# Developing Explainable AI-Assisted Decision-Making Systems

Thao Le
ORCID: 0000-0002-7453-6778

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

School of Computing and Information Systems
THE UNIVERSITY OF MELBOURNE

December 2024

ORCID: 0000-0002-7453-6778

# Abstract

Prior research on AI-assisted human decision-making has explored several different explainable AI (XAI) approaches. A common approach is to provide AI recommendations along with uncertainty measures and explanations. However, this approach can lead to over-reliance on AI, which is a significant concern when the AI is wrong. On the other hand, under-reliance can occur when the human decision-maker does not trust the AI when it is correct. Moreover, human decision-makers only have two options: to follow the AI recommendation or not, in which the latter they often go with their own initial thought. Therefore, challenges remain in building effective AI-assisted systems that can reduce reliance on AI and do not limit the control of human decision-makers.

The goal of my PhD research is to develop these systems by explaining the uncertainty and building a more reliable decision-making approach. I begin by proposing a method for explaining model uncertainty to promote trust and improve end-user understanding. I then define a new decision-making approach based on the Evaluative AI framework. Through human-subject experiments, the new approach has shown promise in reducing over-reliance and allowing users to make better decisions, although there is a small increase in under-reliance compared to the traditional recommendation-driven approach. Finally, I demonstrate the application of the new decision-making approach in supporting skin cancer diagnosis by extending the Weight of Evidence framework to image datasets. The proposed approach is evaluated by conducting experiments with individuals experienced in skin cancer diagnosis. The results show that recommendation-driven and hypothesis-driven approaches have their own advantages and disadvantages, and suggest future research in combining the strengths of both approaches.

# Declaration

I, Thao Le, declare that this thesis titled, "Developing Explainable AI-Assisted Decision-Making Systems" and the work presented in it are my own. I confirm that:

- The thesis comprises only my original work towards the degree of Doctor of Philosophy, except where indicated in the preface;
- due acknowledgement has been made in the text to all other material used; and
- the thesis is fewer than 100,000 limit in length, exclusive of tables, maps, bibliographies and appendices as approved by the Research Higher Degrees Committee.

Signed: Thao Le

Date: 18/12/2024

# Acknowledgements

I have the privilege to pursue my PhD studies at the University of Melbourne and have been surrounded by many amazing and brilliant people. My love for science started in early childhood. Despite having some passion and small achievements in Chemistry and Biology at a young age, I could not become a medical doctor. But "life finds a way", as I am writing this thesis now, I believe the title "Dr." is not far away.

To my supervisors, Prof. Tim Miller, Prof. Liz Sonenberg and Dr. Ronal Singh, I am grateful for your continued guidance and support. You are the role models in academia that I never had. Although there were many uncertainties and challenges, I am thankful for your patience in supporting my research career. And I hope that this PhD thesis might be able to *explain some of those uncertainties*. I could not have hoped for a more supportive team, and I must say, the time that I spent doing PhD is a very happy time in my life so far. Tim, you are the first person who taught me how to do research properly. Even though you already moved to UQ, your support for my PhD studies is tremendous. I am able to finish this thesis in time thanks to your guidance. Liz, your sharp mind always gives me valuable feedback on my research. Although it is still an NP-hard problem to find an available time on your calendar, you would always promptly respond to my emails and make time when I need your help. Thank you for acting as my primary supervisor when Tim left Unimelb. Ronal, your kindness and calmness help me when I am stressed; and I have enjoyed working with you on several projects over the past few years.

To my colleagues and mates in the AI group, Abeer Alshehri, Archana Vadakattu (Archie), Anubhav Singh, Lyndon Benke, Ruihan Zhang, Chao Lei, Michelle Blom, Emma Baillie, Guang Hu, Chenyuan Zhang, Tingxuan Wang, Markus Hiller, Viktoria Schram, Joshua Newn, Hanan Alsouly, Rinu Sebastian. Thanks for our coffee chats, lunches, and

of course, after-work drinks and games in the pub, you guys keep me sane. I will very much miss the banter and how boardgames can turn into wonderful chaos. Also, thank some of you for lending me your strong hands when I move home.

In addition, my PhD journey has been enriched by many more collaborations and mentorships. I would like to thank the teaching team of AI Planning for Autonomy. It was a pleasure to work with you, Prof. Nir Lipovetzky, Prof. Adrian Pearce and other staff on this subject. I am grateful for the opportunity to teach and learn from you. I want to further express my gratitude to Prof. Peter Soyer (UQ) for providing his invaluable insights on my skin cancer project. I thank Prof. Matthew Gombolay (Georgia Tech) for being my mentor at the AAAI doctoral consortium. I also want to thank Prof. Michael Kirley and Kirsten Eccles for organising the Doctoral Academy at Melbourne Centre for Data Science, where I had a great time meeting fellow PhD students and learning from various sessions we attended together.

I have been fortunate to go on various trips both in Australia and overseas thanks to financial support from various sources, especially from my supervisors. I thank CIS Unimelb for offering travel scholarships. I would also like to thank Google for providing me with a conference grant and flying me to Sydney to attend the Google Research Day. I would like to thank Prof. Iyad Rahwan for hosting me at the Center for Human and Machine (CHM), Max Plank Institute for Human Development. I thank everyone at the centre for their hospitality, and thank Joanna and Kerstin for helping me during my visit in Berlin.

Last but not least, to my parents, thank you for your constant love and support. It has been quite a journey raising the first PhD in the family. I am thankful for the ups and downs along the way – they have taught me a lot about resilience. Life would not be the same without a few twists and turns, and I am looking forward to the next colourful chapter of my life.

# Preface

This thesis has been primarily conducted in the School of Computing and Information Systems, Faculty of Engineering and Information Technology, The University of Melbourne. Below is the list of publications and manuscripts that are included in this thesis. For all papers, I am the primary author who has contributed more than 50% of the work, including but not limited to developing the methods, implementation, designing experiments, data collection, data analysis and writing. My co-authors have provided feedback on the research and writing. I note that I use "we" in Chapter 2, 3, 4, and 5 in recognition of the contributions of my co-authors, resulting in publications and manuscripts listed below.

**Conference Proceedings**

- **[C1] (AAAI23 Main Track) [127]** Thao Le, Tim Miller, Ronal Singh, Liz Sonenberg. "Explaining Model Confidence Using Counterfactuals." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 10, pp. 11856-11864. 2023.

- **[C2] (ECAI24 Main Track) [128]** Thao Le, Tim Miller, Liz Sonenberg, Ronal Singh. "Towards the New XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence". In *In 27th European Conference on Artificial Intelligence*, vol. 392, pp. 850-857. 2024.

**Workshops**

- **[W1] (XAI-IJCAI22)** Thao Le, Tim Miller, Ronal Singh, Liz Sonenberg, 2022. "Improving Model Understanding and Trust with Counterfactual Explanations of Model

Confidence." In *International Joint Conference on Artificial Intelligence - Workshop on Explainable Artificial Intelligence*. 2022. *This is an earlier version of [C1].*

**Doctoral Consortium**

- **[DC1] (AAAI-DC23) [126]** Thao Le. "Explaining the Uncertainty in AI-Assisted Decision Making." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, pp. 16119-16120. 2023.

**Working Papers (Under review)**

- **[P1] [130]** Thao Le, Tim Miller, Ruihan Zhang, Liz Sonenberg, Ronal Singh. "Visual Evaluative AI: A Hypothesis-Driven Tool with Concept-Based Explanations and Weight of Evidence."
- **[P2] [129]** Thao Le, Tim Miller, Peter Soyer, Ronal Singh, Liz Sonenberg. "EvaSkan: An Evaluative Skin Cancer Tool for Decision Support."

**Acknowledgement of Funding** I acknowledge the following sources of funding:

- Melbourne Research Scholarship
- Melbourne Centre for Data Science (MCDS) Top Up Scholarship
- Google Conference Grant
- FEIT-CIS Conference Travel Grant
- My supervisors' research funding

# Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **BDI** | Belief-Desire-Intention |
| **CBM** | Concept Bottleneck Models |
| **CBR** | Case-Based Reasoning |
| **CNN** | Convolutional Neural Network |
| **CEM** | Concept Embedding Model |
| **CF** | Counterfactual |
| **CW** | Concept Whitening |
| **DA** | Decision Aid |
| **DNN** | Deep Neural Network |
| **DST** | Decision Support Tool |
| **EvaSkan** | Evaluative Skin Cancer |
| **FC** | Fully Connected |
| **GNB** | Gaussian Naive Bayes |
| **ICE** | Invertible Concept-Based Explanation |
| **LIME** | Local Interpretable Model-Agnostic Explanations |
| **ML** | Machine Learning |
| **NDM** | Naturalistic Decision Making |
| **NLN** | Nearest-Like-Neighbour |
| **NMF** | Non-Negative Matrix Factorization |
| **NUN** | Nearest-Unlike-Neighbour |
| **OOD** | Out-of-Distribution |
| **PCBM** | Post-Hoc Concept Bottleneck Model |

| | |
|---|---|
| **PCA** | Principal Component Analysis |
| **PDP** | Partial Dependence Plot |
| **ResNet** | Residual Neural Network |
| **SHAP** | SHapley Additive exPlanations |
| **TCAV** | Testing with Concept Activation Vectors |
| **XAI** | Explainable Artificial Intelligence |
| **WoE** | Weight of Evidence |

# Chapter 1

# Introduction

AI-ASSISTED decision-support systems have become increasingly popular in various domains, such as healthcare, finance, and transportation. Machine learning models are used to provide recommendations to help users make better decisions. However, these models are often treated as black boxes, making it difficult for users to understand how the model works and why it makes certain recommendations. Furthermore, another issue is that even if the model can have a very high accuracy, it might use the wrong features (or evidence) to make the prediction. This is referred to *Clever Hans phenomenon* [175] in psychology. Therefore, explainable AI (XAI) has shown promise in improving trust and understanding in machine learning models. A common XAI approach is to provide explanations of the model's predictions, which can help users understand why the model makes that prediction. Moreover, uncertainty measures can be provided to indicate the model's confidence in its predictions. This approach is called *recommendation-driven AI*, where the model provides a recommendation and additional information about that recommendation. Human decision-makers can then decide whether they should follow the recommendation or not. In recent research, Miller [153] argues that the recommendation-driven approach has two main issues: (1) by explaining just the AI recommendation, it limits the user's control, especially when few alternatives are offered if the user disagrees with the AI; (2) the recommendation-driven approach does not align with the cognitive processes of human decision-making. Therefore, challenges remain in building more effective AI-assisted systems in the future.

First, users may over-rely on the recommendation because of overtrust in the model, which is problematic when the model is wrong. On the other hand, users may under-

rely on AI recommendations, in which the users do not follow the prediction when it is correct because of distrust in the model. Importantly, research has shown that providing explanations does not always reduce over-reliance on AI recommendations, compared to only providing AI predictions [16, 27]. More troubling is the fact that adding more information (i.e., adding explanations for the AI recommendation) is not always helpful and sometimes can lead to worse performance and overconfidence in the wrong information [167, 168]. For instance, explanations can result in people viewing the model's incorrectness as being correct [103] and people being deceived by incorrect explanations regardless of their expertise [158]. This suggests that the current explainable AI approach might not always be effective in helping users make better decisions.

Second, providing recommendations and explanations does not align with the cognitive processes of human decision-making [82, 112, 153, 174]. Specifically, Hoffman et al. [82] argue that abductive reasoning is an appropriate basis for conceptualising explainable AI, as it involves generating hypotheses to explain an event. Moreover, based on Yates and Potworowski [232]'s definition of cardinal decision issues, Miller [153] identifies six criteria for good decision aid, including: (1) help to identify options, (2) help to identify possible outcomes for each option, (3) help to judge which outcomes are most likely, (4) help to identify impacts on stakeholders, (5) help to make trade-offs between options, and (6) help to understand the machine decision. The current recommendation-driven approach (i.e., giving AI recommendations and explanations) only satisfies criteria (6).

This thesis aims to build more reliable and trustworthy explainable decision-support approaches, which are evaluated and applied across various application domains. First, in the recommendation-driven approach, users are typically provided with AI recommendations and explanations. Most existing research has used uncertainty measures as a means to improve user trust [222, 239]. However, these approaches do not explain the model uncertainty, which represents a promising research direction for further improving users' trust and understanding in the model [202]. Second, as discussed above, even when uncertainty and explanations are provided, the recommendation-driven approach still limits user agency and it does not align with human cognitive processes. Therefore,

this thesis aims to address two main challenges: (1) Explaining model uncertainty, and (2) Designing a decision-support model that can calibrate reliance on AI. In summary, this thesis contributes methods for explaining recommendation uncertainty and designing decision-support systems that enhance user trust, understanding and decision quality.

## 1.1 Research Motivations

I summarise the two challenges and research questions of this thesis as follows:

**Challenge 1: Model Uncertainty** The first challenge is to explain the machine learning (ML) model's uncertainty in the recommendation-driven approach, which can be crucial for building trust and understanding in these models. Specifically, this uncertainty refers to the uncertainty in the ML model's predictions, which can be caused by various factors, such as the internal structure of the ML model or uncertainty in the training data. While recent research has used confidence (uncertainty) measures as a way to improve trust and understanding in ML models, these approaches [222, 239] do not provide explicit explanations of their uncertainty. Therefore, I aim to address this challenge in order to help users understand *why* the model is confident (or not confident) in its predictions, and then decide whether they should trust the model's decisions.

**Challenge 2: Decision-Support Model** The second challenge is to design a decision-support model that is promising in helping users make better decisions by applying the new *Evaluative AI* paradigm [153]. While we gained valuable insights from the positive results of explaining uncertainty, we did not believe that pursuing this line of work would improve human decision-making with AI assistance. The recommendation-driven approach in Challenge 1 still has limitations in terms of user agency and cognitive processes as explained above. Therefore, the new paradigm aims to address the issues of over and under-reliance on decision-support tools by providing evidence for possible hypotheses and allowing users to make informed decisions based on the evidence. This paradigm is built on Peirce's notion of *abductive reasoning* [174], which is argued to best reflect the human decision-making process in the context of human-AI interaction [82]. More specifically, abductive reasoning refers to *hypothesis-driven approach* (as opposed to a

recommendation-driven approach), which is a form of decision-support that starts with a hypothesis and then provides evidence that supports or refutes the hypothesis. For this reason, I aim to address this challenge by proposing a new evidence-based decision-support approach based on the Evaluative AI paradigm and evaluating its effectiveness in improving decision quality and reducing over-reliance on AI recommendations.

To address these challenges, this thesis aims to answer the following research questions. In Chapter 3, I address **Challenge 1** by answering questions **RQ1** and **RQ2**.

**RQ1.** How can we **explain model uncertainty**?

**RQ2.** Can explaining model uncertainty improve user **trust** and **understanding** in the machine learning model?

To answer **RQ1**, I formalise the counterfactual (CF) explanation of confidence score to explain model uncertainty. Counterfactual explanations are chosen because people tend to focus more on counterfactuals than factual ones when seeking explanations [32, 150]. Particularly, the CF model provides explanations to change the confidence score of a specific output class. Then, I address **RQ2** by conducting two user studies to evaluate the effectiveness of this explanation approach. I also compare two different counterfactual explanation approaches, example-based explanation and visualisation-based explanation, and investigate how users perceive and use these explanations differently.

