an indicator of whether the borrower can comfortably <b>afford the payments</b> (in case of a default). <b>Income</b> also affects the ability to afford the payments. In general, having another debtor such as a co-applicant or a guarantor and a higher income lead to better ability to afford payments and a higher chance of approval.	59
Most Valuable Asset	<b>Assets</b> such as real estate and other properties are good indicators of the <b>capital</b> the borrower can use to repay the loan if they are short on income. <b>Account balances</b> can also be used for this purpose. In general, more valuable assets and higher account balances lead to more capital and a higher chance of approval.	56
Chequing Account Balance	<b>Chequing and Savings Account Balances</b> are good indicators of the <b>capital</b> the borrower can use to repay the loan if they are short on income. <b>Assets</b> such as properties can also be used for this purpose. In general, higher account balances and more valuable assets lead to more capital, and a higher chance of approval.	55

Savings Account Balance	<b>Savings and Chequing Account Balances</b> are good indicators of the <b>capital</b> the borrower can use to repay the loan if they are short on income. <b>Assets</b> such as properties can also be used for this purpose. In general, higher account balances and more valuable assets lead to more capital and a higher chance of approval.	55
Credits At This Bank	<b>Debts</b> such as <b>Credits At This Bank</b> are considered when assessing whether the borrower can comfortably <b>afford the payments</b> . <b>Income</b> also affects the ability to afford the payments. In general, lower debt and higher income lead to better ability to afford the payments and a higher chance of approval.	49
Housing	<b>Expenses</b> such as <b>rent</b> and the number of <b>dependents</b> are considered when assessing whether the borrower can comfortably <b>afford the payments</b> . In general, not paying rent and fewer dependents lead to better ability to afford the payments and a higher chance of approval.	43

### A.1.2 Ethics Approval

This study was approved by the University of Waterloo Research Ethics Committee (Figure A.1).

### A.1.3 Recruitment

The following recruitment material was posted on Amazon Mechanical Turks' job board as a HIT (Human Intelligence Task):

**Title: Survey about Artificial Intelligence in Loan Decisions**

**Description:** This is a survey study conducted by researchers from the University of Waterloo, Ontario, Canada. In this study, you will see loan approval/rejection decisions made by an Artificial Intelligence system, and you

## UNIVERSITY OF WATERLOO

### Notification of Ethics Clearance to Conduct Research with Human Participants

---

Principal Investigator: Catherine Burns (Systems Design Engineering)

Student investigator: Murat Dikmen (Systems Design Engineering)

File #: 41245

Title: CWA-based Explainable AI

---

The Human Research Ethics Committee is pleased to inform you this study has been reviewed and given ethics clearance.

**Initial Approval Date: 10/07/19 (m/d/y)**

University of Waterloo Research Ethics Committees are composed in accordance with, and carry out their functions and operate in a manner consistent with, the institution's guidelines for research with human participants, the Tri-Council Policy Statement for the Ethical Conduct for Research Involving Humans (TCPS, 2nd edition), International Conference on Harmonization: Good Clinical Practice (ICH-GCP), the Ontario Personal Health Information Protection Act (PHIPA), the applicable laws and regulations of the province of Ontario. Both Committees are registered with the U.S. Department of Health and Human Services under the Federal Wide Assurance, FWA00021410, and IRB registration number IRB00002419 (HREC) and IRB00007409 (CREC).

This study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

**Expiry Date: 10/08/20 (m/d/y)**

Multi-year research must be renewed at least once every 12 months unless a more frequent review has otherwise been specified. Studies will only be renewed if the renewal report is received and approved before the expiry date. Failure to submit renewal reports will result in the investigators being notified ethics clearance has been suspended and Research Finance being notified the ethics clearance is no longer valid.

Level of review: Delegated Review

Signed on behalf of the Human Research Ethics Committee



This above named study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Documents reviewed and received ethics clearance for use in the study and/or received for information:

file: Explanation\_of\_Revisions\_Version1\_20190827.pdf

file: FeedbackLetter\_Version2\_20190827.pdf

Figure A.1: Research Ethics Committee approval for Study I.

will be asked to provide your opinions. The study is estimated to take 60 minutes, and you will receive a remuneration of \$4. There will be about 80 questions. Some of the questions will be presented in a multiple-choice format, others as open-ended questions. In addition to questions about your opinions on the Artificial Intelligence system presented in the survey, there will be demographics questions such as age, gender, and education, and additional questions regarding your knowledge about and experience with loan decision process and technology use.

When a potential participant accepted the HIT, they were shown a link that took them to the survey platform (Qualtrics). For the HIT, we targeted Amazon Mechanical Turk users from the United States and Canada who had over 95% HIT approval rating.

#### **A.1.4 Information Letter and Consent Form**

Before each session, an information letter and a consent form were presented on the survey platform.

##### **Information Letter**

###### **Study Overview**

You are invited to participate in a research study conducted by Murat Dikmen, under the supervision of Catherine Burns at the Systems Design Engineering Department of the University of Waterloo, Ontario, Canada. The objective of the research study is to understand people's opinions on loan decision maker artificial intelligence (AI) systems. The study is for a PhD thesis.

###### **What You Will Be Asked to Do**

If you decide to volunteer, you will be asked to complete a 60-minute online study (in the survey format) that is completed anonymously. The study focuses on your opinions on explanations provided by an artificial intelligence system. There will be about 80 questions. Some of the questions will be presented in a multiple-choice format, others as open-ended questions. In addition to questions about your opinions on the Artificial Intelligence system presented in the survey, there will be demographics questions such as age, gender, and education, and additional questions regarding your knowledge about and experience

with loan decision process and technology use. Participation in this study is voluntary. You may decline to answer any questions that you do not wish to answer and you can withdraw your participation at any time by not submitting your responses. There are no known or anticipated risks from participating in this study.

### **Remuneration**

In appreciation of your time, you will be paid \$4 through the Amazon Mechanical Turk system. To receive your payment, you need to use the code that you receive at the end of the survey. If you wish to withdraw at any time, you can still receive the payment by clicking next until you see the last page of the survey and use the code.

### **Personal Benefits of the Study**

There are no direct benefits to the participants of this study.

### **Confidentiality and Data**

When information is transmitted over the internet confidentiality cannot be guaranteed. University of Waterloo practices are to turn off functions that collect machine identifiers such as IP addresses. The host of the system collecting the data such as Qualtrics may collect this information without our knowledge and make this accessible to us. We will not use or save this information without your consent.

Because this is an anonymous survey the researchers have no way of identifying you or getting in touch with you should you choose to tell us something about yourself or your life experiences. Also, after the survey responses submitted, it is not possible to withdraw them as we don't know which response belongs to the you.

The data, with no personal identifiers, collected from this study will be maintained on a password-protected computer database in a restricted access area of the university. As well, the data will be electronically archived after completion of the study and maintained for at least 10 years.