Subsequently, in Chapter 4, I address **Challenge 2** with **RQ3**.

**RQ3.** How can we design an effective **evidence-based decision-support** model?

I address **RQ3** by proposing a new decision-support approach called *evidence-informed hypothesis-driven decision-making*. This decision-support approach aims to help users make better decisions by providing evidence for possible hypotheses. Through human behavioural experiments, the new approach has been shown to improve decision quality and reduce over-reliance on AI recommendations. Moreover, using qualitative analysis, I explore the limitations and challenges of the new approach and compare it with two base-

line decision-support approaches, namely recommendation-driven and AI-explanation-only.

I further apply the new decision-support approach to build a decision-aid tool for supporting skin cancer diagnosis in Chapter 5, called *Visual Evaluative AI (VisE)*. In this chapter, I respond to **RQ4** and **RQ5**.

> **RQ4.** Based on the new decision-support paradigm, how can we build a **decision-aid tool** for **image datasets**?

> **RQ5.** How do different decision-support approaches impact **human decision-making** in **skin cancer diagnosis**?

For **RQ4**, I introduce the *Visual Evaluative AI tool*, which combines concept-based explanations and the weight of evidence (WoE) framework to provide hypothesis-driven decision-support for image datasets. The original WoE framework has only addressed tabular data, which is different from image data. Extracting features from images is more challenging and requires techniques like convolutional neural networks (CNNs). Therefore, I extend the WoE framework to support image data by combining it with concept-based explanations. Concept-based explanations will find human-understandable high-level concepts in the image data, which represent the *features* of the image. These features are then put into the WoE framework to generate evidence for possible hypotheses of the image.

Moreover, I apply this decision-support tool in a case study of skin cancer diagnosis, but it can be used in other computer vision domains as well. To address **RQ5**, I conduct a user study with participants who have backgrounds in the skin cancer field to understand how different decision-support interfaces (recommendation-driven and hypothesis-driven) can impact their decisions differently. Study participants use the tool to make diagnosis decisions and provide feedback on the tool's effectiveness through a semi-structured interview. Based on this study, I aim to measure the effectiveness of the two decision-support interfaces in terms of decision quality and user experience.

## 1.2 Methods

To address the research questions, I use a mixed-methods approach by combining quantitative and qualitative methods. The quantitative method includes both computational and human experiments to measure numerical data such as completion time, model's accuracy, users' performance and users' satisfaction. The qualitative method includes interviews and content analysis to understand users' perceptions and behaviours. It focuses on analysing non-numerical and textual data to identify codes and themes in the text.

Regarding the selection of domains, when conducting user studies with laypeople on crowdsourcing such as Amazon Mechanical Turk (Amazon MTurk) [28] and Prolific[1], the tasks should not require expertise. Moreover, the tasks should not be too easy that participants can complete them without the decision-aid [213]. Therefore, I choose the domains of *income prediction* [143, 217, 228], *resignation prediction* [99, 199], and *housing price prediction* [43, 179] for the user studies in Chapter 3 and 4. These domains are everyday knowledge and are suitable for participants with different backgrounds. Example tasks are provided in Appendix A and B.

Domain experts can make decisions very differently from laypeople as they can incorporate their prior domain-specific knowledge with the information provided by the aid. For that reason, I aim to improve the experiment further by involving human expertise in the decision-making process. Therefore, I choose the *skin cancer domain* [17, 37, 205] for the user study in Chapter 5. Participants are recruited through professional networks, who have backgrounds in the skin cancer field (PhD students, postdoctoral researchers, doctors and melanographers). The study is conducted with both experienced (doctors, melanographers) and inexperienced participants (PhD students, postdoctoral researchers) in skin cancer diagnosis to understand how different decision-support approaches can impact human decision-making differently. Moreover, participants are asked to provide their opinions on the decision-support tool through a semi-structured interview. Details of this experiment are included in Appendix C. The results help explore how participants would use the tool and what they think about the tool.

[1] https://www.prolific.com

## 1.3 Thesis Contributions

In this section, I will summarise the contributions of this thesis. The papers' notations (**C1**, **C2**, **P1**, and **P2**) refer to the notations in Preface.

1. The thesis proposes a method to explain model uncertainty, and therefore, promote trust and improve end-user understanding. Specifically, I formalise the *counterfactual explanation of confidence (uncertainty) score*. User studies indicate that providing counterfactual explanations of confidence scores can help users better understand and trust the model. Through qualitative analysis, I identify some limitations of the two explainability approaches (example-based explanation and visualisation-based explanation). These limitations suggest directions for improving presentations of counterfactual explanations. This work has resulted in **C1** [127].

2. The thesis defines the *evidence-informed hypothesis-driven decision-making* model based on the hypothesis-driven approach and the Weight of Evidence (WoE) framework. I conduct two human behavioural experiments to compare our (1) *hypothesis-driven* approach with two baseline decision-making approaches (2) *recommendation-driven* and (3) a form of *cognitive forcing* that provides only AI explanations and withholds the AI recommendations. The results indicate that the hypothesis-driven approach improves decision quality and reduces over-reliance, with an increase in under-reliance. Our qualitative analysis further identifies some limitations and challenges in the three approaches and shows that participants used the hypothesis-driven approach in a materially different way than the recommendation-driven or AI-explanation-only conditions. This contribution has resulted in **C2** [128].

3. The thesis proposes and studies an *Visual Evaluative AI* tool for image datasets by combining concept-based explanations and the weight of evidence (WoE) framework. This tool offers hypothesis-driven decision-making by generating evidence for possible hypotheses of an image. I provide public access to this tool as a Python package so other researchers can use it. Its application is further demonstrated in supporting skin cancer diagnosis. Through a user study with experienced participants in the skin cancer field, I explore how different decision-making interfaces

(*recommendation-driven* and *hypothesis-driven*) can impact human decision-making differently. This work has resulted in **P1** [130] and **P2** [129].

## 1.4 Thesis Overview

| Chapter | Title | Research Questions | Contribution Summaries |
|---|---|---|---|
| 1 | Introduction | | Research motivations and thesis overview |
| 2 | Background | | Literature review on model uncertainty, explainability, trust, decision-making and skin cancer |
| 3 | Explaining the Uncertainty | RQ1, RQ2 | Counterfactual explanation of confidence score |
| 4 | Hypothesis-Driven Decision-Making Model | RQ3 | Evidence-informed hypothesis-driven decision-making model |
| 5 | Visual Evaluative AI | RQ4, RQ5 | Evaluative AI tool for image datasets. Application of this tool in skin cancer diagnosis |
| 6 | Conclusion | | Summary of contributions and future work |

Table 1.1: Thesis Overview

Table 1.1 provides an overview of the thesis structure. In Chapter 2, I review the background and related work on model uncertainty and explainability. Chapter 3 presents the counterfactual explanation of confidence score to explain model uncertainty. Chapter 4 introduces the evidence-informed hypothesis-driven decision-making model. Chapter 5 demonstrates the Evaluative AI tool for image datasets and its application in skin cancer diagnosis. Finally, Chapter 6 concludes the thesis and discusses future work.

# Chapter 2
# Background and Related Work

In this chapter, I provide an overview of the background and related work that is relevant to this thesis. I review the literature on uncertainty, explanations, trust, AI-assisted decision-making and skin cancer. Figure 2.1 shows the relationship between the literature review and the corresponding chapters in this thesis.

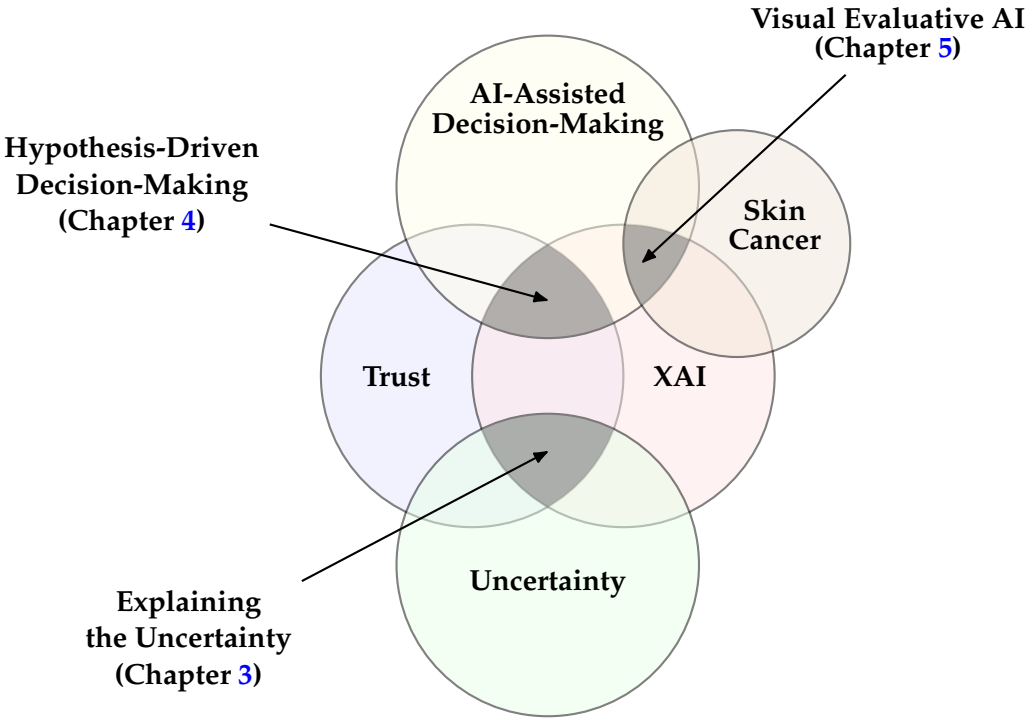



Figure 2.1: Literature Review and Corresponding Chapters

First, I discuss the literature on uncertainty (Section 2.1), an important concept for understanding how AI systems can measure and communicate uncertainty in their predictions. Moreover, uncertainty has been used to foster trust and support decision-making.

This review lays the groundwork for Chapter 3, where I explore how uncertainty can be explained in AI systems.

Next, I examine the literature on explanations (Section 2.2), which is the foundation of all chapters in this thesis. Various explanation forms are considered, including counterfactual and contrastive explanations (Chapter 3), example-based explanations (Chapter 3 and 5), feature-based explanations (Chapter 4 and 5) and concept-based explanations (Chapter 5). The Weight of Evidence (WoE) framework and why we choose to use it in Chapter 4 and 5 are also discussed in Section 2.2.4. Following this, I delve into the literature on trust (Section 2.3), which is a key factor in evaluating user acceptance of AI decision support systems (Chapter 3 and 4).

I then review the literature on AI-assisted decision-making, beginning with human cognitive processes that form the basis for implementing abductive reasoning in decision support [82, 153]. Argumentation theory is also discussed as a complementary framework to abductive reasoning, offering an approach to choose the best decision based on arguments for and against a decision. Subsequently, I review different AI-assisted decision-making paradigms, which inform the experimental designs in Chapter 4 and 5. This section concludes with an overview of how XAI has previously been applied in decision support. Finally, Section 2.5 provides essential background on XAI in support of skin cancer diagnosis for Chapter 5, in which I evaluate different AI-assisted decision-making approaches in this context.

## 2.1 Confidence (Uncertainty) Measures

A common approach to measuring uncertainty in prediction is to use the prediction probability [21, 49]. To evaluate the uncertainty quality, we need the uncertainty to be *calibrated* [21, 118]. A well-calibrated uncertainty of an output reflects the true probability of that output. For example, in a calibrated system, if it predicts that an employee will resign with a probability of 70%, then, in reality, 70% of the time, the employee will leave. In fact, prediction probabilities are often poorly-calibrated [76, 92] [65, p55], leading to either overconfident or underconfident in the prediction. This problem can result in users

having a false sense of trust in the corresponding uncertainty (or confidence) measures.

Another approach to measure the uncertainty is through *uncertainty sampling* [132]. This approach queries unlabelled instance $x$ with maximum uncertainty to get human feedback. There are four types of uncertainty sampling such as: *least confidence*, *margin of confidence*, *ratio of confidence* and *entropy* [195, p12],[157, p70]. Formally, assuming a classification prediction probability is $P(y|x)$ and $\mathcal{Y}$ is a set of classes, the uncertainty measure $U(x)$ can be defined as follows:

- *Least confidence* is the difference between 100% probability and the highest probability returned by the model:

$$U(x) = 1 - \max_{y \in Y} P(y|x) \tag{2.1}$$

- *Margin of confidence* is the difference between the first and second highest probabilities.

$$U(x) = P(y = y_1|x) - P(y = y_2|x) \tag{2.2}$$

  where $y_1 = \arg\max_{y \in Y} P(y|x)$ and $y_2 = \arg\max_{y \in Y \setminus y_1} P(y|x)$

- *Ratio of confidence* is the ratio between the first and second highest probabilities.

$$U(x) = \frac{P(y = y_2|x)}{P(y = y_1|x)} \tag{2.3}$$

- *Entropy*:

$$U(x) = -\sum_{y \in Y} P(y|x) \log P(y|x) \tag{2.4}$$

We will now evaluate the differences among the four measures above [195, p14], [157, p93]. *Least confidence* only considers the confidence information of the most likely class. Thus, it overlooks the information of the remaining classes. *Margin of confidence* and *ratio of confidence* overcome this shortcoming by measuring the margin between the two most probable classes. However, these two measurements still ignore most information on the output probability distribution when we have many label classes (more than two classes). For that reason, *entropy* is better to find the confidence (or uncertainty) among all classes.

In a binary classification task, *margin of confidence* or *ratio of confidence* is similar to *entropy* because they already consider the difference of all classes (i.e., only two classes in this case).

### 2.1.1 Types of Uncertainty Measures

Fundamentally, there are two types of uncertainty, (1) aleatoric uncertainty and (2) epistemic uncertainty [208]. Aleatoric uncertainty is called indirect uncertainty and comes from the noise in the data. Epistemic uncertainty is direct uncertainty that stems from whether we chose the right model that best explains the data. This is also referred to as *model specification uncertainty* or *architecture uncertainty* [21].

In the context of machine learning, uncertainty estimation algorithms can be classified as (1) *intrinsic method* and (2) *extrinsic method* [68]. Intrinsic methods mean the uncertainty estimation is implicitly provided along with the output predictions. By contrast, extrinsic methods are used to provide post-hoc uncertainty estimates or improve the existing uncertainty estimates. Moreover, uncertainty algorithms can also be categorised based on the type of machine learning model such as (1) classification (entropy, mutual information) and (2) regression (confidence intervals, quantiles) [21].

### 2.1.2 Uncertainty Communication

Communicating uncertainty is important to build a trustworthy AI system and support trust calibration for end-users [21, 239]. Uncertainty is often expressed in one (or a combination) of the following ways: *visual*, *numerical* or *verbal (words)* [208]. Choosing the right format to communicate uncertainty is crucial to ensure that the audience can understand the uncertainty, and therefore help build a trustworthy system and support decision-making.

A key issue of using *numerical* expression is that it might be challenging for people with low numeracy skills. For example, in some experiments [184], participants rate $2/3$ as being smaller than $3/5$ (known as *ratio bias* or *denominator neglect*). Bhatt et al. [21] also identify that humans have cognitive biases, which can impede them from understanding

uncertainty. A kind of cognitive bias is called *framing*, in which people prefer an option depending on the context of the information. For example, people would feel less worried about this statement "about 85% chance of surviving after a surgery" than "about 15% chance of dying after a surgery" even though these two statements mean exactly the same thing. However, it is important to note that, despite these limitations, numerical expression is more precise than other forms of expression.