Data may be deposited in an online public repository/database. Data will be de-identified prior to submission to the repository/database. This process is integral to the research process as it allows other researchers to verify results and avoid duplicating research.

Should you have any questions about the study, please contact either Murat Dikmen at [murat.dikmen@uwaterloo.ca](mailto:murat.dikmen@uwaterloo.ca) or Catherine Burns at



catherine.burns@uwaterloo.ca. Furthermore, if you would like to receive a copy of the results of this study, please contact either investigator. The survey results will be available in November 2019.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #41245). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

Thank you for considering participation in this study.

## Consent Form

**Please read carefully and indicate your consent by clicking one of the buttons below.**

Study Title: Opinions on Artificial Intelligence in Loan Decisions

I have had the opportunity to contact researchers and ask any questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted. I am aware that I may withdraw from the study without penalty at any time.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #41245). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

For all other questions contact Murat Dikmen at murat.dikmen@uwaterloo.ca. You may also contact Professor Catherine Burns at 519-888-4567 Ext. 33903. (catherine.burns@uwaterloo.ca)

By submitting the attached survey, you are not waving your legal rights or releasing the investigators or involved institution from their legal and professional responsibilities.

With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

I agree to participate.

I do not wish to participate (please return the HIT).

### A.1.5 Procedure

After a participant clicked on the survey link presented in the HIT, they were redirected to a Qualtrics survey. They read the information letter and indicated their consent on the consent form by clicking on a button. After giving the consent, the next page provided a brief introduction to the study, followed by a set of demographics and background questions. The following questions were displayed in this section:

What is your age? (open-ended)

What is your gender?

- Female
- Male
- Other (open-ended)

What is the highest degree or level of school you have completed? (If you are currently enrolled in school, please indicate the highest degree you have received).

- Less than high school
- Graduated high school
- Trade/technical school
- Some college, no degree
- Associate degree
- Bachelor's degree
- Advanced degree (Master's, Ph.D., M.D.)

Are you currently a student?

- Yes
- No

[If Student is Yes] Please indicate the degree you are studying for.

- High School
- Trade/technical school

- Associate degree
- Bachelor’s degree
- Advanced degree (Master’s, Ph.D., M.D.)

Next, participants were shown the Loan Knowledge Quiz described in [A.1.1](#). Finally, a number of questions were asked about the participant’s experience with AI:

Please indicate your agreement with the following statements. (from **Strongly Disagree** to **Strongly Agree** on a 7-point scale)

- I am confident using computers.
- I can make use of computer programming to solve a problem.
- I understand how Amazon recommends products for me to purchase.
- I understand how my email provider’s spam filter works.
- I understand how self-driving cars drive on their own.
- I understand how autopilot works on an airplane.
- I understand how YouTube recommends videos for me to watch.
- I understand how Facebook Newsfeed chooses which posts to show before everything else.

Have you had any experience with an AI program that makes loan approval / rejection decisions?

- Yes
- No

[If Yes] Please describe your experience with the AI program that makes loan approval / rejection decisions. (open-ended)

How much do you know how Artificial Intelligence algorithms work? (from **Nothing** to **A lot** on a 5-point scale)

How much programming knowledge do you have? (from **Nothing** to **A lot** on a 5-point scale)

How much knowledge of computer algorithms do you have? (from **Nothing** to **A lot** on a 5-point scale)

After completing the background section, the following description was presented:

### **Information about the Next Section**

In the next section of the study, you will see several loan applicant profiles (hypothetical people who applied for a loan). The details about the person's finances will be provided, such as the loan amount they requested, their account balance, their employment status, and so on. You will also see loan approval or rejection decisions made by an Artificial Intelligence (AI) program. The AI program will also provide explanations regarding its decisions. After you get familiarity with the AI program, the loan approval and rejection decisions, and individual factors that play a role in these decisions, you will be asked several questions regarding the AI program and its decisions.

### **What is Artificial Intelligence?**

Artificial Intelligence (AI) refers to capable programs that can analyze large amounts of data and find patterns. This process is also called training the AI system. Once the AI system is trained, it can make predictions on the data it has never seen before. A common AI program that we use everyday is when email providers label certain emails as "spam". The AI programs that label the emails as "spam" have seen lots and lots of emails, and whether or not they are labelled as "spam" by users. Over time, the AI programs learn how to differentiate "spam" emails from normal ones. For example, the AI programs may understand how the words used in "spam" emails differ from the words used in regular emails. When you receive an email, these AI programs check whether the email looks like the "spam" emails it has learned in the past, and if so, they label the email as "spam".

### **The AI Program**

In this study, we introduce an AI program that has learned how people's financial situation affects their loan applications. This program is trained on customer data of more than 1000 people who applied for a loan and it has learned to predict whether the loan application should be approved or rejected.

The AI program acts like a bank and makes a loan approval or rejection decision based on the applicant's financial situation. It can also provide explanations regarding the decisions and why it made a certain decision. Below are annotated images that show various information AI program provides.

After this introduction, the study proceeded as described in Chapter 2. At the end of the study, the following questions were presented:

From 1 (None at all) to 7 (A great deal), how much did the explanations provided by the AI program help you learn about loan decision process?

From 1 (Too little) to 7 (Too much), overall, the information provided by the AI program when explaining the decisions was...

Please explain how loan decision process works. Provide as much detail as you can. (open-ended)

Please rate the following statement: I can predict how the AI program will behave when making loan decisions. (from **Strongly Disagree** to **Strongly Agree** on a 7-point scale)

What are your thoughts on the loan AI program? Do you have any suggestions to improve it? (open-ended)

Which parts of the AI program performed as you expected?

Which parts of the AI program did not perform as you expected?

At the end of the study, the following feedback letter was shown to appreciate their participation:

I would like to thank you for your participation in this study. As a reminder, the purpose of this study was to understand people's opinions on artificial intelligence and explanations of its decisions. This study will help inform the future research on making artificial intelligence more understandable and trustworthy. Below you can find the code you need to complete the HIT on Amazon Mechanical Turk.

The study is designed such that half of the participants saw limited AI explanations (without additional explanations at the end of the profile page) and the other half with enhanced AI explanations (with additional explanations at the end of the profile page). We hypothesized that participants who read enhanced explanations will rate the quality of the explanations higher than participants who read limited explanations.

The data, with no personal identifiers, collected from this study will be maintained on a password-protected computer database in a restricted access area of the university. In addition, the data will be electronically archived after completion of the study and maintained for at least 10 years.

If you are interested in receiving more information regarding this study or a summary of its results, or if you have any questions or concerns, please contact

us at the email addresses listed at the bottom of this page. The results of this study will be available in November 2019.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #41245). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

## **A.1.6 Data Analysis**

In this section, we present additional data analysis that are complementary to the results presented in Chapter 2.

### **Perceived Ability to Predict AI**

We asked participants to rate the following question, from 1 to 7: I can predict how the AI program will behave when making loan decisions. There was a marginally significant difference,  $t(95) = 1.80$ ,  $p = .075$ . Participants in the AH-based condition provided slightly lower ratings ( $M = 4.54$ ,  $SD = .99$ ) than participants in the baseline condition ( $M = 4.94$ ,  $SD = 1.13$ ).

### **Trust in AI**

We used Checklist for Trust between People and Automation [92] to measure trust towards AI. There was no difference between conditions,  $t = .66$ ,  $p = .51$ .

### **Perceptions of Explanations and Domain Knowledge**

We measured participants' existing knowledge about loans and credit using the loan knowledge quiz described earlier in this Appendix. To investigate how existing knowledge affected perceptions of explanations, we split participants into two groups (less domain knowledge, more domain knowledge) based on their performance in the quiz (using a median-split), and performed t-tests on explanation quality questions. We did not report these findings as the measurement instrument we used was not an established scale and lacked psychometric properties such as reliability and validity, however we used it in an exploratory fashion to gain insights into potential effects of existing domain knowledge on perceptions of AI and to inform future studies.

Participants who had less domain knowledge reported higher ratings in confidence in AI ( $t = 2.78, p = .007$ ), human-likeness ( $t = 2.96, p = .004$ ), and adequate justification ( $t = 2.83, p = .006$ ) than participants who had more domain knowledge. These results suggest that domain knowledge might be an important factor in how people perceive XAI, however it is difficult to draw conclusions without using an instrument that has good reliability and validity.

## Demographics

We conducted multiple 2(gender) x 2(condition) analyses to examine the effect of gender, however we found, no significant effect of gender in any metrics we used in the study, all  $p$ 's  $> .05$ .

We examined the effect of age by conducting correlational analyses, however, we found no significant correlations between age and study metrics, all  $p$ 's  $> .05$ .

## A.2 Study 2

In this section, we present the study materials used in the second study (Chapter 3) and the follow-up study (Chapter 4).

### A.2.1 Study Materials

#### Tooltip Descriptions

Part of conveying domain knowledge in the P2P lending app was to utilize tooltip descriptions for each factor shown on a borrower's profile. Participants in all conditions had access to these tooltips, however participants in the baseline condition did not see the **bold** portions of the descriptions. The **bold** text represented additional domain knowledge gained through expert interviews and decision ladders.

Table A.2: Tooltip Descriptions

Tooltip	Description
---------	-------------

Utilization	<p>The utilization rate is the ratio of credit balance to credit limit.  Total Utilization: The ratio of total balance to total available credit limit.  Credit Card Utilization: The ratio of total credit card balance to available credit card limit.  Loan Utilization: Current loan balances divided by original loan amounts.</p> <p><b>Having a high utilization rate (more than 30-35%) is a red flag that implies that the borrower is not managing debt well. If at any time this person's income is gone, they will be left with lots of debt, which increases the risk of default.</b></p>
Inquiries	<p>Inquiries occur when the person applies for credit (e.g. mortgage) and the lender checks the credit report of the person (also known as hard inquiries).  An inquiry does not necessarily mean that the person was able to get the credit product. It only tells that the person has applied for a credit product.</p> <p><b>Inquiries are very important. If a borrower has multiple inquiries, this may be an indication of two things:</b></p> <p><b>(1) They have debt to pay therefore they are seeking money, or</b>  <b>(2) Their credit applications are rejected.</b></p> <p><b>Both of these are red flags. Up to 2 inquiries per year is acceptable, however more than 2 inquiries is considered as a red flag.</b></p>
Late Payments	<p>If the borrower doesn't make credit payments on time (e.g. not paying the credit card bill before the due date), it is considered as a late payment (even if the borrower pays the bill eventually).</p> <p><b>Late payments are critical for both credit scores and lenders. Note that human lenders can be more flexible in terms of how they assess whether late payments constitute risk while credit scoring companies punish late payments heavily.</b></p> <p><b>From a lender's perspective, the type and the amount of late payments are more important than the number of late payments in the credit profile.</b></p>



Employment & Income	<p>Annual Income is used to calculate the borrower's debt to income ratio which is 'total monthly debt obligations (excluding mortgage) divided by monthly income'.</p> <p><b>As long as the debt the income ratio is acceptable (less than 35%), the income does not matter as much. The only case where it matters is that if the annual income is very high, a higher debt to income ratio might be acceptable as the amount of money left after paying debt will be high.</b></p> <p><b>In terms of employment length, the lenders will look for at least 2 years of continuous employment. This will ensure that the borrower's income is stable.</b></p>
Number of Accounts	<p>Number of accounts show the number of credit card accounts or loan accounts the borrower has. Number of accounts come into play when calculating the utilization rate.</p> <p><b>Since utilization rate is "balance divided by credit limit", having multiple accounts increases the credit limit and may result in a lower utilization rates.</b></p> <p><b>However, having more or less accounts does not necessarily how how well the borrower can manage debt.</b></p>
Account Balances	<p>Account balances show the amount of debt the person has. Account balances play a role in calculating the utilization rates which is "balance divided by credit limit".</p> <p><b>As long as the borrower has acceptable utilization rate and debt to income ratio, having high or low balances is not as important.</b></p>
Account Activity	<p>Account activity parameters shows the age of the most recently opened credit card or loan accounts.</p> <p><b>Every new credit account reduces the average age account which is important for credit scoring.</b></p> <p><b>Higher average account age is considered better as it shows that the borrower is able to manage debt longer (assuming they don't have late payments etc.)</b></p>

Account Age	<p>Account age parameters the age of the oldest credit card or loan accounts the borrower has.</p> <p><b>Account age is an indicator of the borrower’s history with credit. Ideally one should have at least 2 year on one of their accounts.</b></p> <p><b>Higher average account age is considered better as it shows that the borrower is able to manage debt longer (assuming they don’t have late payments etc.)</b></p>
Limits	<p>Credit limit refers to the maximum amount of credit the borrower is allowed to use and plays a role in calculating the utilization rates which is "balance divided by credit limit".</p> <p>Credit Card Limit: The total limit across all credit cards the borrower has.</p> <p>Loan Accounts Limit: The total limit across all loan accounts the borrower has.</p> <p><b>As long as the borrower has acceptable utilization rate and debt to income ratio, having high or low credit limits are not as important.</b></p>

In addition to financial factors, the AI Assistant and the Key Indicators Panel had their own tooltips to explain how to use them:

Table A.3: AI Assistant and Key Indicators Panel Descriptions

Panel	Description
-------	-------------

<p style="text-align: center;">AI Assistant</p>	<p>This section shows the analysis made by the AI Assistant.</p> <p>Prediction: AI Assistant makes a prediction of default chance (default means that the borrower will not be able to pay back the loan fully), ranging from 0% to 100%.</p> <p>The Most Important Factor: If the predicted default chance is 50% or high, the AI Assistant will pick the factor that increased the risk most as "the most important factor". If the predicted default chance is lower than 50%, the AI Assistant will pick the factor that decreased the risk most as "the most important factor".</p> <p>Contribution of Each Factor: These two graphs show the factors that increased or decreased the risk, according to AI Assistant. The AI Assistant makes its prediction by taking all these factors into account. You can hover over the bar charts to see how much they increased or decreased the default chance.</p> <p>It is recommended that you review both 'risk-decreasing factors' and 'risk-increasing factors' to better understand the AI Assistant.</p>
<p style="text-align: center;">Key Indicators Panel</p>	<p>This section shows the borrower's standings in key dimensions that were identified by talking to lending industry experts.</p> <p>In these graphs, 6 key indicators are presented. For each indicator, the green zone represents the 'acceptable' or 'safe' range, and the red zone represents the 'risky' range.</p> <p>If a value is in the green zone, it means that the borrower satisfies the criteria, and may be considered as low risk.</p> <p>If a value is in the red zone, the borrower fails to satisfy the criteria and may be considered as high risk.</p> <p>NOTE: This analysis is separate from the AI Assistant, however it is recommended to review both.</p> <p>There are multiple ways you can use these graphs:</p> <ol style="list-style-type: none"> <li>1. You can, at a quick glance, see where the borrower stands in key dimensions that are established in the lending industry.</li> <li>2. You can use these graphs in conjunction with the AI Assistant to evaluate the borrower and the loan.</li> </ol>

## A.2.2 Ethics Approval

Studies presented in Chapters 3 and 4 were approved the University of Waterloo Research Ethics Committee (Figure A.2).

## A.2.3 Recruitment

For recruitment, we used social media (r/uwaterloo Subreddit and a private Discord server for University of Waterloo students) and University of Waterloo mailing lists. The following recruiting material was posted on social media and shared with department coordinators to forward to students:

**Become an Investor on a Peer-to-peer platform. Evaluate Loans. Invest. Make Profit.**

Hello, my name is Murat. I am a PhD Candidate in the Department of Systems Design Engineering.

You are invited to participate in a 2-hour virtual study that investigates artificial intelligence-powered tools to help people make better financial decisions. The study consists of surveys, interviews, and interacting with an app that mimics a peer-to-peer lending platform where you will play the role of an investor, make investments in loans and maximize your profit. The study will take place on Microsoft Teams at an agreed upon date and time. In appreciation of your time, you will receive \$30.

For more information about the study or to participate, please contact Murat Dikmen at [murat.dikmen@uwaterloo.ca](mailto:murat.dikmen@uwaterloo.ca)

You can also read more about the study here: [Link to Information Letter]

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #42758). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or [ore-ceo@uwaterloo.ca](mailto:ore-ceo@uwaterloo.ca).

Thank you for considering participation in the study!

## A.2.4 Information Letter and Consent Form

Before each session, participants were asked to read the information letter and provide their verbal consent.

## UNIVERSITY OF WATERLOO

### Notification of Ethics Clearance to Conduct Research with Human Participants

---

Principal Investigator: Catherine Burns (Systems Design Engineering)

Student investigator: Murat Dikmen (Systems Design Engineering)

File #: 42758

Title: Investigation of Explainable Artificial Intelligence in a Peer-to-Peer Lending Context - Phase 2

---

The Human Research Ethics Committee is pleased to inform you this study has been reviewed and given ethics clearance.

**Initial Approval Date: 01/06/21 (m/d/y)**

University of Waterloo Research Ethics Committees are composed in accordance with, and carry out their functions and operate in a manner consistent with, the institution's guidelines for research with human participants, the Tri-Council Policy Statement for the Ethical Conduct for Research Involving Humans (TCPS, 2nd edition), International Conference on Harmonization: Good Clinical Practice (ICH-GCP), the Ontario Personal Health Information Protection Act (PHIPA), the applicable laws and regulations of the province of Ontario. Both Committees are registered with the U.S. Department of Health and Human Services under the Federal Wide Assurance, FWA00021410, and IRB registration number IRB00002419 (HREC) and IRB00007409 (CREC).

This study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

**Expiry Date: 01/07/22 (m/d/y)**

Multi-year research must be renewed at least once every 12 months unless a more frequent review has otherwise been specified. Studies will only be renewed if the renewal report is received and approved before the expiry date. Failure to submit renewal reports will result in the investigators being notified ethics clearance has been suspended and Research Finance being notified the ethics clearance is no longer valid.

Level of review: Delegated Review

Signed on behalf of the Human Research Ethics Committee



This above named study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Documents reviewed and received ethics clearance for use in the study and/or received for information:

file: PostStudyInterviewGuide\_Version1\_20201120.pdf

file: WebsiteAndTask\_Version1\_20201120.pdf

Figure A.2: Research Ethics Committee approval for Study II and III.

## **Information Letter**

### **Study Overview**

You are invited to participate in a research study conducted by Murat Dikmen, under the supervision of Catherine Burns at the Systems Design Engineering Department. The objective of the research study is to understand how people may use an artificial intelligence assistant when making investments on a peer-to-peer lending website. The study is for a PhD thesis.

### **What You Will Be Asked to Do**

If you decide to volunteer, you will be asked to complete a 2-hour long online study that involves surveys, interviews, and interacting with a website. The study will be in virtual format. You will be invited to a virtual meeting on Microsoft Teams to facilitate the study. The study will consist of:

Consent: You will be asked to provide verbal consent to participate in the study.

Pre-study survey: 40 questions about your background, knowledge and experience related to finance, lending, and artificial intelligence. ( 20 minutes)

Task Orientation: Clarify questions as well as introduce the peer-to-peer lending app and provide training ( 20 minutes)

Main task: Interacting with an app that mimics a peer-to-peer lending website without real money or transactions. The task involves playing the role of an investor, evaluating the loan requests, and investing in loans to make profit. ( 45 minutes)

Post-study survey: 20 questions about your experience with the app. ( 10 minutes)

Post-study interview: In-depth interview about your experience with the app. ( 25 minutes)

Payment: Payment

The study will involve approximately 60 participants, divided into 2 groups. Groups will differ in the data and the visualizations available in the app part of the study. If you participate in the study, you will be randomly assigned to one of the 2 groups.

### **Virtual Meeting**

With your permission, the interviews will be audio recorded using OBS (Open Broadcaster Software) to facilitate data analysis.

We do advise participants to refrain from unnecessarily disclosing personal information, make use of Teams' backgrounds feature to protect any information that could be visible around them, and wear headphones or isolate themselves in a separate room to prevent other people from overhearing the conversation.

### **Your Participation**

Participation in this study is voluntary. You may decline to answer any questions that you do not wish to answer and you can withdraw your participation at any time by not submitting your responses. If you chose to withdraw from the study at any point, please contact the student investigator in Microsoft Teams or at [murat.dikmen@uwaterlo.ca](mailto:murat.dikmen@uwaterlo.ca).