In contrast, *visualisation* is more attractive and easier to understand the trends and patterns in the data. While there are not many designs for *numerical* and *verbal* format, there are many ways to visualise uncertainty. Some common uncertainty visualisation techniques are: error bars, box plots, icon arrays, violin plots, quantile dot plots and hypothetical outcome plots (HOP). Padilla et al. [169] describe different visualisation methods and explore in depth the advantages and disadvantages of each method using cognitive theories. For instance, icon arrays can address *denominator neglect* mentioned above. Besides the obvious advantage of visualisation, a challenge of this format is *deterministic construal error (DCE)* [95, 191], in which people mistake uncertainty information as a simpler but wrong interpretation, to reduce cognitive load. For example, people can incorrectly interpret the error bar as a representation of the maximum and minimum values. It is worth mentioning that deterministic construal error has only been found with visualisation, not with other formats.

Since all formats have their own pros and cons, there is no one-size-fits-all uncertainty communication for all domains. Therefore, it is important to consider the study context and the audience when choosing the right format to communicate uncertainty.

### 2.1.3 Uncertainty and Trust

Uncertainty is a complementary form of communicating transparency and therefore can be an advantage to build trust between stakeholders and systems [6, 21, 188]. Some concerns are that communicating uncertainty can undermine people's trust in the facts. But Van Der Bles et al. [209] showed that communicating uncertainty has a minor effect on trusting news articles. Furthermore, Zhang et al. [239] found that confidence score can help people calibrate their trust in the AI system, and know when to trust or not trust

the AI recommendation. Wang et al. [222] also improved model understanding and trust by showing feature attribution uncertainty. In this case, they use LIME [185] to express the feature attribution, which measures how much each feature contributes to the model prediction. They then show the uncertainty of the feature attribution by using violin plots.

Some studies have pointed out that the effect of uncertainty communication on trust is not always positive. Displaying uncertainty might only promote trust in AI recommendations under low cognitive conditions; however, it might decrease trust under high cognitive load [243]. When combining both model confidence and model accuracy in human-AI interaction settings [182], model accuracy has more impact on people's belief in the model recommendation, as well as their self-reported trust in the model.

### 2.1.4 Uncertainty and Decision-Making

Uncertainty helps people to better understand the system output and then combine the prediction output from the system with their own judgment. Therefore, presenting uncertainty effectively can help improve human decision-making when interacting with a recommendation system. For example, combining decision aid with uncertainty information can result in better users' performance, compared to using the decision aid alone [96, 161]. A key in developing a successful AI-assisted decision-making system is to help users know when to trust or distrust the model's recommendation, and therefore, form a correct mental model of the model's error boundaries [15]. Specifically, in the medical application, we can have medical professionals intervene when our model is incorrect. This is called *reject option* [22, 162].

But showing uncertainty is not always helpful in improving decision-making [222, 239]. Wang et al. [222] suggested that suppressing uncertainty can, in fact, improve decision-making. In this example, suppressing uncertainty involves minimising the attribution of inputs with high uncertainty, and relocating that attribution to other input features. Moreover, it is important to consider whether people can bring their knowledge in the decision-making process, to recognise the model's errors [239]. Otherwise, it is challenging to improve the performance of human-AI interaction overall. Moreover, since

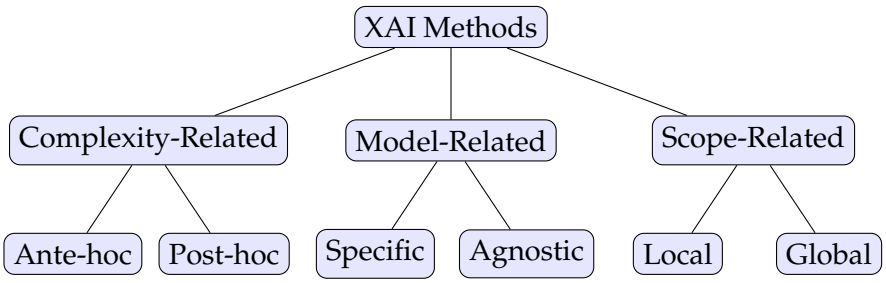



Figure 2.2: XAI methods classification

there are different types of uncertainty, there is still limited work in empirical studies about how people use aleatoric versus epistemic uncertainty to better their decision [21].

Besides the model uncertainty, Zhang and Hußmann [240] indicated the focus on *output uncertainty* instead of the complexity and opaqueness in human-AI interaction. Output uncertainty refers to the user's uncertainty about the AI system output. Users can remain uncertain about the model output regardless the model is highly confident in its prediction or not. The current decision-making approach is to show fully automatically generated outputs to the end users. They are then responsible for understanding the reasons behind these outputs, deciding whether they want to reject the decision and updating the algorithm accordingly. Therefore, we need to address the output uncertainty to truly build human-centred AI systems, instead of algorithm-centred AI systems.

## 2.2 Explanations

Explanations in human-AI interaction can facilitate understanding, trust [140, 146, 238], fairness judgement [53] and decision-making [75]. In this section, we will review different forms of explanations in explainable AI (XAI) and how they are applied and evaluated in various domains.

XAI methods can be classified based on the complexity of the model, the model-related method or the scope-related method. First, we can classify explanation (or interpretability) methods into *intrinsic (ante-hoc)* and *post hoc* explanations. Intrinsic (ante-hoc) explanations refer to the models that are simple in their structure and complexity and therefore easy to interpret (e.g. linear models, decision trees, K-nearest neighbours).

By contrast, post hoc explanations are applied to analyse a more complex model after its training [185]. Examples of models that need post hoc explanations are neural networks and random forests.

Another way to categorise explanation methods is based on *model-specific* and *model-agnostic*. Model-specific explanations are explanation methods limited to a certain model, while model-agnostic methods are independent of the model. All model-agnostic methods are post-hoc explanations. Moreover, the explanation can either be *global* or *local*. Local explanations refer to a single prediction [69, 142, 218]. Global explanations explain the entire model behaviour (e.g. Partial Dependence Plot [62]).

There are four common forms of explanations: (1) counterfactual and contrastive explanations, (2) example-based explanations, (3) feature-based explanations and (4) concept-based explanations. We will review each of these explanations as follows.

### 2.2.1 Counterfactual and Contrastive Explanation

#### Counterfactual Explanation

A counterfactual explanation is described as the possible smallest changes in input values to change the model prediction to the desired output [218]. It has been increasingly used in explainable AI (XAI) to facilitate human interaction with the AI model [31, 150, 151]; for example, applied in credit application [72]. Counterfactual explanations are expressed in the following form: "You were denied a loan because your annual income was $30,000. If your income had been $45,000, you would have been offered a loan" [218]. To generate counterfactuals, Wachter et al. [218] suggest finding solutions for the following loss function.

$$\arg\min_{x'} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') \tag{2.5}$$

where $x'$ is the counterfactual solution; $(f(x') - y')^2$ presents the distance between the model's prediction output of counterfactual input $x'$ and the desired counterfactual output $y'$; $d(x, x')$ is the distance between the original input and the counterfactual input; and $\lambda$ is a weight parameter. A high $\lambda$ means we prefer to find counterfactual point

$x'$ that gives output $f(x')$ close to the desired output $y'$, a low $\lambda$ means we aim to find counterfactual input $x'$ that is close to the original input $x$ even when the counterfactual output $f(x')$ can be far away from the desired output $y'$. This loss function can be solved by iteratively increasing $\lambda$ until a close solution $x'$ is found. In this model, $f(x)$ would be the output, such as a denied loan, and $y'$ would be the desired output – the loan is granted. The counterfactual $x'$ would be the properties of a similar customer that would have received the loan.

A key property of counterfactual explanations is *coherence* [61, 150], i.e., the counterfactual explanations should be realistic and consistent with prior beliefs. Russell [189] proposed a search algorithm to generate counterfactual explanations based on mixed-integer programming, assuming that input variables can be continuous or discrete values. They defined a set of linear integer constraints, which is called *mixed polytope*. These constraints can be given to Gurobi Optimization [77] and then an optimal solution is generated. They find the counterfactual point $x'$ by solving this function.

$$\arg\min_{x'} ||\hat{x} - x'||_{1,w} \tag{2.6}$$

where $\hat{x}$ is the mixed encoding of $x$; $x'$ lies on the mixed polytope; $||.||_{1,w}$ is a weighted $l_1$ norm with weight $w$ is defined as the inverse median absolute deviation (MAD) [218].

A main drawback of Russell [189]'s model is that it can only be applied to linear classifiers. Mothilal et al. [159] propose another CF search engine by addressing both *feasibility* and *diversity* of the counterfactual explanations, so-called *Diverse Counterfactual Explanations (DiCE)*. They formulate four necessary constraints for counterfactual explanations, which are: (1) *diversity*: the counterfactual explanations should be diverse; (2) *sparsity*: the counterfactual explanations should require changes in fewer number of features; (3) *proximity*: the counterfactual explanations should be closest to the original input; and (4) *user constraints*: the counterfactual explanations should satisfy user-defined constraints.

#### Evaluating Counterfactual Explanation

Keane et al. [102] provided a survey of 100 distinct counterfactual explanation methods

to determine five key deficits in evaluating counterfactual explanation approach as follows: (1) neglecting users, (2) lacking plausibility, (3) addressing sparsity, (4) coverage assessment and (5) comparative testing. First, neglecting human studies is a common issue when Keane et al. [102] found only 31% of papers have user studies (36 out of 117) in their survey. More importantly, some of these studies have unreproducible designs.

The second problem is *plausibility*. The definition of plausible explanations can vary depending on different techniques. Here we have two types of plausible explanations: plausibility-as-proximity and plausibility as more-good-features.

- *Plausibility-as-proximity*: find counterfactual $x'$ point such that the distance between the counterfactual point and the fact point $x$ is minimum and $f(x') = y'$ where $y'$ is a new target and $f$ is a classifier function [218]. However, finding the minimum distance can be problematic. For example, using a low-distance score would not be enough if the counterfactual explanation is meaningless (e.g. a class of 25.2 students is very close to a target class of 25 students. However, this comparison violates common sense as in reality, we never have a class of 25.2 students). Also, distance metrics (e.g. $L_1$, $L_2$ or others) should be *psychologically grounded*; that is, people should find these metrics acceptable. Keane et al. [102] note that there are currently no user studies that can confirm which distance metric people would prefer the most.

- *Plausibility as more-good-features*: determine the number of "good" features that are needed for the counterfactual explanation. For example, immutable features are not good features. We can consider actionable and mutable features as "good" features.

Third, a good counterfactual explanation should be sparse; that is, we change the *least* number of features to get counterfactuals. Sparsity is important due to human memory limits and people only care for some reasons instead of all reasons when they ask for explanations [150]. It is also difficult to find the best number of features that need to be modified in counterfactuals as it varies in many research. Keane and Smyth [101] select one or two feature changes as *good* counterfactuals. Additionally, Warren et al. [226] propose a psychologically plausible method by prioritising categorical features over con-

tinuous features to find the best counterfactuals. User studies further show that people find explanations referring to categorical features easier to understand compared to those involving continuous features [227].

Fourth, the coverage refers to the guarantee that the counterfactual method will produce *good* explanations as a whole. Explanations are considered *good* if they are psychologically acceptable, which can be measured by many metrics. For instance, some can use out-of-distribution (OOD) measures to differentiate between valid and invalid CFs. Formally, the *explanatory coverage* is defined as follows:

$$\text{XP\_Coverage\_Set}(X) = \{x' \in X | \exists x \in X \setminus \{x'\} \& \text{explains}(x, x')\}$$

$$\text{XP\_Coverage}(X) = |\text{XP\_Coverage\_Set}(X)|/|X| \tag{2.7}$$

where: $X$ is a dataset, $x$ is the test instance, $x'$ is the counterfactual instance, $\text{explains}(x, x')$ is psychologically acceptable counterfactual explanations; XP_Coverage is the *explanatory coverage*, which is the size of the coverage set divided by the size of the dataset.

Fifth, they found that papers on counterfactual explanations lack comparative testing between different methods. With a set of available measures, they encourage researchers to use these measures to evaluate their counterfactual explanation methods and make their code publicly available.

In summary, Keane et al. [102] identified five key deficits in evaluating counterfactual explanations. They also propose a roadmap and evaluative benchmarks to address these issues. To benchmark evaluative methods, counterfactual explanations should be evaluated based on the following criteria as mentioned above, including *proximity*, *sparsity* and *coverage*.

#### Contrastive Explanation

Miller [150] highlighted an important factor of explanation is that explanations are contrastive. That is, they do not address why an event happened but rather why it happened instead of another event. For example, they do not ask *Why A?* (factual question) but instead they ask *Why A instead of B?* (*contrastive question*). We refer $A$ as the *fact* and $B$

as the *foil* (or can be called *counterfactual*). Miller [151] argue that contrastive explanation is important according to researchers in social science because of two reasons: (1) Contrastive explanation helps identify what people expect to happen when they are surprised; (2) Presenting contrastive explanations is simpler to both explainer and explainee. The main difference between *contrastive* explanations and *counterfactual* explanations is that *contrastive* explanations are the differences between the fact and the foil, whereas *counterfactual* explanations only address the foil [138, 151].

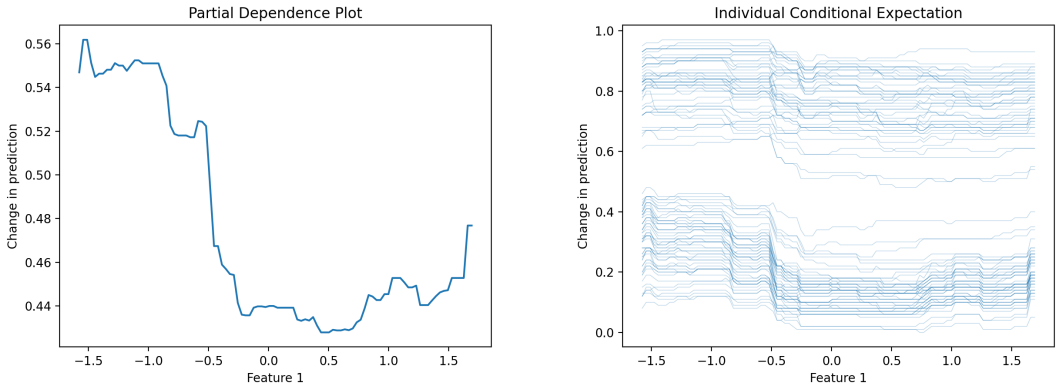
A challenge of designing contrastive explanations is to find a *good foil*. In the case of binary classification which only has two output classes, the foil is easily identified. However, when we have more than two possible outcomes, we need other methods to find which options people consider so that we can have a right contrastive explanation [64].

In practice, Lucic et al. [140] proposed *Monte Carlo Bounds for Reasonable Predictions* to understand why there are large errors in a model prediction that aims to explain errors in regression prediction in a sales forecasting problem. This new approach determines (1) important features (find the reasonable bounds) and (2) directions between each feature and the output. van der Waa et al. [210] proposed a decision tree trained on generated data points that belong to the foil (counterfactual) class. Followed by identifying the fact-leaf in which the data points of the fact class reside. Next, they identified the foil-leaf which contains the data point of the foil class by choosing the closest leaf to the fact-leaf. To evaluate this new approach, they applied four classification models (a random forest, logistic regression, support vector machine and a neural network) on three benchmark classification tasks to prove the model-agnostic essence. Another example of generating contrastive explanations for neural networks is by highlighting the pertinent positives (minimally sufficient presence to justify the prediction) and pertinent negatives (necessary absence to justify the final outcome) [51].

### 2.2.2 Feature-Based Explanation

Feature-based explanations describe how input features contribute to the model prediction. This explanation assigns a score to each feature to indicate its importance in the prediction. Some common feature-based explanation methods are Partial Dependence

Plot (PDP) [62], Individual Conditional Expectation (ICE) [69], Accumulated Local Effects (ALE) [8], Local Interpretable Model-agnostic Explanations (LIME) [185] and SHapley Additive exPlanations (SHAP) [142].




(a) Partial Dependence Plot (b) Individual Conditional Expectation

Figure 2.3: Feature-based explanations: (a) Partial Dependence Plot (PDP) and (b) Individual Conditional Expectation (ICE).

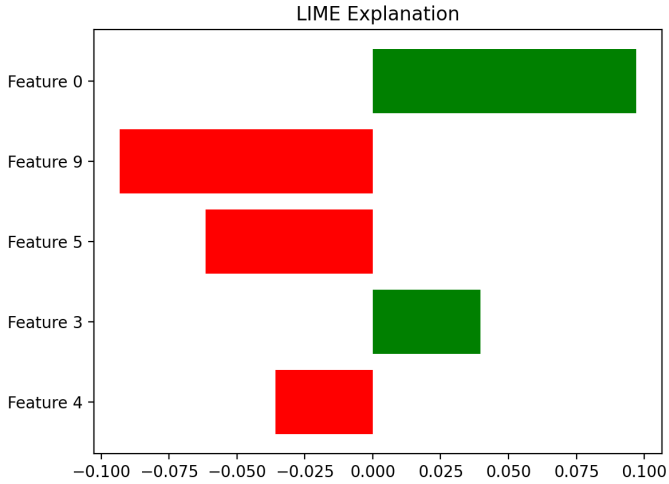



Figure 2.4: Local Interpretable Model-agnostic Explanations (LIME)

The Partial Dependence Plot (PDP [62]) is a global method that shows the relationship between a set of input features and the model prediction (e.g., Figure 2.3a). To calculate a PDP, fix the feature(s) of interest at specific values and get the average predictions across all instances in the dataset. Due to people commonly visualising in two dimensions, showing a PDP is often limited to only one or two features of interest. Moreover, PDP

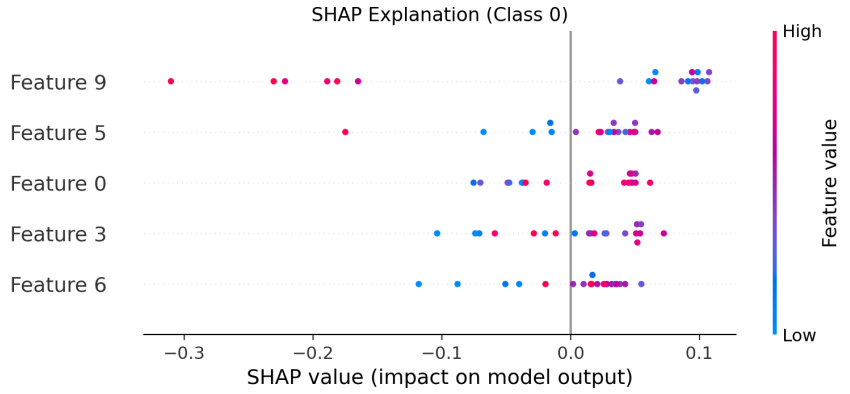



Figure 2.5: SHapley Additive exPlanations (SHAP)

assumes that the features are independent of each other.

The Individual Conditional Expectation (ICE [69]) is a local method that shows the relationship between a feature and the model prediction for a single instance (e.g., Figure 2.3b). To calculate Individual Conditional Expectation (ICE), vary the feature of interest for a single instance across a range of values while keeping all other features fixed. Features are also assumed to be independent of each other. Furthermore, plotting too many ICE curves for many instances can be overwhelming and difficult to interpret.

The Accumulated Local Effects (ALE [8]) is a global method that aims to address the shortcomings of PDP. ALE plots also show how features influence the model prediction. However, they are less computationally expensive than PDP and can handle the case where features are correlated with each other. Instead of getting the average predictions, ALE plots are calculated by taking the difference between the predictions at two values of a given feature.

The Local Interpretable Model-agnostic Explanations (LIME [185]) is a local method that explains the model prediction by using approximate models (e.g. linear models) to quantify the importance of features (e.g., Figure 2.4). LIME generates a set of perturbed instances by changing the input features and then fits a linear model to these instances. The coefficients of the linear model are used to explain the model prediction locally, but cannot guarantee to provide a good global explanation.

The SHapley Additive exPlanations (SHAP [142]) is a game-theoretic method that explains the model prediction by calculating the Shapley values for each feature (e.g.,

Figure 2.5). The Shapley value is a measure of the feature's contribution to the model prediction. SHAP can provide both local and global explanations. However, compared to LIME, SHAP is more computationally expensive.

### 2.2.3 Example-Based (Case-Based) Explanation


Figure 2.6: Example-based explanation

Case-based reasoning [116] provides prediction based on similar past cases of the current instance. This method can also be called *example-based* explanation, where we provide similar examples in the training dataset to explain the model's behaviour. Figure 2.6 shows an example of how an example-based explanation can be presented. Images are taken from the CUB dataset [219]. It shows similar images in the training set to explain the class prediction (e.g., white pelican) of the test image. Example-based explanations are often used in image data. However, they can also be applied to other types of data, such as text and tabular data. For example, a doctor can present similar patient cases to explain the diagnosis of a new patient. Nonetheless, presenting examples for tabular data can be challenging, particularly when there are significant variations in the high-dimensional input features across different examples.

Besides providing similar cases, which are found by using *nearest-like-neighbour (NLN)* explanations, we can find *nearest-unlike-neighbour (NUN)* [165] to present case-based explanations. NUNs can also be referred to as counterfactual explanations, in which we find a minimally different case that has been discussed previously.

### 2.2.4 Evidence-Based Explanation

We describe key differences between the Weight of Evidence approach (WoE) and other feature attribution approaches such as LIME [185] and SHAP [142].

An approach to generate evidence is the *Weight of Evidence (WoE)* framework [148], which can be adapted to meet human-centred design principles. Formally, given the predicted output $y$ for input $x$, WoE seeks how much *evidence* input feature $x_i$ gives in favour of (or against) $y$. WoE is similar to feature importance explanations. However, the main difference is that *Weight of Evidence* uses log likelihoods and log odds ratios to generate explanations, whereas LIME [185] and SHAP [142] find feature importance by modifying the predictive posterior probability in various ways.

We choose the Weight of Evidence approach (WoE) in Chapter 4 and 5 because WoE follows human-centred design principles [148], in which explanations should be *contrastive* (i.e., why the model predicted $y$ instead of alternative $y'$), *exhaustive* (i.e., justify on why every alternative $y'$), *compositional* (i.e., be able to break down into simple components in the prediction), *easily-understandable* (i.e., understandable components) and *parsimonious* (i.e., only provide most relevant facts). Kumar et al. [121] argue that SHAP [142] has several human-centred issues, including non-contrastive and non-actionable explanations, and that most people do not have a correct mental model of Shapley values. LIME [185] does not follow human-centred design principles either. Moreover, LIME [185] uses a surrogate model to approximate the original one, leading to lower performance and explanations that do not reflect the original model. SHAP [142] can have a high computational cost due to the need to compute Shapley values for each instance. In contrast, WoE [148] is a probabilistic approach that does not require a surrogate model and can be computed more efficiently.

A closely related work to Melis et al. [148] is from Poulin et al. [177]. Poulin et al. [177] proposed a framework called *ExplainD* that uses *additive evidence*. The framework also measures the weight of evidence using a Naive Bayes classifier along with highlighting the negative and positive evidence for a decision. However, the problem being considered is a binary classification. Furthermore, there is still room for improvement by conducting experiments to evaluate the framework. Kulesza et al. [119, 120] introduced

*EluciDebug* in email classification using Multinomial Naive Bayes classifier (MNB). The EluciDebug prototype provides an interface that includes important words and the folder size that both contribute to the email classification. In this example, important words can be referred to *strength of evidence*. The folder size which describes the number of instances in each class can be referred to *weight of evidence*. Yet the prototype did not specifically give positive and negative evidence in decision-making situations.

In the literature, there also exist concepts called *strength of evidence* and *weight of evidence*. The strength of evidence is defined as the proportion of evidence that favours one hypothesis, and the weight of evidence is defined as the total number of evidence [122]. Alternatively, the strength of evidence can be understood as the confidence score of the model, and the weight of evidence can be defined as the sample size being considered [63]. When assessing the plausibility of a model, which refers to how likely a model is to be accepted by a user, the strength of evidence positively increases the plausibility; however, the weight of evidence is often ignored [63]. Moreover, Griffin and Tversky [74] indicated that *overconfidence* is when the strength of evidence is high and the weight of evidence is low. In contrast, *underconfidence* happens when we have low strength but strong weight of evidence.

Having more evidence does not always have positive effects on decision-making. Ratcliff and Smith [181] proposed stopping rules that can be applied when additional evidence does not change the final decision. More specifically, *relative stopping rule* refers to when the balance between hypotheses reaches a threshold, meaning evidence for one alternative inherently counts against the other. On the other hand, *absolute stopping rule* leads to a decision when evidence for a single hypothesis reaches a threshold, with alternatives being considered independently.

### 2.2.5 Concept-Based Explanation

Concept-based techniques are commonly used to explain image data. They provide explanations using human-defined concepts that are related to parts (a group of pixels) of images [67, 104]. The explanation is visualised as a segmentation of the image that represents a specific concept. The concept-based model can be classified into two cate-

gories: (1) supervised concept learning (concepts are labelled on each image in the training dataset) and (2) unsupervised concept learning (not having concept labels in the training dataset). Supervised concept learning requires labelled concepts in the training set, or the concepts can be transferred using another labelled dataset [234]. Unsupervised learning concept methods do not require the concepts to be labelled during the training process. This method is helpful when labelling concepts can be laborious, require expertise, or are not always available. Moreover, unsupervised learning can give users more agency as they can find a new concept that has not been labelled, but is still used by a machine learning model.

Kim et al. [105] found that study participants mostly preferred part-based explanations, which highlight the areas representing the concept on images, along with concept scores that indicate how the concept can negatively or positively affect the model prediction (Figure 2.7). Additionally, the score can be represented as a similarity score with the same concept in another image (Figure 2.9). To quantify the importance of each concept to the final classification, a popular approach is called *testing with CAVs (TCAV)* [104]. TCAV uses directional derivatives to measure the sensitivity of the model's prediction to a concept. Some examples of how concept-based explanations may look are shown in Figure 2.7, 2.8 and 2.9. Photos are taken from the CUB dataset [219].

In the following sections, we will review some popular concept-based explanation methods.

#### Supervised Concept Learning

Supervised concept methods require labelled concepts in the dataset. This method is commonly used and often yields good performance. However, it can be expensive to label the concepts. An example of how the found concepts are being presented to users is shown in Figure 2.7. These concepts are segmented on the image and labelled as *beak* and *leg* with their corresponding scores. These scores indicate how concepts contribute to the model prediction.

There are two common approaches for supervised concept methods, depending on whether the concept labels are available in the training data or not, specifically: (1) The

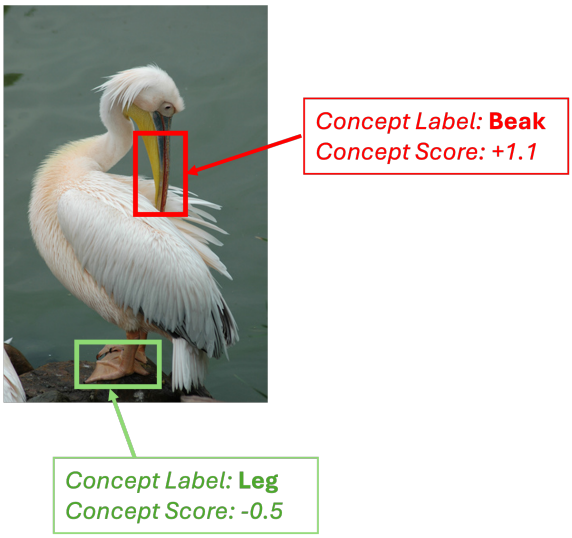



Figure 2.7: An example of supervised concept learning.

concept labels are available within the same training data; (2) The training data does not have concept labels, but the concept labels are available in an external dataset. According to Poeta et al. [176], the former is called *joint concept training* and the latter is called *concept instillation*.

***Joint Concept Training*** An example of joint concept training is the Concept-Bottleneck Model (CBM [115]), where a concept bottleneck layer learns concepts from input features. This model allows *concept intervention*, in which human users can adjust the concepts and see how the model prediction changes. Concept Embedding Models (CEM [57]) improve CBM and overcome the accuracy-vs-interpretability tradeoff in concept-incomplete settings. CEM uses high-dimensional embeddings to represent each concept, and therefore, obtains state-of-the-art accuracy while requiring fewer concept labels compared to CBM.

***Concept Instillation*** Concept instillation is a method that transfers concepts from an auxiliary dataset to the training dataset. In this case, a given layer in the neural network is modified to represent concepts, which can be called the *concept embedding* layer. This method is particularly useful because the concept labels are not always available in the training dataset. For example, Yuksekgonul et al. [234] propose a post-hoc concept bottleneck model (PCBM) that can be applied to any neural network without sacrificing model

performance. This method aims to address the shortcoming of CBM [115] as CBM requires concept annotations in the training set. PCBM learns the concept bank from other datasets and transfers this information to the unlabelled dataset. PCBM also allows users to update the global model, which is more efficient than other works that only allow fixing a specific prediction.

Concept Whitening (CW [41]) method demonstrates how a concept is represented at a particular layer of the neural network by altering the layer using a mechanism called concept whitening, which decorrelates and normalises the latent space. The concepts used in CW can be learned from an external dataset, which is different from the training dataset used for the classification task. CW layer can help us better understand how concepts are built over the layers of the neural network. Moreover, it can be applied in any layer without sacrificing the predictive performance.

#### Unsupervised Concept Learning

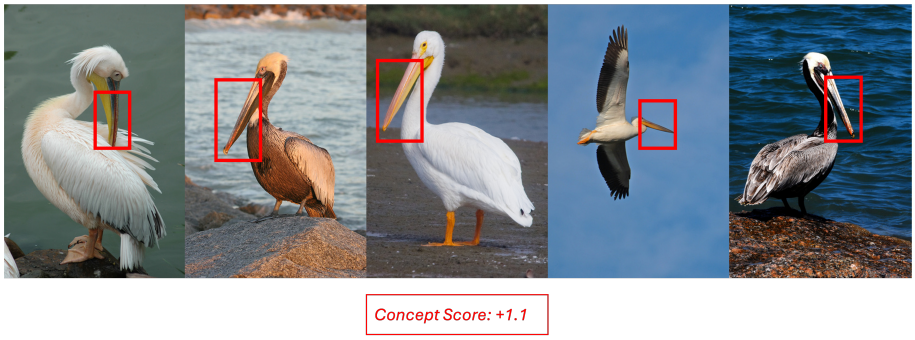



Figure 2.8: An example of unsupervised concept learning (prototype-based explanation).