**Remuneration** In appreciation of your time, you will be paid \$30. The payment will be made via e-transfer or any appropriate method if e-transfer is not an option. You will be asked to sign Acknowledgement of Receipt of Remuneration and Self-Declared Income form when you receive the payment.

### **Risks**

There are no known or anticipated risks from participating in this study.

### **Personal Benefits of the Study**

There are no direct benefits to the participants of this study. However, the results obtained from this study will contribute to the society by advancing our knowledge about building more user-friendly artificial intelligence applications.

### **Data Confidentiality and Security**

When information is transmitted over the internet, confidentiality cannot be guaranteed. University of Waterloo practices are to turn off functions that collect machine identifiers such as IP addresses. The host of the system collecting the data such as Qualtrics or Mongo Inc. may collect this information (such as your IP address) without our knowledge and make this accessible to us. We will not use or save this information without your consent. Furthermore, the list that links participant names with participant IDs will be kept separately from the study data to ensure the study data does not have any personal identifiers.

The data, with no personal identifiers, collected from this study will be maintained on cloud platforms. During the data collection and analysis phases, survey data will be stored on the Qualtrics platform, log data (from the app)

will be stored on a secure server in the US (provided by the database service, Mongo Inc.), and audio recordings will be stored in OneDrive. After the completion of the study, all of the data will be moved to OneDrive storage provided by UW and maintained for at least 10 years.

We may use anonymous quotes from the interviews in future publications.

Data from this study may be deposited in an online public repository/database. Data will be de-identified prior to submission to the repository/database, and in the case of audio recordings, only transcripts will be submitted. This process is integral to the research process as it allows other researchers to verify results and avoid duplicating research.

If you wish to withdraw your data at any time, please contact [murat.dikmen@uwaterloo.ca](mailto:murat.dikmen@uwaterloo.ca) or [catherine.burns@uwaterloo.ca](mailto:catherine.burns@uwaterloo.ca). Note that it is not possible to withdraw your data once the results of this study are published in a publication (e.g. a journal article) or a PhD dissertation.

If you are interested in receiving more information regarding this study or a summary of its results, or if you have any questions or concerns, please contact us at the email addresses listed at the bottom of this page. The results of this study will be available in February 2021.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #42758). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or [ore-ceo@uwaterloo.ca](mailto:ore-ceo@uwaterloo.ca).

Thank you for considering participation in this study.

## **Consent Form**

By providing your consent, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities.

I have had the opportunity to contact researchers and ask any questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted. I am aware that I may withdraw from the study without penalty at any time.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #42758). If you have questions



for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

For all other questions contact Murat Dikmen at murat.dikmen@uwaterloo.ca. You may also contact Professor Catherine Burns at 519-888-4567 Ext. 33903., catherine.burns@uwaterloo.ca

I am aware that the virtual meeting will be audio recorded to ensure accurate transcription and analysis.

I give permission for the use of anonymous quotations in any thesis or publication that comes from this research.

I agree of my own free will to participate in the study.

## A.2.5 Procedure

### Remote Study

The studies were conducted remotely in a Microsoft Teams meeting. The P2P lending website was developed such that it could be accessed by participants using a link and a password on their own computer. Throughout the study, the researcher was also able to log in to the website using the participant's unique number to see their investments. Both the researcher and the participant were always present in the meeting, however participants were allowed to turn off their camera, especially when they were working on a task (filling out questionnaires or interacting with the P2P website). We did not use screen sharing because of the privacy concerns as participants could inadvertently reveal personal data on their computer. For the same reason, we did not turn on video recording as almost all participants participated to the study from their home due to COVID-19.

Except from the interview at the end of the study, all questions were asked in a survey format (pre-study questionnaire and post-study questionnaire). The link to these questionnaire were shared using the chat functionality of Microsoft Teams. The procedure was the same for both studies presented in Chapters 3 and 4.

### Procedure

After participants provided their verbal consent, the researcher started the audio recording and asked for the consent again to ensure the consent is recorded. Next, the researcher assigned a participant ID and shared this along with a link to the pre-study questionnaire in

the chat. The participant could visit the questionnaire and complete the questions on their own. They were also able to ask questions to the researcher if they required clarification. This pre-study questionnaire included the following questions:

What is your age? (open-ended)

What is your gender?

- Male
- Female
- Prefer to self-identify (open-ended)
- Prefer not to answer

What is your major? (open-ended)

Have you taken any finance related courses at UW?

- Yes (please indicate which finance related course(s) you have taken)
- No

Next, the Loan Knowledge Quiz described earlier in [A.1.1](#) and risk attitudes questionnaires (DOSPERT and SIRI) were presented. Finally, a number of questions were asked about participant's experience with AI:

Please indicate your agreement with the following statements. (from **Strongly Disagree** to **Strongly Agree** on a 7-point scale)

- I am confident using computers.
- I can make use of computer programming to solve a problem.
- I understand how Amazon recommends products for me to purchase.
- I understand how my email provider's spam filter works.
- I understand how self-driving cars drive on their own.
- I understand how autopilot works on an airplane.
- I understand how YouTube recommends videos for me to watch.
- I understand how Facebook Newsfeed chooses which posts to show before everything else.

Have you had any experience with an AI program that makes loan approval / rejection decisions?

- Yes
- No

[If Yes] Please describe your experience with the AI program that makes loan approval / rejection decisions. (open-ended)

How much do you know how Artificial Intelligence algorithms work? (from **Nothing** to **A lot** on a 5-point scale)

How much programming knowledge do you have? (from **Nothing** to **A lot** on a 5-point scale)

How much knowledge of computer algorithms do you have? (from **Nothing** to **A lot** on a 5-point scale)

After completing the pre-study questionnaire, the link to the P2P lending website was shared in the chat. Participants were asked to input their participant ID on the landing page. Upon entering their ID, the app created a new session by randomizing the order of loan requests. The main functionality of website is described in Chapter 3. Additionally, the website had the following properties:

- After participants entered their participant ID, they were shown a “Home” screen (Figures A.3, A.4, and A.5).
- Participants could browse the available loans and click on any one of them to see the borrower’s financial profile. The order of the loans was randomized, but participants were not given any instructions regarding which ones to look at. They were free to check all or some the loan requests in any order they wanted.
- Participants could, at any time, navigate to different sections of the app: Home, Loans, and Portfolio (Figure A.6).
- When participants made an investment in a loan, the website user interface was updated to show the current balance and the invested amount. Additionally, the portfolio visualizations were also automatically updated.
- Actions such as investments were stored in a database. This allowed the researcher to log in to the app using the participant ID to see the current status of investments.

- When participants made an investment in a loan, they could “reset” the investment by clicking on a button on the user interface which set the invested amount to 0 for that loan. They could then reinvest if they preferred.

Participants completed a training session with five loans. When they were done with their investments (defined as “The participant is happy with their portfolio, they invested in the loans they wanted to invest in, and they don’t want to make further changes”), they clicked on a “Submit” button to finish the session. Upon completing the training session, the app automatically pulled the set of loans to use in the main task, reset balances and investments, adjusted allocated funds, and redirected the user to the home page. At this point, they told the researcher that they have finished the training session. No feedback was given about their performance during the training task, however the researcher clarified any questions they had, and then they proceeded with the main task. This time, when they “Submit” their investments, the website showed a link. Upon clicking on the link, participants were taken to the Qualtrics platform for the post-study questionnaire, which included the following questions:

From 1 (Too little) to 5 (Too much), overall, when evaluating loans, the information presented on the website was:

From 1 (None at all) to 5 (A great deal), how much did the Machine Learning Assistant help you when making investments? From 1 (None at all) to 5 (A great deal), how much did the explanations and contribution of each factor provided by the Machine Learning Assistant help you when making investments?

From 1 (Never) to 5 (Always), how frequently did the consult the Machine Learning Assistant when evaluating the risk of a loan?

Please rate the following statement: I can predict how the Machine Learning Assistant will behave when making loan decisions. (from **Strongly Disagree** to **Strongly Agree** on a 7-point scale)

Explanation quality questions, described in Section [3.2.3](#)

A checklist for Trust between People and Automation [\[92\]](#)

In predicting the default risk, from 0% to 100%, how accurate do you think the Machine Learning Assistant was in this study? (a slider was used)

Overall, how satisfied are you with the tools that we provided to help you make better investments? (from **Extremely Dissatisfied** to **Extremely Satisfied** on a 5-point scale)

From 1 (None at all) to 5 (A great deal), how helpful were the tooltips?

When making your final decision about whether to invest in a loan, how much did you rely on the Machine Learning Assistant's predictions? (from 1-None at all to 5-A great deal)

How important is Income in assessing the creditworthiness of a borrower? (from 1-Not at all important) to 5-Extremely important)

How important is Credit Card Account Balance in assessing the creditworthiness of a borrower? (from 1-Not at all important) to 5-Extremely important)

How important is Total Utilization Rate in assessing the creditworthiness of a borrower? (from 1-Not at all important) to 5-Extremely important)

How important is Debt-to-Income Ratio in assessing the creditworthiness of a borrower? (from 1-Not at all important) to 5-Extremely important)

How important is Number of Inquiries in assessing the creditworthiness of a borrower? (from 1-Not at all important) to 5-Extremely important)

How important is Late Payments in assessing the creditworthiness of a borrower? (from 1-Not at all important) to 5-Extremely important)

How important is Months Since Oldest Credit Card Opened in assessing the creditworthiness of a borrower? (from 1-Not at all important) to 5-Extremely important)

How important is Employment Length in assessing the creditworthiness of a borrower? (from 1-Not at all important) to 5-Extremely important)

From 1 (Extremely easy) to 7 (Extremely difficult), how difficult was the study? (from 1-Not at all important) to 5-Extremely important)

From 1 (Extremely unclear) to 7 (Extremely clear), how clear were the instructions? (from 1-Not at all important) to 5-Extremely important)

After the participant filled out the questionnaire, the researcher conducted a semi-structured interview, and the following questions were used as a guide:

- How confident are you with your decisions?
- What was the most challenging aspect of the task?
- How did you make your decisions? Which factors did you take into account?

- How did you use machine learning predictions, if at all?
- What did you like about AI? What did you dislike about AI
- What are your thoughts on the AI Assistant? Do you have any suggestions to improve it?
- Were there cases where your own evaluation and the machine learning prediction didn't match? How did you resolve those cases?
- Did you use the key indicators chart? How?
- Did you follow the industry guidelines?
- Were there cases where the Machine Learning model and Lending Industry Guidelines didn't match? What did you do?
- Did you use Portfolio page? How? What do you think about the portfolio page?
- If this was a real website with real money involved, would you use the same strategy? What would you need to know to invest more comfortably?

In addition to the question-answer setup, during the interviews, we went over some of the loan decisions participants made, and asked questions about how they made their assessment for each case. We specifically focused on cases where participants lost money (the borrower defaulted), and cases where the AI made incorrect risk predictions. To achieve this, both the researcher and the participant logged in to the website using the participant ID, and the researcher asked the participant to click on loans that were of interest. The participant was asked to articulate on why they made an investment (or not), and how they made their assessments. This was repeated for at least five loans (loans where the AI failed to predict the default risk correctly), and for loans where the researcher observed an unusual activity (for example, not investing in the safest loan, or investing fully in the most risky loan). During the session, we also gave feedback about the performance of the participant and showed where they made a profit or a loss. After the interview session, the participant was paid via online money transfer, and a link to the following feedback letter was shared in the chat:

### **Thank You!**

I would like to thank you for your participation in this study. As a reminder, the purpose of this study was to understand how people interact with artificial

intelligence and perceive explanations of AI predictions in a peer-to-peer lending environment. This study will help inform the future research on making artificial intelligence more understandable and trustworthy.

The study is designed such that half of the participants saw only the AI Assistant tool and the other half saw both the AI Assistant tool and the Key Indicators graph. We hypothesized that participants who had access to both visualizations will make better investment decisions (resulting in higher profit) and will perceive the AI Assistant differently than participants who had access to only the AI Assistant.

### **Data Confidentiality and Security**

As a reminder, when information is transmitted over the internet, confidentiality cannot be guaranteed. University of Waterloo practices are to turn off functions that collect machine identifiers such as IP addresses. The host of the system collecting the data such as Qualtrics or Mongo Inc. may collect this information (such as your IP address) without our knowledge and make this accessible to us. We will not use or save this information without your consent. Furthermore, the list that links participant names with participant IDs will be kept separately from the study data to ensure the study data does not have any personal identifiers.

The data, with no personal identifiers, collected from this study will be maintained on cloud platforms. During the data collection and analysis phases, survey data will be stored on the Qualtrics platform, log data (from the app) will be stored on a secure server in the US (provided by the database service, Mongo Inc.), and audio recordings will be stored in OneDrive. After the completion of the study, all of the data will be moved to OneDrive storage provided by UW and maintained for at least 10 years.

As we mentioned in the information letter, we may use anonymous quotes from the interviews in future publications.

Data from this study may be deposited in an online public repository/database. Data will be de-identified prior to submission to the repository/database, and in the case of audio recordings, only transcripts will be submitted. This process is integral to the research process as it allows other researchers to verify results and avoid duplicating research.

If you wish to withdraw your data at any time, please contact [murat.dikmen@uwaterloo.ca](mailto:murat.dikmen@uwaterloo.ca) or [catherine.burns@uwaterloo.ca](mailto:catherine.burns@uwaterloo.ca). Note that it is

not possible to withdraw your data once the results of this study are published in a publication (e.g. a journal article) or a PhD dissertation.