Unsupervised concept learning methods do not require concept labels in the training dataset. These methods often present a concept using a set of prototypes that have the same concept. This is referred to *prototype-based explanations*. An example is shown in Figure 2.8. The concept is represented as a set of five prototypes, highlighting the areas of interest. This concept is not labelled by the model. However, it can be understood by human users as the concept *beak*. Moreover, another way to present the concept is by drawing boxes around the relevant parts on the test image and adding similarity scores

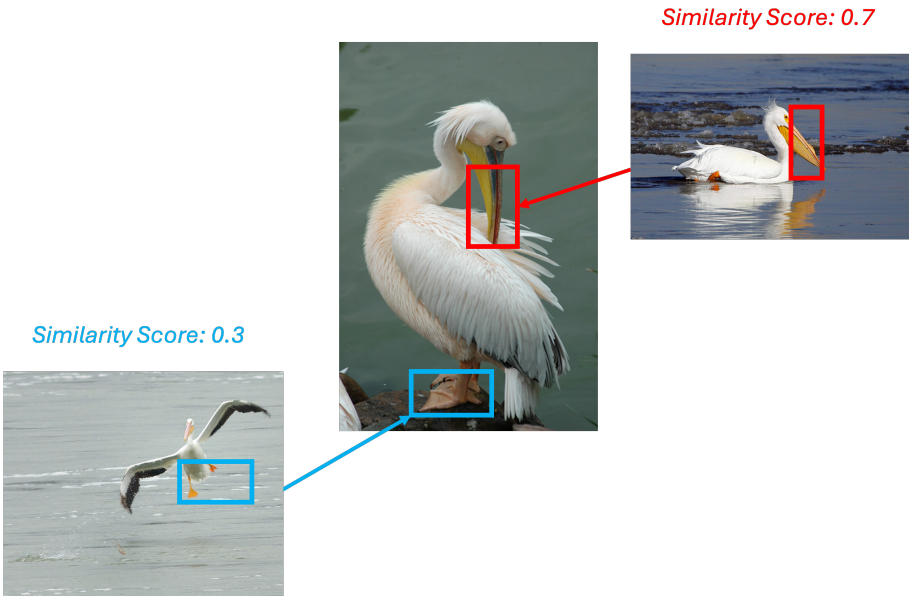



Figure 2.9: Another way to present prototype-based explanation, different from Figure 2.8.

compared to other prototypes as shown in Figure 2.9. We will now review some methods that use unsupervised concept learning as follows.

The Automatic Concept-based Explanations (ACE [67]) method is a global method that explains a classifier class without human supervision. A set of segments is segmented from a set of images from the same class and their resolutions. Similar segments are then clustered into concepts. For each concept, its TCAV score [104] is calculated to determine its importance to the classifier.

The Invertible Concept-based Explanation (ICE [238]) method is built based on ACE by using a range of matrix factorization methods instead of clustering segmentations to address the limitations of ACE. ICE applies Non-negative Matrix Factorization (NMF) for extracting concepts, which offers better interpretability and fidelity compared to those derived from PCA or K-means clustering.

Yeh et al. [233] introduced concept *completeness*, which measures how sufficient the concepts are to predict the model's outcome. The authors also propose a method for concept discovery to maximise finding complete concepts. The results show the proposed method can find complete and interpretable concepts, as well as outperform the previous

methods, including ACE.

## 2.3 Trust

In this section, we will review the concept of trust, focusing on trust in human-AI interactions. We will start by defining trust. We will also discuss the causes of trust and methods to evaluate trust.

### 2.3.1 Definitions of Trust

Trust is a concept that has been studied in various disciplines, including psychology, sociology, economics, and computer science. To define trust, we start to review the definition of trust in human-human scenarios (interpersonal trust). Then we extend the definition to trust in human-machine (human-AI) interactions.

**Definition 2.1** (**Interpersonal trust [147]**). *Mayer et al. [147] define interpersonal trust as: "The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party."*

Mayer et al. [147] also proposed a classic model of trust, the *ABI (Ability, Benevolence, Integrity)* model, which is widely used in the literature. *Ability* refers to the trustee's skills and competencies, which can decide whether the individual can be trusted to complete tasks in some specific domains. *Benevolence* means a trustee is believed to act in the trustor's best interests, rather than acting solely from self-interested motives. *Integrity* is the extent to which the trustee follows a set of principles that are accepted by the trustor.

The above definition is originally considered in the context of human-human trust. For trust in human-machine interaction, we need to understand some challenges. First, trust in human-machine interaction lacks intentionality [131]. For example, an individual may intentionally act in a trustworthy manner to receive affirmation from others. If we consider the *ABI* model, machine agents do not have the same ability, benevolence and integrity as humans. Moreover, Parasuraman and Riley [170] identified flaws in human-

machine partnership in terms of *disuse* and *misuse*, which refer to under-reliance and over-reliance on automation. *Disuse* indicates the neglect of the potential of automation. *Misuse* refers to when people over-rely on automation which can result in inappropriate decisions. To address these shortcomings, Lee and See [131] might be the first to propose a conceptual model of trust in automated systems that can be applied in improving the design of automation systems to promote appropriate reliance.

Based on the definitions of interpersonal trust, Jacovi et al. [87] define trust in human-AI scenarios by considering two notions, *anticipation* and *vulnerability*. Similar to benevolence, anticipation is the degree to which human (trustor) believes that the machine (trustee) will act in the human's favour. Vulnerability refers to the presence of risk and uncertainty in the interaction. They argue that "trust does not exist if human does not perceive risk".

**Definition 2.2** (**Human-AI Trust [87]**). *Jacovi et al. [87] define human-AI trust as: "If H (human) perceives that M (AI model) is trustworthy to contract C, and accepts vulnerability to M's actions, then H trusts M contractually to C."*

**Definition 2.3** (**Human-AI Distrust [87]**). *Jacovi et al. [87] define human-AI distrust as: "If H (human) perceives that M (AI model) is not trustworthy to contract C, and therefore does not accept vulnerability to M's actions, then H distrusts M contractually to C."*

**Definition 2.4** (**Trustworthy AI [87]**). *Jacovi et al. [87] define trustworthy AI as: "An AI model is trustworthy to contract C if it is capable of maintaining the contract."*

Jacovi et al. [87] also clearly separated *trust* (relate to trustor) and *trustworthy* (relate to trustee). If we trust the AI when it is trustworthy, it is *warranted trust*. Otherwise, it is *unwarranted trust*. In addition, if we do not trust the AI when it is trustworthy, it causes *unwarranted distrust*. Of course, we should aim to avoid unwarranted trust and unwarranted distrust, though avoiding unwarranted trust is more important.

**Definition 2.5** (**Overtrust [131]**). *"Trust exceeds system capabilities, leading to misuse."*

**Definition 2.6** (**Distrust [131]**). *"Trust falls short of system capabilities, leading to disuse."*

**Definition 2.7** (**Calibrated Trust [131]**). *"Trust matches system capabilities, leading to appropriate use."*

The concepts of warranted trust and unwarranted trust are quite similar to *overtrust* and *distrust* discussed in [131]. But we believe that the definitions in [87] are more precise because they consider trust on a case-by-case basis, rather than just comparing the trustor's trust with the trustee's capabilities. In human-machine, or more precisely in this case, human-AI interaction, does trust simply refer to the belief that the AI will perform well? A model can have a very high accuracy, but should we rely on it in all cases? Jacovi et al. [87] argue that the "trust in model correctness" refers to "the patterns that distinguish the model's correct and incorrect cases are available to users". This means that a model capable of communicating to users when it is correct is more reliable than one that only shows high performance on certain datasets without providing such transparency. This is where *Explainable AI (XAI)* comes into play.

### 2.3.2 Causes of Trust

Trustworthiness is not a prerequisite for trust; however, it is a necessary requirement for warranted trust [87]. But what can cause trust in human-AI interactions? There are two possible reasons: 1) human's prior knowledge matches the model output, and 2) evaluation methods, which are referred to *intrinsic trust* and *extrinsic trust* [87]. In particular, for intrinsic trust, if users understand the underlying reasoning of the model and the model output matches their prior knowledge, they are more likely to trust the model. However, we should be cautious, as this can cause confirmation bias too. Also, it is not possible for users to have intrinsic trust if they do not have background knowledge of the domain. For extrinsic trust, users trust the model based on the evaluation of the model's performance. This can be done by observing the behaviour of the model from a history of interactions and/or using some evaluation metrics to measure the performance of the model on a set of test data.

In the next section, we will discuss some methods to evaluate trust, focusing on the context of explainable AI.

### 2.3.3 Trust Evaluation

Trust can be divided into two types: *perceived trust (self-reported trust)* and *demonstrated trust*. To measure *self-reported trust*, researchers often measure this in a controlled environment where human participants conduct surveys and/or interviews based on Likert Trust Scale [81, 91, 223] to rate participants' trust on various factors. On the other hand, *demonstrated trust* is measured by observing the participants' behaviour and actions during an interaction, often in controlled settings. The most common example is *trust game* and otherwise called *investment game* [19]. This game is an economic experiment that examines trust and reciprocity between two players. The first player (trustor) is given some money and must decide how much they want to give to the second player (trustee), highlighting trust behaviour. The amount sent is multiplied, and the trustee decides how much they want to return to the trustor, exhibiting the willingness to reciprocate. Trust game has been further applied in the context of human-agent interactions [80, 216] or agent-agent interactions [229].

We should acknowledge that there is a distinction between *trust* and *reliance*. Scharowski et al. [192] advocated for a clearer distinction, arguing that trust is an *attitude*, while reliance is a *behaviour*. More specifically, trust refers to unobservable features, while reliance is observable. For example, in the context of human-AI interaction, users may rely on the system in certain cases, but not necessarily form a trusting attitude towards it. More recent work has highlighted this distinction in conducting their user studies [98, 114].

Miller [152] presented several requirements to evaluate demonstrated trust in human-AI interaction. First, we need to be able to measure the *task performance* of participants when they complete the tasks in the experiment. Second, there must be *presence of risk* and penalty to having unwarranted trust (or unwarranted distrust) when doing the tasks. As mentioned earlier, trust does not exist if there is no risk. However, designing controlled experiments that have risk is not straightforward as the stakes are often low. Some solutions include having monetary bonuses when participants do the tasks well and designing the experiment as a competitive game to encourage participants to engage. However, we have to be clear that the stakes in controlled laboratory environments do not reflect the real-world stakes. Third, participants must have *choices* in the experiment, meaning

they can choose to rely on the AI or not. Importantly, Miller [152] proposed an additional requirement, *manipulated trustworthiness*, which means we can manipulate the level of trustworthiness of a chosen technique. This can be achieved by adding random noise or bias to the selected model to modify its accuracy. For instance, in a within-subject design, for each condition (i.e., each AI technique), participants experience different agents with different levels of trustworthiness [84, 85]. In a between-subject design, each condition has only one agent, and participants choose between the agent or do the tasks themselves.

## 2.4 AI-Assisted Decision-Making

In this section, we review the literature on human decision-making processes and the role of AI in supporting decision-making. We first discuss the cognitive processes in human decision-making, followed by different AI-assisted decision-making paradigms. We then discuss the application of Explainable AI (XAI) in decision support.

### 2.4.1 Cognitive Processes in Human Decision-Making

A well-known psychology theory is the *dual process theory* [97], which suggests that humans have two different systems for processing information: (1) *System 1* is fast, automatic and intuitive; and (2) *System 2* is slow, deliberate and more accurate. Aligning with this theory, there are two different types of decisions that can be made by humans [117]: (1) reflexive decision and (2) multi-attribute decision. A reflexive decision is a simple decision made instantly in a short time. It neither involves many attributes in the input nor demands conscious thoughts. By contrast, a multi-attribute decision is a complicated decision that involves many attributes and numerous alternatives.

We might think that System 2 (slow) is *better* than System 1 (fast) in decision-making and should aim to avoid System 1. However, Miller [153] argues against this view and suggests that we should use the strengths of both systems in designing decision aids. Specifically, the *Naturalistic Decision Making (NDM)* community considered intuition as prior experience that can be used to make rapid and accurate decisions without having

to evaluate all options [109, 113]. The value of intuition can be overlooked if we conduct experiments with laypeople who have no experience in the task and, therefore, fail to capture the advantages of prior knowledge in decision-making [109].

Then *how should we design decision-support systems that can leverage both System 1 and System 2?* Klein et al. [110, 111] examined *sensemaking* from various psychological perspectives, in which *sensemaking* refers to how people make sense of the world. Klein et al. [111, 112] presented a theory of sensemaking known as ***Data/Frame Theory***. The *data* is the information and observations that we use to reconstruct the *frame*, which is a generalisation of a hypothesis. We can question the frame and seek new data to adjust the frame. This process is iterative and done using both System 1 and System 2. People make their decisions first by using their prior knowledge (System 1, System 2) and then search for evidence and make deliberate judgements among plausible options (System 2). Along the same lines, Hoffman et al. [82] argue that Peirce's notion of ***abductive reasoning*** [174] best reflects the cognitive processes in the XAI model. Abductive reasoning is the cognitive process when we give hypotheses to explain an occurred event. Therefore, based on this foundation, we will discuss the idea of *Evaluative AI* [153] in the next section.

#### Cognitive Biases in Decision-Making

Cognitive biases, introduced by Tversky and Kahneman [206], represent systematic errors in judgment, affecting how individuals perceive input information. These biases can lead to inaccurate and irrational decisions in decision-making contexts. Therefore, understanding the effects of cognitive biases on human-AI decision-making settings, with a particular focus on the application of XAI techniques, is important.

Tversky and Kahneman [206] outlined several heuristics and biases such as availability, representativeness, and anchoring. The availability heuristic refers to the tendency to rely on the information that is easiest to recall in memory during the evaluation process. Representativeness involves evaluating the probability of an event based on how similar it is to a typical case, rather than measuring the true statistical probability of the event. Anchoring refers to the reliance on the first piece of information encountered when making decisions. Tversky and Kahneman [206]'s foundational work has continued to

inform much of the literature on cognitive biases such as confirmation bias [108, 163], automation bias [131], framing [207], fixation [112], etc.

The goal of XAI is to enhance the interpretability and transparency of complex AI models, thereby improving human understanding and trust in AI. Despite the benefits of XAI, human cognitive biases can still influence the decision-making process. For instance, Bertrand et al. [20] provided a systematic review of the connection between cognitive biases and XAI. They found that cognitive biases can affect or be affected by XAI in various ways, which were classified into four categories: (1) cognitive biases that affect how XAI methods are designed (e.g., explanatory heuristics); (2) cognitive biases that occurred in user studies (e.g., preference for usability over accuracy); (3) cognitive biases that can be mitigated by XAI (e.g., providing prototypes to mitigate representativeness bias); and (4) cognitive biases that can be worsened by XAI (e.g., confirmation bias can lead to over-reliance on the AI). The recommendation-driven approach is an example of how confirmation bias can lead to over-reliance.

These biases can compromise decision quality even with the help of XAI methods; therefore, de-biasing strategies are needed to reduce the impact of human cognitive biases on decision-making. The idea is to provide explanations to not only the AI's prediction but also other alternative hypotheses, as mentioned in [20, 153]. This approach refers to *hypothesis-driven* decision-making, which can help counter automation bias and fixation by encouraging users to consider evidence of multiple hypotheses. Furthermore, Rastogi et al. [180] proposed a time-based strategy to address anchoring bias by allocating time according to AI confidence.

### 2.4.2 Argumentation Theory

Along with the dual process theory, argumentation theory is another important foundation for designing decision support systems. Argumentation theory deals with constructing arguments and understanding the relationships between them, which have been applied to support XAI [46, 214].

In early work, Toulmin's argumentation model [203] consists of six components: (1) claim, (2) grounds, (3) warrant, (4) backing, (5) rebuttal and (6) qualifier. Specifically, a

*claim* is the main conclusion being asserted, while *grounds* are the evidence used to support the claim. The *warrant* is the reasoning that connects the grounds to the claim, and *backing* provides additional support for the warrant. A *rebuttal* is a counter-argument that challenges the claim, and a *qualifier* indicates the strength of the claim (e.g., "probably" or "definitely"). Toulmin's model is the foundational work in the field of argumentation theory, and therefore many argumentation frameworks have been built on it. The first approach is to consider arguments as *abstract* entities, starting from Dung [56]'s *abstract argumentation (AA)* framework. This work defines the *attack* relation between them, focusing on determining acceptable arguments that are conflict-free. An extension of this framework is the *bipolar argumentation framework (BAF)* [36], which introduces the notion of *support* in addition to attack. Secondly, arguments can be represented as *structured* entities, which are composed of premises and rules. Some examples of structured argumentation frameworks are: *ABA* [24] builds arguments from assumptions and strict rules, and *ASPIC+* [155] constructs arguments from premises, strict and defeasible rules.

In the field of XAI, argumentative frameworks can be used to develop argumentative explanations. Čyras et al. [46] categorised argumentative explanations into two types: (1) intrinsic (i.e., models that are inherently argumentative) and (2) post-hoc (i.e., models that are non-argumentative). For example, Amgoud and Prade [5] used AA to select the best decision based on the acceptance of arguments for and against each decision. The AA framework refers to intrinsic explanations. For post-hoc AF-based explanation, ABA frameworks have been applied to explain decision [236, 241].

### 2.4.3 AI-Assisted Decision-Making Paradigms

In the literature, there are two workflows that are often used in AI-assisted decision-making: (1) AI-first decision-making; and (2) human-first decision-making. AI-first workflow provides the AI recommendation first and then humans decide if they want to accept or not the recommendation, whereas human-first decision-making requires humans to make a provisional decision before they are provided with any AI recommendations.

**AI-first Workflow**

In the AI-first decision-making workflow, it has been demonstrated that participants feel more confident and are also faster in decision-making [60]. These participants also rated AI as more practical. In terms of limitations, the anchoring effect is reported to occur more often in the AI-first workflow [27, 60, 180] in which people overly rely on the AI recommendation (also called *over-reliance*). The anchoring effect [206] refers to giving stronger preference to the earlier knowledge rather than doing a full revision and considering the latest evidence. By contrast, Fogliato et al. [59] did not find any significant difference in the participants' performance, which is measured by accuracy between the two workflows (AI-first and human-first). However, they also found that participants are 65% more likely to revise their answers in the human-first setting than those in the AI-first setting.

**Human-first Workflow**

Human-first decision workflow has been shown to help reduce the over-reliance on erroneous AI recommendations [27, 60]. However, experts may interact with decision-making systems differently from laypeople (crowdworkers). For example, Fogliato et al. [60], Gaube et al. [66] ran studies with radiologists who were the experts. Their task is to review patients' X-ray images. The studies conclude that in human-first workflow with expert participants, they are less likely to leverage AI advice even though the AI is more accurate. In fact, this is referred to *algorithm aversion* [52] or *under-reliance*.

A human-first approach called *cognitive forcing*, based on earlier ideas in psychology for interventions that elicit human thinking at decision-making time [125], has been proposed as a way to improve users' engagement and also increase their learning when interacting with the AI [27, 64]. Four cognitive forcing designs have been introduced: (1) *On demand*: Participants can only see the AI recommendation when they request it; (2) *Update*: Participants first made a decision without seeing the AI recommendation. Then, they were shown the AI prediction and could update their decision later; (3) *Wait*: Participants had to wait for 30 seconds before the AI decision was shown; and (4) *Only*

*AI explanation*: Providing just the AI explanation and *no AI recommendation*, on the basis that this may help people process the AI explanation more carefully and therefore, improve their knowledge and make better decisions [64]. Importantly, *cognitive forcing* has been shown to reduce *over-reliance* compared to the standard AI suggestion approach, although that study had a limitation in that the AI prediction was always correct. There are also some recognised trade-offs of cognitive forcing designs: more time-consuming [64], and less trust [27].

#### Evaluative AI (Hypothesis-driven) Paradigm

Miller [153] argues that AI-assisted decision support is on the cusp of a paradigm shift. This shift is away from the idea of human-first or AI-first, and into a framework he calls **evaluative AI**, which is built around the Data/Frame model [112]. The key insight of evaluative AI is to not necessarily provide a recommendation, and instead to support the human cognitive decision-making process by providing evidence for or against the particular hypothesis that a human decision-maker is considering. This would help to prevent over- and under-reliance, and would help the decision maker to retain their *internal locus on control* [198].

### 2.4.4 Explainable AI (XAI) in Decision Support

Machine Learning (ML) technologies have been used in various domains to support decision-making. They are used in healthcare application [13, 18, 83, 149], financial investment [29, 190], law [12], hiring [133], and even in daily-life tasks such as ingredient detection [94]. There are two methods often used to support decision-making: (1) using uncertainty or confidence measures [211, 239] and (2) using explanation AI techniques [186].

When users' epistemic uncertainty increases (i.e., lack of knowledge about the task increases), prediction rationale (i.e., explaining the link between the inputs and outputs) can be helpful to aid users learning [93] and boost their confidence [242]. By contrast, only showing the confidence score and alternative advice may impede users from accepting

the AI advice as their lack of knowledge about the AI system is high. When people are not given any explanations, they are more likely to agree with the model confidence [39], regardless of its correctness. In this section, we will focus on the role of XAI in supporting decision-making.

#### Human Decision-Making Reliance on AI Support

There is no straightforward position regarding when humans are well calibrated to accept AI-generated advice [215]. Overall, study participants appear more likely to accept an AI's recommendation when provided with explanations, regardless of the model's correctness [16, 30, 86, 212], whereas less detailed explanations can lead to self-reliance [30]. Explanations can increase the accuracy of the human-AI team when the AI is correct, but *decrease* it when it is wrong, resulting in over-reliance on the AI's recommendations. Arguably, this is because the current explanation forms do not provide details of the underlying rationale of the AI model behaviour [212]. We therefore should be careful when selecting the explanation type as it can have a significant effect on whether users decide to rely on them [35, 139]. Moreover, Vasconcelos et al. [213] found that explanations can reduce over-reliance by increasing the task difficulty, easing the explanation difficulty or increasing the benefit of task completion via monetary rewards.

#### Are Current Explanation Approaches Helpful in Decision Making?

Prior research has shown that some explanations are not always helpful in decision-making tasks. For example, neither contrastive rule-based and example-based explanations improve task performance over a baseline of no explanation [212]. LIME feature importance explanations [185] along with the AI's recommendation may not improve users' decision-making compared to only providing the AI's recommendation [3]. Jacobs et al. [86] suggested that feature-based explanations can exacerbate the issues of incorrect recommendations compared to a baseline condition when the explanations are not given. Additionally, even though a counterfactual explanation is considered to be similar to humans' explanations [31], it is shown that counterfactuals may not help trust calibra-

tion [224], and neither factual nor contrastive counterfactual explanations are appropriate in case of incorrect predictions [186].

Furthermore, showing predicted classes greatly improves human performance more than showing only explanations [123]. Human performance can further be improved by showing predicted classes with suggesting high accuracy. When predicted classes are shown, explanations and accuracy induce similar human performance accuracy.

In summary, study participants are more likely to accept the AI's recommendation when provided with explanations, regardless of the model's correctness [16, 86, 212]. Therefore, explanations increase the accuracy of the human-AI team when the AI is correct but *decrease* it when it is wrong, resulting in over-reliance on the AI's recommendations. This is due to the fact the current explanation styles do not provide details of the underlying rationale of the AI model behaviour [212]. We therefore should be careful when selecting the explanation type as it can have a significant effect on whether users decide to blindly trust them. Simple and informative explanations are important to help humans easily find false and unreliable reasoning in the explanation [35, 139]. Importantly, explanations should not be *persuading* in case of incorrect AI recommendations [16].

#### Explanations and Cognitive Engagement

Shang et al. [197] found that users do not often look actively for explanations in low-stake decision-making tasks such as everyday recommendations (e.g. restaurant recommendations). This study also shows that the utility of decisions is a major factor that affects users' needs for counterfactual explanations.

To improve users' engagement and also increase their learning when interacting with the AI, Gajos and Mamykina [64] suggest that providing just the AI explanation and *no AI recommendation* can help people process the AI explanation more carefully and therefore, improve their knowledge and make better decisions. This study however has a key limitation that the AI prediction is always correct. Designing informative explanations in situations of incorrect AI recommendations remains a challenge.

Interactive interfaces can improve users' understanding of how the AI algorithm

works [42] and reduce over-reliance on the AI with *cognitive forcing functions* [27]. However, *cognitive forcing functions* can only improve the performance of human+AI teams compared to only explanations given in situations when the AI prediction is incorrect. There are also some trade-offs when designing interactive interfaces, such as more time-consuming [42] and less trust [27].

#### Designing Explanations in AI-Assisted Decision Making

It is a myth that having more information can lead to better decision-making [110]. In fact, having more information can improve performance to a point. But after that point, more information can cause adverse effects [137]. Additional information can lead to an increase in people's confidence in their decisions, while a decrease in their accuracy [167, 168]. For example, Poursabzi-Sangdeh et al. [178] argue that showing people a *clear* model (i.e., showing model internals that lead to a prediction) impedes them from detecting the model's sizable mistakes. Clear models may be detrimental due to information overload. They suggested that we should let people make their predictions before giving them the model's prediction as it can be helpful for people's comprehension of feature values [106]. Moreover, Lim and Dey [135] suggested allowing users to explore more details *on demand* rather than providing all the information at once.

Wang and Yin [224] summarised three needed factors of AI explanations: (1) improve people's understanding, (2) help people be aware of the AI prediction uncertainty and (3) help people to calibrate their trust appropriately. They evaluated four different explanations (e.g., feature importance, feature contribution, nearest neighbours and counterfactuals). Interestingly, none of the three mentioned factors is satisfied by these four types of explanations when people have limited domain expertise. However, in situations where people have more domain expertise, feature contribution is proved to satisfy the most in three factors. Additionally, we can consider explanation modality in the design by combining visualisations with text and audio explanations [187].

In earlier work, Lim et al. [136] showed that answering the *why* question (i.e., why did the system do X?) resulted in better understanding and perceived trust than answering the *why not* question (i.e., why did the system not do Y?). Lim and Dey [134] found

that the following intelligibility types are recommended in general circumstances: (1) *answering why* as mentioned above; (2) *answering how* (i.e., how does the system do X?); (3) *certainty* (i.e., the system's confidence in its decisions); (4) *control* (i.e., how to change settings or thresholds in the system); and (5) *visualisation* (i.e., providing visualisation of the explanations).

In the context of medical settings, Bussone et al. [30] suggested that showing differential diagnoses is a necessary explanation as it helps users to weigh the positives and negatives of each diagnosis. Along the same lines, Wang et al. [221] proposed a theory-driven conceptual framework for connecting XAI methods with human reasoning, which was applied in the medical domain. They suggested that XAI designs should support *forward reasoning* by showing input feature values and attributions before hypotheses in order to avoid confirmation bias. Furthermore, Cai et al. [34] found that clinicians expect from their AI assistant much like what they do from their colleague, revealing the following desired properties: (1) its *strengths* and *limitations*, particularly in well-known edge cases that humans often make mistakes; (2) its *point-of-view* in relation to their views (e.g., the AI might be more conservative in diagnosis); and (3) its *functionality*, which is the information that the AI has access to and how it uses that information to make a prediction. Moreover, the AI should easily identify case-by-case recommendation confidence such that medical doctors can follow a Bayesian-like procedure to make a better decision (i.e., they are aware of the performance between them and the AI in the past, and also the confidence/uncertainty of the AI recommendation for a specific case) [183].

### 2.4.5 Human and AI Complementary

Zhang et al. [239] argue that a key factor of decision-making systems is to help decision-makers decide if they should trust or not trust the AI model's prediction. In other words, humans can *calibrate trust* and therefore form a *joint decision outcome* with the AI model that leads to improved overall performance than what could be done by only either humans or AI models. A common assumption is that a human-AI team may outperform either the human or AI acting independently. However, existing research challenges this assumption, suggesting that humans interacting with the AI can actually *decrease* the per-

formance compared to the AI model working solo [16, 27, 86]. As a result, helping users to form a mental model of the AI mode's error boundaries is also more important than just improving the AI accuracy. For example, Bansal et al. [15] show that improving the AI accuracy can actually decrease the accuracy in AI-human team performance when the updates in the AI model violate the user's mental model.

In line with the idea of defining the human mental model, Chen et al. [39] define three core concepts to measure understanding in human-AI interaction: *task decision boundary*, *model decision boundary* and *model error*. Task decision boundary separates all instances into correct classes and it represents the ground truth. Model decision boundary represents the AI model prediction which can misclassify some instances. Model error is instances where the AI model predicts incorrectly. Chen et al. show that humans are more likely to accept the AI prediction when they do not use human intuitions in the task prediction. However, in tasks where they can apply their own intuitions, they agree more with the AI prediction when the model explanations are consistent with their own intuitions. In addition, Bansal et al. [14] show that when humans are aware of an AI's error boundary (i.e. the instances where the AI is correct), humans can accept or decline the AI's recommendation; therefore human-AI performance can be optimal. They define two properties of the AI's error boundary, *parsimony* and *stochasticity* that can affect humans' ability when creating their mental model. Parsimony is the complexity of mathematical logic that defines all conditions where the AI model gives incorrect predictions. A non-stochastic error boundary separates strictly between incorrect and correct predictions. Bansal et al. [14] also highlight a property of the task called *dimensionality*, which refers to the number of features used in the data set.

The overlap between the human mental model and the model explanation can affect the human's confidence (or uncertainty) in the model predictions [220]. In Table 2.1, only when there is a large overlap does the confidence in the model prediction increase. On the other hand, when the model explanation is provided but there is a small overlap between the human's prior beliefs and the explanation, their confidence in the final prediction decreases, and the epistemic uncertainty increases. In other words, the human's prior beliefs would lead to *unwarranted confidence* in the explanation evaluations, resulting in

biased decisions. Moreover, humans trust and rely on the AI more when there is perfect complementary in expertise (human and AI are good at different tasks) than when the human and AI have perfect overlap in expertise (both are good at the same task and bad at the same other task) [237].

| | Show Explanation | Not Show Explanation |
|---|---|---|
| Large overlap | Increase confidence | Decrease confidence (Overlapping |
| Small overlap | Decrease confidence | is not applicable) |

Table 2.1: Summary of the findings from Wan et al. [220]. Decrease/Increase the confidence in the AI prediction depending on the overlap between the mental model and the AI explanation.