If you are interested in receiving more information regarding this study or a summary of its results, or if you have any questions or concerns, please contact us at the email addresses listed at the bottom of this page. The results of this study will be available in February 2021.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #42758). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

## A.2.6 Data Analysis

In this section, we present additional data analysis that was not reported in Chapters 3 and 4.

There were a number of subjective measures we collected. These include perceived helpfulness of AI assistant (“How much did the AI assistant help you in making your assessments?”), perceived helpfulness of explanations (“How much did explanations help you in making your assessments?”), frequency of using AI assistant (“How frequently did you consult the AI assistant?”), AI predictability (“How well can you predict how AI assistant will assess future borrowers?”), and the usefulness of tooltips. These questions were asked on 5-point scales. ANOVAs and Kruskal-Wallis tests showed no significant differences between conditions, all  $p$ 's  $> .05$ .

We also asked about the importance of several financial factors (e.g., “How importance is income in determining the risk of a borrower”) after the experimental sessions. There were no differences in income, debt-to-income ratio, account age, and employment length. There was a difference in card balance,  $H(2) = 7.21$ ,  $p = .03$ , but post hoc tests were not significant. The importance of credit utilization was also different,  $H(2) = 6.77$ ,  $p = .03$ . Participants in the *embedded* condition ( $M = 3.9$ ,  $SD = 1.02$ ) reported higher ratings than participants in the *baseline* condition ( $M = 3.1$ ,  $SD = 1.02$ ),  $p = .05$ . There was a marginally significant difference in the importance of inquiries,  $H(2) = 5.63$ ,  $p = .06$ , however post hoc tests were not significant. Similarly, there was an overall difference in the importance of late payments,  $H(2) = 7.01$ ,  $p = .03$ , but post hoc tests were not significant.



## Interviews

In this section, we want to share additional insights gained from the interviews conducted after each study session.

Regarding the Key Indicators panel, (Figure 3.3) participants generally expressed that it was useful, however, some participants said that they did not consider it much. It looks like the visual form was handy for some, as one participant explained: *“I was just using the colors to help me. It was really easy to understand and read all of these things. It was just quickly open up the loan, read the financial details really quickly, and then it takes maybe 2 seconds and you’ll understand where everything is at. [P10]”*.

The presence of domain knowledge had some effects on the assessment process. Most participants said that they used it in conjunction with the AI: *“I used both sides, but mainly the stuff on the right [key indicators panel] to determine if I wanted to give a loan and then on the left [AI assistant] based on the default risk and the factors, I decided how much I would want to give them access.[P10]”*. In this case, it appears that the domain knowledge was used to understand the risk (i.e., whether or not the borrower will default), and the AI was used to calibrate the action (i.e., how much to invest). Overall, the presence of domain knowledge helped them make better assessments: *“I used it [key indicators panel] because the data was presented for someone like me who doesn’t have a lot of financial knowledge. Sometimes it’s hard to interpret what do these values. So by looking at the right [key indicators panel] it just gave me an idea.” [P31]*. The mismatch between AI and domain knowledge was also explicitly mentioned by some participants: *“I found that interesting disconnect between what a financial expert would say, and what the model says.” [P47]*

Regardless of the presence of domain knowledge, some participants were still confused about how to interpret the financial factors, which is captured well by one participant: *“Like I know someone made 400K and I was like oh yeah, they can make enough money to pay me back, but also they have so many credit cards right? And I don’t know. I don’t know how to interpret income. [P50]”*.

We also observed a range of strategies used when evaluating the explanations. Some participants expressed strategies such as comparing the number of positive and negative factors, the magnitude of positive and negative factors, and so on. For example, one participant said: *“I simply looked at the visualization and from that I sort of decided if there was more green [factors that decreased the risk] than red [factors that increased the risk], but at the same time I also looked at what each green bar represented and how much. I think that matters to me.”*. Another participant was concerned with magnitudes: *“I*

*looked at the ones that appeared to be larger bars. The ones that were lower down on each chart were not significant so I kind of just skimmed over there.” [P7].* In general, looking at the number of factors and their magnitudes was a good starting strategy. However, without understanding what they mean in the context of the borrower, these approaches could easily lead to premature, and often incorrect assessments, especially if the AI was incorrect and the explanations were misleading.

The explanations could also create confusion. Some participants were confused and were not able to make sense of the explanations, as illustrated by this comment: *“I was getting frustrated because I like this information [referring to explanations], but I just didn’t know what would give me the best outcome still. (P23)”*.

### **A.3 Impact of COVID-19**

COVID-19 had some impact on this work. Initially, we designed the studies in Chapters 3 and 4 as in-person studies, however the pandemic started while we were developing the study materials. We had to switch to remote studies, and that brought some additional challenges. For example, we had to switch to a web-based platform, including automated data collection (as opposed to screen recording in a lab environment). This approach resulted in significant delays and limited visibility into how participants behaved during the experiments.

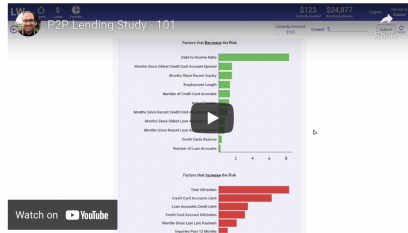
The global economic impact of COVID-19 also might have affected how participants behaved during the experiments, as financial decision-making was at the core of the studies. Credit industry was particularly sensitive to the economic impact of the pandemic. As the uncertainty about the future was extremely high during the first few months of the pandemic, we have seen that financial institutions took extra care when it comes to allocating funds to credits, and adjusting the lending policies. We have also seen some of the largest P2P companies switching from consumer lending to business lending because of the economic risks. While we did not observe the impact of these developments during the interviews, it is possible that participants’ risk perceptions might have been implicitly affected by the ongoing pandemic and economic uncertainty.

### **A.4 Reflections**

As the main author of this dissertation, I’d like to take this opportunity to reflect on the past four years, and share some of the learnings that might help future students.

## Welcome to Lending World

Play the role of an Investor. Assess the loans. Invest.  
Make Profit!



Watch on YouTube

### Study In a Nutshell

In this study, we ask you to play the role of an investor on this peer-to-peer lending website. A peer-to-peer lending platform provides a convenient way for individuals and small businesses to access loans which are funded by other individuals, also known as "investors". The borrowers can apply for loans and provide their financial information. If a loan application is approved by the website, the borrower's loan request is listed on the website, and the website asks "investors" to fund/invest in these loans at the set interest rate. Investor can invest in loans and as the borrowers pay back the loan, the investors earn interest.

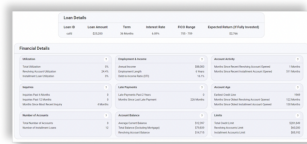
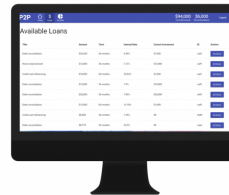
### Your Task

In this study, you will play the role of an "investor" and your job is to use the amount of money provided to you to make investments in loans that are approved by the website but looking for investors. \*Note that this is not real money or the loans you will see are not real loans. However, we still ask you to approach this as if you were a real investor who wants to maximize their profit on this website.

## How It Works

### Browse Loan Listings.

Browse and review the loans that you can invest in as well as see which ones you have already invested in. During the training phase (in this phase, there will be 5 loans listed. During the main study phase, there will be 20 loans listed.



### See Loan Details.

See all the details about the loan, including the amount, term, and the interest rate. Review the borrower's credit report and financial details, from late payments to credit utilization, and other credit score components. You can also see the expected return if you invest fully in that loan.

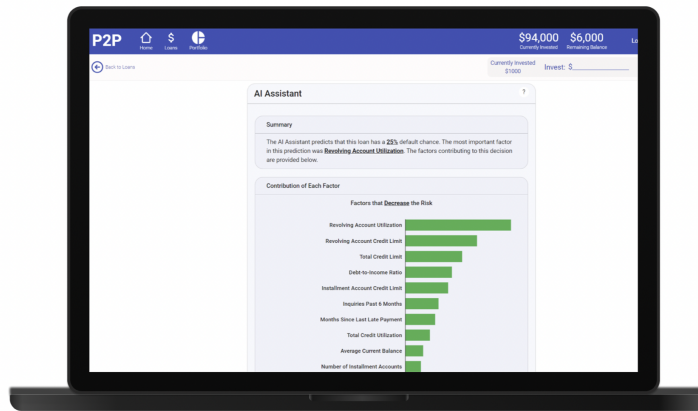
Figure A.3: Home page.

### Use Powerful Artificial Intelligence (AI) Tools.

When evaluating the loan and the borrower, take advantage of the **Machine Learning Assistant**.

The Machine Learning Assistant predicts whether the borrower will be able to pay back the loan as well as provide detailed graphs and information that explain "Why and How" the **Machine Learning Assistant** made that prediction.

The Machine Learning Assistant was built using Machine Learning and was trained on hundreds of thousands of loan transactions. Of course, as with most AI programs, the Machine Learning Assistant is not always accurate. It is your job and your responsibility to evaluate the suggestion provided by the Machine Learning Assistant when making your investment decisions.



### Make Investments.

Using the amount of cash allocated to you, make investments in loans for profit. Note that the money is virtual money, however we want you to experience how it feels to be an investor.

By investing in loans, you take a risk. If the borrower pays back, you will earn an interest at the interest rate shown on each loan. If the loan defaults or is not paid back, **you may lose money**. You can choose how much to invest in a loan. You don't need to fully fund a loan, you can partially invest. This allows you to diversify your portfolio. Remember your goal is to earn as much money as possible and invest in loans that will be (hopefully) paid back so that you can earn the interest.



### Review Your Portfolio.

The portfolio page allows you to easily view your current investments and present an overview of your investment portfolio. You can find the distribution of your investments, expected returns, and more.

Make sure to hover over the graphs to see the detailed information.



Figure A.4: Home page (cont'd).

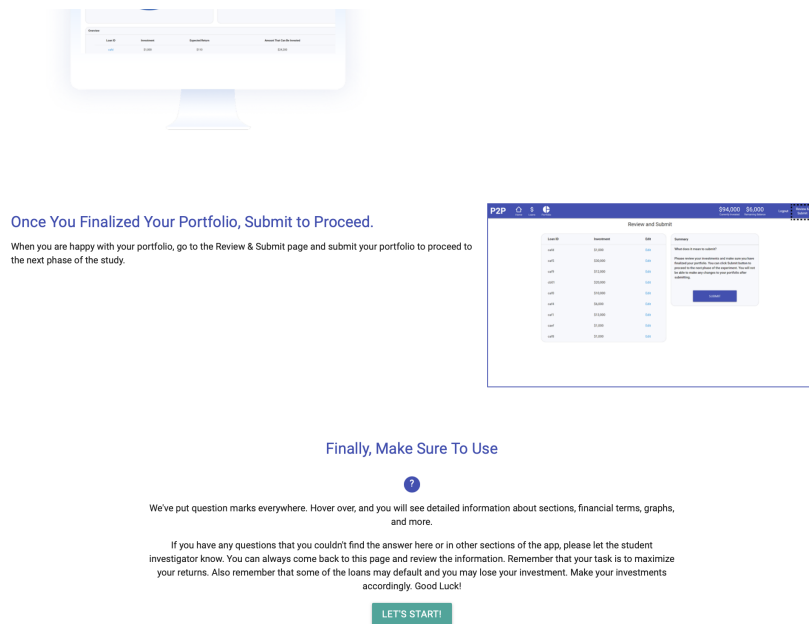


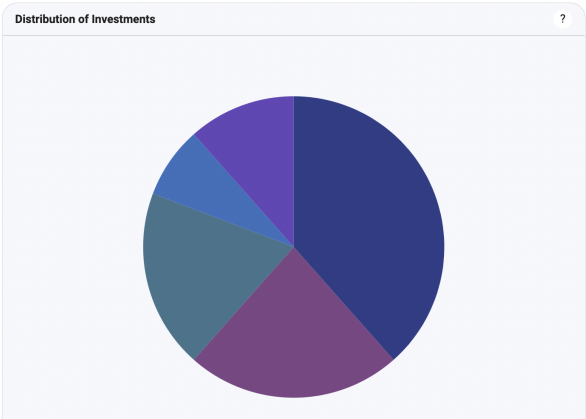
Figure A.5: Home page (cont'd).

One of the challenges I encountered initially was the lack of research, tools, approaches and ideas in this field. XAI is a fairly young field. This means that while the opportunities were abundant, finding the right problem to focus could be difficult as there were a lot of unknowns. On top of that, the lack of established research resulted in many decisions being made with limited information. On the bright side, this meant that a lot experimentation could be done, however as a PhD student, making progress was as important as experimenting. I believe a promising approach to deal with this problem is iterating early and often.

I also learned to appreciate the value of engaging in diverse research fields throughout my PhD journey. I was lucky to get involved in a range of research projects including self-driving cars, navy automation, task interruptions, augmented/virtual reality, unmanned aerial vehicles, and AI. Having the opportunity to observe automation problems and opportunities in diverse fields helped me significantly in developing my research approach, consolidating my ideas and gaining an interdisciplinary perspective as there are many commonalities in the problems faced in seemingly unrelated fields.

Portfolio

Expected Net Return ?  
**\$2,257 (17.36%)**



Overview ?

Loan ID	Investment	Expected Return	Amount That Can Be Invested
<a href="#">9a43</a>	\$5,000	\$1,116	\$5,000
<a href="#">9a46</a>	\$3,000	\$425	\$7,000
<a href="#">9a44</a>	\$2,500	\$293	\$22,500
<a href="#">9a47</a>	\$1,000	\$187	\$9,400
<a href="#">9a45</a>	\$1,500	\$236	\$5,500

Figure A.6: Portfolio page.