Besides the human mental model, their confidence and experience with the decision aid can also affect the decision-making process. Zhang et al. [237] argue that trust calibration is not needed in human-AI interaction when humans doing tasks in which they are confident. The human only needs AI recommendations when they have low expertise about the task. In this case, interactive designs can select the information that humans want to know on demand (e.g. uncertainty measures, explanations). Moreover, the experience of interacting with ML models may influence how users want to use the ML recommendation or not. Jacobs et al. [86] found that clinicians who are more familiar with ML are *less* likely to accept the ML recommendation than those who have less experience with ML.

## 2.5 Explainable AI (XAI) in Supporting Skin Cancer Diagnosis

We will now review the literature on the role of explainable AI in supporting skin cancer diagnosis.

Applying AI in supporting skin cancer detection has become more prevalent and potentially improved decision-making accuracy. The sensitivity of dermatologists who correctly diagnose melanoma rarely exceeds 80%, and general practitioners have a much lower sensitivity [235]. The examination often depends on the dermatologists' experience and can be subjective. In clinical settings, AI is still not widely applied due to the

lack of trust in the system, but some technologies are used to aid doctors. For example, a user study has found that the majority of healthcare practitioners (58%) who are based mostly in Australia, have used mobile dermatoscopy for lesion monitoring and record keeping [90]. More research in applying AI in supporting skin cancer diagnosis is needed to build trust in the system. With the help of AI, we aim to aid dermatologists in making more accurate diagnoses.

A common task is to apply deep learning to classify between benign and malignant images. For example, Esteva et al. [58] trained a CNN model that has comparable dermatologist-level competency. Tschandl et al. [205] showed that human-AI collaboration can improve diagnosis accuracy over that of either human-alone or AI. Although AI can benefit non-expert clinicians, wrong AI recommendations can mislead people who have different levels of clinical expertise, including experts [205]. Barata et al. [17] used a reinforcement learning model to incorporate human preferences into the decision-aid algorithms.

To provide more information than only an AI-recommended diagnosis, XAI has been applied to explain further the AI's decision-making process. Explanations need to be grounded in users' goals and needs. For instance, XAI can extract meaningful concepts from the dermatoscopic images [47, 73, 141, 171, 231], which are referred to the concept-based explanations that have been discussed previously in Section 2.2. Chanda et al. [37] showed that Grad-CAM [194] can improve trust in the AI system significantly compared to the system without explanations. The results also found a strong alignment between the Grad-CAM method and explanations from dermatologists. However, despite improving trust, Grad-CAM did not significantly increase diagnostic accuracy compared to AI recommendation alone.

Different types of explanations can influence the effectiveness of the decision aid. Schoonderwoerd et al. [193] investigated different post-hoc local explanations in a controlled environment of child health diagnosis. Clinicians evaluated different interface designs using different types of explanations, including, general information about patients, evidence that supports or contradicts the diagnosis, contrastive explanations, counterfactual explanations, case-based explanations and the certainty of the diagnosis. The study

found that the following information elements are rated as highly important: the information being used to make the diagnosis by the system, evidence for and against the diagnosis, certainty level and how to increase the certainty of the diagnosis, and the performance of the system. Moreover, study participants consistently agreed that including case-based explanations by showing a typical case with the same diagnosis and how that relates to the current case, would improve their decision-making. Additionally, Tschandl et al. [205] study different interaction modalities, including: (1) *AI-based multiclass probabilities*: People see all probabilities of all classes, calculated by the AI; (2) *AI-based probability of malignancy*: People only see the probability of malignant diagnosis; (3) *AI-based content-based image retrieval (CBIR)*: People see a few example of images that have similar diagnosis, so-called *example-based explanations*; (4) *Crowd-based multiclass probabilities*: The probabilities are collected from human raters for each diagnosis. They found that AI-based multiclass probabilities outperformed others.

# Chapter 3
# Explaining the Uncertainty

*Displaying confidence scores in human-AI interaction has been shown to help build trust between humans and AI systems. However, most existing research uses only the confidence score as a form of communication. As confidence scores are just another model output, users may want to understand why the algorithm is confident to determine whether to accept the confidence score. In this chapter, we show that counterfactual explanations of confidence scores help study participants to better understand and better trust a machine learning model's prediction. We present two methods for understanding model confidence using counterfactual explanation: (1) based on counterfactual examples, and (2) based on visualisation of the counterfactual space. Both increase understanding and trust for study participants over a baseline of no explanation, but qualitative results show that they are used quite differently, leading to recommendations of when to use each one and directions for designing better explanations.*

## 3.1 Introduction

EXPLAINING why an AI model gives a certain prediction can promote trust and understanding for users, especially for non-expert users. While recent research [222, 239] has used confidence (or uncertainty) measures as a way to improve AI model understanding and trust, the area of explaining why the AI model is confident (or not confident) in its prediction is still underexplored [202].

This chapter is based on the following published paper:
**[C1] (AAAI23 Main Track) [127]** Thao Le, Tim Miller, Ronal Singh, Liz Sonenberg. "Explaining Model Confidence Using Counterfactuals." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 10, pp. 11856-11864. 2023.

In Machine Learning (ML), the *confidence score* indicates the chances that the ML model's prediction is correct. In other words, it shows how *certain* the model is in its prediction, which can be defined as the predicted probability for the best outcome [239]. Another way to define the *confidence score* is based on uncertainty measures, which can be calculated using entropy [21] or using *uncertainty sampling* [132], [157, p93].

In this chapter, we complement prior research by applying a counterfactual (CF) explanation method to generate explanations of the confidence of a predicted output. It is increasingly accepted that explainability techniques should be built on research in philosophy, psychology and cognitive science [31, 150]. Specifically, we focus on counterfactual explanations because they align with how humans seek explanations in daily life [32, 150]. That is, people often ask about counterfactuals rather than factual ones when they want to understand why an event happened. Moreover, the evaluation process of explanation should involve human-subject studies [61, 103, 154, 212]. We, therefore, evaluate our explanation to know whether counterfactual explanations can improve *understanding*, *trust*, and *user satisfaction* in two user studies using existing methods for assessing understanding, trust and satisfaction. We present the CF explanation using two designs: (1) providing counterfactual examples (example-based counterfactuals); and (2) visualising the counterfactual space for each feature and its effect on model confidence (visualisation-based counterfactuals).

Our contributions are:

- We formalise two approaches for the counterfactual explanation of confidence score: one using counterfactual examples and one visualising the counterfactual space.

- Through two user studies we demonstrate that showing counterfactual explanations of confidence scores can help users better understand and trust the model.

- Using qualitative analysis, we observe limitations of the two explainability approaches and suggest directions for improving presentations of counterfactual explanations.

## 3.2 Formalising Counterfactual Explanation of Confidence

This section describes two methods for CF explanation: one based on counterfactual examples [7] and one based on counterfactual visualisation as in Figure 3.1.

### 3.2.1 Generating Counterfactual Explanation of Confidence

In this section, we show how to generate counterfactual explanations of the confidence score in data where input variables can take either categorical or continuous values. The tool is similar to regression counterfactuals; however, we focus on testing the ability to explain confidence, not regression. The counterfactual model can generate explanations to either increase or decrease the confidence score of a specific class. For example, when the AI model predicts that an employee will leave the company with confidence of 70%, a person may ask: *Why is the model 70% confident instead of 40% confident or less?*. This person could ask why the model prediction did not have a lower confidence score when they were sceptical about the high confidence score. We aim to generate counterfactual inputs that bring the confidence score to 40% or lower. An example of counterfactual explanation, in this case, is: *"One way you could have got a confidence score of 40% instead is if Daily Rate had taken the value 400 rather than 300"*. Therefore, from this counterfactual explanation, we know that we can achieve lowering of the confidence in them resigning from the company by increasing the employee's daily rate.

We now describe our approach to generate counterfactuals for confidence scores. We follow [189] in proposing an algorithm to search for counterfactual points of output confidence. Importantly, we modify this approach to find counterfactual points that change the confidence score but do not change the predicted class.

Formally, given a question: "Why does the model prediction have a confidence score of $U(x)$ rather than greater than (or less than) $T$?" where $T$ is a user-defined confidence threshold, $x$ is the input instance, $U(x)$ is the confidence score of the original prediction, we want to find the counterfactual explanation of confidence $U(x')$ generated by data point $x'$ such that $U(x') > T$ or $U(x') < T$ depending on the question. In case the user cannot give a threshold T, the default threshold T value is the original confidence score

$U(x)$ of the prediction. We seek the counterfactual point $x'$ by solving Equation 3.1:

$$\arg\min_{x'}(||x - x'||_{1,w} + |U(x') - T|) \tag{3.1}$$

such that:

$$U(x') > T \qquad \text{if } T > U(x) \tag{3.2}$$

$$U(x') < T \qquad \text{if } T < U(x) \tag{3.3}$$

$$\begin{cases} P(y = k \mid x') < D & \text{if } P(y = k \mid x) < D \\ P(y = k \mid x') \geq D & \text{if } P(y = k \mid x) \geq D \end{cases} \tag{3.4}$$

where $||.||_{1,w}$ is a weighted $l_1$ norm with weight $w$ defined as the inverse median absolute deviation (MAD) [218]; $D$ is the decision boundary that classifies the class.

We apply Equation 3.2 when we want to find counterfactual $x'$ that increases the confidence score, and Equation 3.3 for a counterfactual $x'$ that decreases the confidence score. Since $x$ and $x'$ will give the same output prediction as class $k$ but different confidence scores $U(x)$ and $U(x')$, $P(x)$ and $P(x')$ must be in the same space according to the decision boundary, defined as Equation 3.4.

### 3.2.2 Example-Based Counterfactual Explanation

Given the original instance input shown in column *Original Value* in Table 3.1, the AI model predicts that this person has an income of *Lower than* $\$50,000$ with a confidence score of 57.8%. Note that in this example, the user chooses the threshold $T = 45\%$. A counterfactual input $x'$ is then searched for such that $U(x') < T$. An example of a counterfactual explanation generated using our method is: *"One way you could have got a confidence score of less than* 45% *(*30.1%*) instead is if Occupation had taken value Manager rather than Service."*

We presented counterfactuals in a table, such as in Table 3.1. We show the details of a person in column *Original Value* and the prediction that their income is lower than $\$50,000$. When we change the value of feature *Occupation* as in columns *Alternative 1*

| Attribute | Alternative 1 | Alternative 2 | Original |
|---|---|---|---|
| Marital status | - | - | Married |
| Years of education | - | - | 9 |
| **Occupation** | **Manager** | **Skilled Specialty** | **Service** |
| Age | - | - | 63 |
| Any capital gains | - | - | No |
| Working hours per week | - | - | 12 |
| Education | - | - | High School |
| **Confidence score** | **30.1**% | **42.1**% | **57.8**% |
| **AI prediction** | **Lower than $50,000** | | |

Table 3.1: Example-based counterfactual explanation presented in a table. In alternative columns, notation (-) means the value is unchanged from the original value, we only highlight the values that changed.

and *Alternative 2*, the confidence score changes but the prediction is still lower than $\$50,000$. From this table, we can find the correlation between the *Occupation* and the confidence score; the occupation *Service* gives the prediction with the highest confidence score among all three occupations.

### 3.2.3 Visualisation-Based Counterfactual Explanation

In this section, we propose a method for visualising the counterfactual space of a model and how this affects the model's confidence as shown in Figure 3.1 and 3.2. The idea is to visualise how varying a single feature affects the model's confidence, relative to the factual input $x$. For example, Figure 3.1 shows the visualisation based on Table 3.1 in the income prediction task. Here we can see the prediction reaches maximum confidence score at *Service* occupation. The title of this graph shows the output prediction *Lower than* $\$50,000$ and the feature name *Occupation* which we used to change the values.

This visualisation technique is based on the idea of **Individual Conditional Expectation (ICE)** [69]. ICE is often used to show the effect of a feature value on the predicted probability of an instance. In our study, we show how changing a feature value can change the *confidence score* instead of changing the predicted probability as in the original ICE. There are two types of variables in the dataset: (1) categorical variable, and (2)

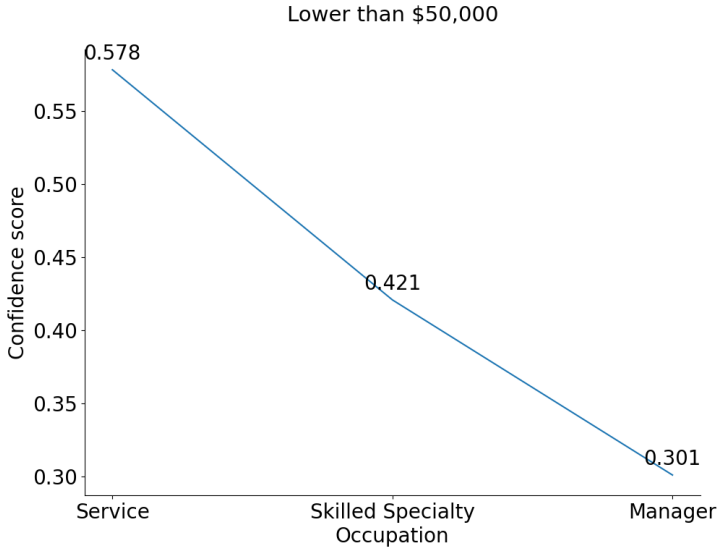



Figure 3.1: Counterfactual visualisation: Categorical variable

| | **Control (C)** | **Example-Based (E)** | **Visualisation-Based (V)** |
|---|---|---|---|
| Phase 1 | Participants are given plain language statement, consent form and demographic questions (age, gender) | | |
| Phase 2 | Participants are provided with | | |
| | Input instances;<br>AI model's prediction class. | Input instances;<br>AI model's prediction class;<br><br>Counterfactual examples. | Input instances;<br>AI model's prediction class;<br><br>Counterfactual visualisation. |
| Phase 3 | Nothing | 10-point *Explanation Satisfaction rating scale* | |
| Phase 4 | 10-point *Trust rating scale* | | |

Table 3.2: Summary of participants' tasks in our three experimental conditions

continuous variable. So we define the ICE for confidence score of a single feature $x_i$ of instance $x$ such that $F(x_i) = U(x_i)$ for all $x_i$, where:

- $x_i \in D$ if $x_i$ is a categorical value and $D$ is the categorical set
- $x_i \in [c_{\min}, c_{\min} + t, \ldots, c_{\max}]$ if $x_i$ is a continuous value; $c_{\min}$ and $c_{\max}$ are the minimum and maximum values of a continuous range and $t$ is a fixed increment.

If we use only a 2-dimensional graph, we can visualise the changes of only one feature, whereas counterfactual examples can explain how changing multiple features simultaneously affect confidence. However, visualising the counterfactual space allows us

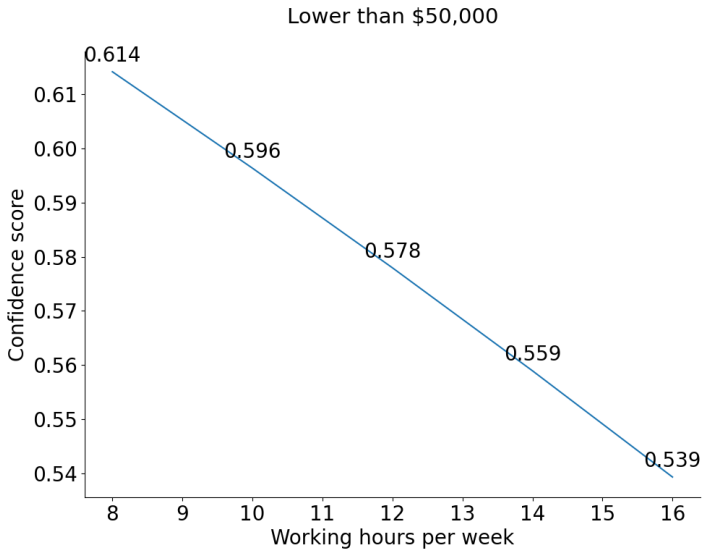



Figure 3.2: Counterfactual visualisation: Continuous variable

to easily identify the lowest and highest confidence values for categorical values and the trend of continuous values.

## 3.3 Human-Subject Experiments

Our user experiments test the following hypotheses.

- **Hypothesis 1a/b (H1a/b)**: **Example-based/Visualisation-based** counterfactual explanations help users better **understand** the AI model than when they are not given explanations.

- **Hypothesis 2a/b (H2a/b)**: **Example-based/Visualisation-based** counterfactual explanations help users better **trust** the AI model than when they are not given explanations.

It is necessary to test against the baseline of no explanation because providing explanations is not always useful compared to not providing any explanations [16, 123]. We then evaluate the difference between example-based counterfactual explanations and visualisation-based counterfactual explanations based on hypotheses H3a/b/c. Previous

research has shown that visual explanations are more effective than textual ones, making them a powerful learning tool [23].

- **Hypothesis 3a/b/c (H3a/b/c)**: *Visualisation-based* counterfactual explanations help users better **understand/trust/be satisfied with** the AI model than *example-based* counterfactual explanations.

To evaluate **understanding**, i.e., H1a, H1b and H3a, we use *task prediction* [81, p11]. Participants are given some instances and their task is to decide for which instance the AI model will predict a higher confidence score. Thus, task prediction helps evaluate the user's mental model about their understanding of model confidence.

To evaluate **trust**, i.e., H2a, H2b and H3b, we use the 10-point *Trust rating scale*. For **satisfaction**, i.e., H3c, we use the 10-point *Explanation Satisfaction rating scale*. Unlike the Likert scale used in [81], the 10-point rating scale in our user study is continuous, with a midpoint of 5.5.

### 3.3.1 Experimental Design

#### Dataset

We ran the experiment on two different domains from two different datasets, which are *income prediction domain* and *HR domain*. Both datasets are selected so that experiments can be conducted on general participants with no requirement of particular expertise. The data used for the income prediction task is the Adult Dataset published in UCI Machine Learning Repository [55] that includes 32561 instances and 14 features. This dataset classifies a person's income into two classes (below or above $50K) based on personal information such as marital status, age, and education. In the second domain, we use the IBM HR Analytics Employee Attrition Performance dataset published in Kaggle [172], which includes 1470 instances and 34 features. This dataset classifies employee attrition as yes or no based on some demographic information (job role, daily rate, age, etc.). We selected the seven most important features for both datasets by applying the Gradient Boosting Classification model over all data.

| **Attribute** | **Employee 1** | **Employee 2** | **Employee 3** |
|---|---|---|---|
| Marital status | Married | Married | Married |
| Years of education | 15 | 15 | 15 |
| Occupation | **Service** | **Manager** | **Skilled Specialty** |
| Age | 25 | 25 | 25 |
| Any capital gains | No | No | No |
| Working hours per week | 30 | 30 | 30 |
| Education | Bachelors | Bachelors | Bachelors |
| **AI model prediction** | **Lower than \$50,000** | | |

Table 3.3: Example input instances provided in the question. The question is: "For which employee the AI model predicts with the highest confidence score?"

### Model Implementation

In our experiments, we use logistic regression to calculate the probability of a class, so $P(x) = \frac{1}{1+e^{-y}}$ where $y = wx$ is a linear function of point $x$. We chose logistic regression because of its simplicity so that we can easily define the confidence score. Moreover, although logistic regression models are considered intrinsically interpretable models [156], it is still challenging to reason about their behaviour when we want to have a lower (or higher) confidence score. In future work, our studies can be extended to using counterfactual tools for more complex models, such as CLUE [7].

We choose *margin of confidence*, which is the difference between the first and the second highest probabilities [157, p93] as the formula of confidence score $U(x)$. The higher the difference between two class probabilities, the more confident the prediction is in the highest probability class.

### Procedure

Before conducting the experiments, we received ethics approval from our institution (ID: 23208). We recruited participants on Amazon Mechanical Turk (Amazon MTurk), a popular crowd-sourcing platform for human-subject experiments [28]. The experiment was designed as a Qualtrics survey[1] and participants can navigate to the survey through the

[1] https://www.qualtrics.com/

Amazon MTurk interface. We allowed participants 30 minutes to finish the experiment and paid each participant a minimum of USD $7 for their time, plus a maximum of up to USD $2 depending on their final performance.

We use a between-subject design such that participants were randomly assigned into one of three groups: (1) *Control (C)*; (2) *Treatment with Example-Based Explanation (E)*; or (3) *Treatment with Visualisation-Based Explanation (V)*. For each group, there are four phases that are described in Table 3.2. The difference between the control group and the treatment group is that in the control group, participants were not given any explanations. In the task prediction (phase 2), participants in the control group were only shown input values along with the AI model prediction class as in Table 3.3. In the treatment group, participants were provided with either example-based explanations (e.g. Table 3.1) or visualisation-based explanations (e.g. Figure 3.1). The participants each received the same 10 questions. For each question, they were asked to select an input instance out of *3 instances* that the AI model would predict with the highest confidence score. A question can have either one or two explanations depending on the number of modified attributes in the question. For instance, the question in Table 3.3 changes only one attribute *Occupation* so participants were given a single explanation of treatment conditions. An explanation can either present a *categorical variable* (e.g. Figure 3.1) or a *continuous variable* (e.g. Figure 3.2).

We scored each participant using: 1 for a correct answer, -2 for a wrong answer and 0 for selecting "I don't have enough information to decide". To imitate high-stake domains, the loss for a wrong choice is higher than the reward for a correct choice [15, p2433]. They are also asked to briefly explain why they chose that option in a text box, which is analysed later in the qualitative analysis. The final compensation was calculated based on the final score — a score of 0 or less than 0 received $7 USD and no bonus. A score greater than 0 received a bonus of $0.2 for each additional score.

#### Participants

We recruited a total of 180 participants for two domains, that is 90 participants for each domain from Amazon MTurk. Then 90 participants were evenly randomly allocated

into three groups (30 participants in each group). All participants were from the United States. We only recruited Masters workers, who achieved a high degree of success in their performance across a large number of Requesters [2]. For the *income prediction domain*, 41 participants were women, 1 was self-specified as non-binary, and 48 were men. Between them, 4 participants were between Age 18 and 29, 34 were between Age 30 and 39, 27 were between Age 40 and 49, and 25 were over Age 50. For the *HR domain*, 43 participants were women, 47 were men. Age-wise, 4 participants were between Age 18 and 19, 37 were between Age 30 and 39, 26 were between Age 40 and 49, and 23 were over Age 50.

We performed *power analysis* for two independent sample t-tests to determine the needed sample sizes. We calculate Cohen's d between the control and treatment group and obtain the effect size of 0.7 and 0.67 in the income and HR domain. Using a power of 0.8 and significant alpha of 0.05, we get sample sizes of 26 and 29 in the two domains. Thus, we determine the sample size needed for a group is 30 and the total number of samples needed is 90 for one domain.

### 3.3.2 Results: Summary of Both Domains

In this section, we present the results from our experiment for two domains that used the income and HR datasets. We tested for data normality by using the Shapiro-Wilks test and found that our data was not normally distributed. Therefore, we applied the Mann–Whitney U test, which is a non-parametric test equivalent to the independent samples t-test to perform pairwise comparisons between two groups. Table 3.4 summarises our results of testing the seven hypotheses. Figure 3.3 and 3.4 show the results of the two studies.

**The results show that counterfactual explanations of confidence scores help users understand and trust the AI model more than those who were not given counterfactual explanations**. We conclude that **H1a, H1b, H2a** and **H2b** are supported in both studies ($p \leq 0.005, r > 0.21$).

**There is no statistically significant difference in improving users' understanding between example-based explanations and visualisation-based explanations — H3a** is

[2] https://www.mturk.com/worker/help

| Measure | Hypothesis | Domain 1 (Income) | Domain 2 (HR) |
|---|---|---|---|
| Understanding | H1a (Control vs. Example) | ✓($p = 0.005, r = 0.42$) | ✓($p < 0.001, r = 0.85$) |
| | H1b (Control vs. Visualisation) | ✓($p < 0.001, r = 0.62$) | ✓($p < 0.001, r = 0.87$) |
| | H3a (Example vs. Visualisation) | ×($p = 0.13, r = 0.23$) | ×($p = 0.86, r = 0.03$) |
| Trust | H2a (Control vs. Example) | ✓($p < 0.001, r = 0.21$) | ✓($p < 0.001, r = 0.34$) |
| | H2b (Control vs. Visualisation) | ✓($p < 0.001, r = 0.51$) | ✓($p < 0.001, r = 0.43$) |
| | H3b (Example vs. Visualisation) | ✓($p < 0.001, r = 0.26$) | ×($p = 0.10, r = 0.09$) |
| Satisfaction | H3c (Example vs. Visualisation) | ✓($p < 0.001, r = 0.28$) | ×($p = 0.06, r = 0.10$) |

Table 3.4: Summary of hypothesis tests in two domains. ✓ represents the hypothesis is supported, × represents the hypothesis is rejected. Since we use the Mann-Whitney U test, we report the effect size $r$ as the rank-biserial correlation.

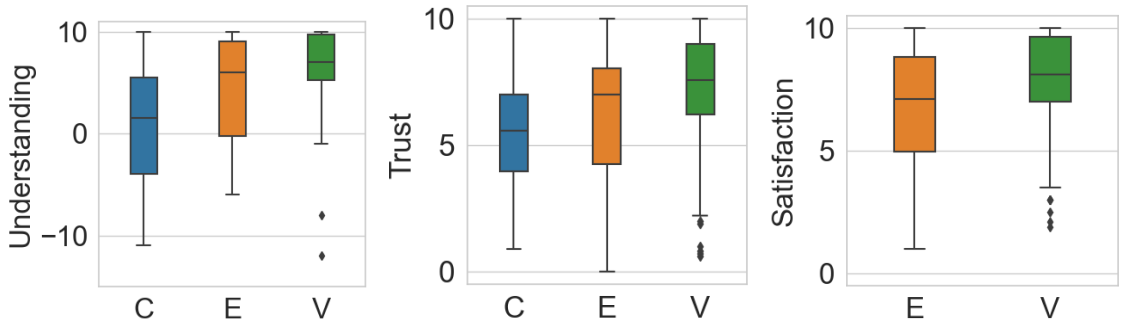



Figure 3.3: Domain 1 (Income). *C = Control*; *E = Example-Based Explanation*; *V = Visualisation-Based Explanation*.

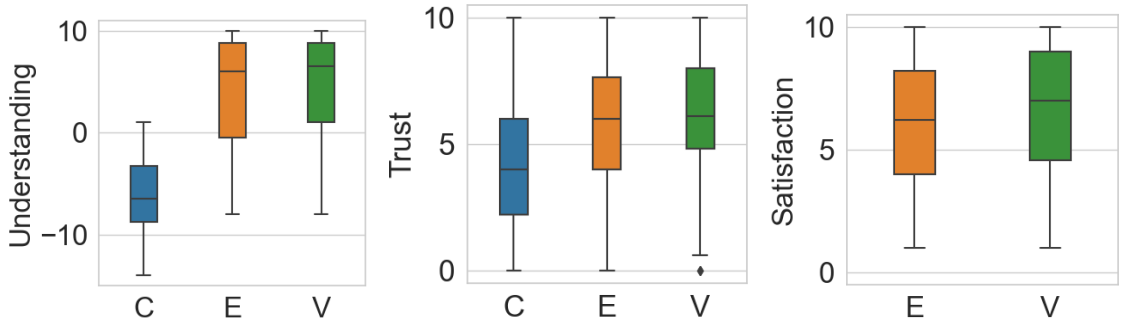



Figure 3.4: Domain 2 (HR). *C = Control*; *E = Example-Based Explanation*; *V = Visualisation-Based Explanation*.

rejected. In domain 1, the difference in the task prediction between the two treatment groups is larger than that in domain 2. Specifically, the effect size in domain 1 is $r = 0.23$

| Code | Definition |
|---|---|
| **W-Reversed category** (CAT) | The participant selected the instance that has the lowest confidence score instead of the highest confidence score among all instances |
| **W-Linear assumption** (CAT) | Assumed the correlation between confidence score and attribute values was linear when it was not (e.g. the feature was categorical) |
| **W-Small differences** (CAT & CON) | Selected a wrong answer due to small differences in the explanation and/or the question |
| **W-Reversed correlation** (CON) | Reversed the trend of the explanation of a continuous variable |
| **W-Case-based** (CON) | Used case-based reasoning when the correlation was linear |
| **D-No correlation** (CAT & CON) | Could not find the trend of the confidence score |
| **D-Different attribute values** (CAT & CON) | Argued that the values of instances in the explanations are not the same as values in the question |
| **D-Outside range** (CON) | The modified values in the question are beyond the lowest and highest values in the explanation |
| **C-Correlation-based** (CON) | Found the correlation in the explanation |
| **C-Case-based** (CAT) | Got the correct answer based on examples in the explanation without mentioning the correlation |

Table 3.5: The codebook for participants' responses to evaluate how they understand the provided explanations. *CAT*, *CON* mean the code is applied for categorical variables and continuous variables, respectively. *W* corresponds to wrong answers. *D* corresponds to the "do not have enough information to decide". *C* corresponds to correct answers.

($p = 0.13$) and in domain 2 is $r = 0.03$ ($p = 0.86$).

**There are some discrepancies between domains 1 and 2 when comparing example-based and visualisation-based explanations in terms of trust and satisfaction**. In the first domain, **H3b** ($p < 0.001, r = 0.26$) and **H3c** ($p < 0.001, r = 0.28$) are supported. However, in domain 2, **H3b** ($p = 0.1 > 0.05$) and **H3c** ($p = 0.06 > 0.05$) are both rejected. We envision the discrepancies between H3b and H3c may be because prior knowledge of participants could affect them doing the tasks in two different domains. Future work could test this idea further.

As observing no statistically significant difference between example-based and visualisation-based explanations, we then used qualitative analysis to find the limits of both designs and suggest directions to design effective explanations.

### 3.3.3 Qualitative Analysis

| | | | Income | | | | HR | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Example | | Visualisation | | Example | | Visualisation | |
| Wrong Answer | Categorical Variables | W-Linear assumption | 0 | (0%) | 0 | (0%) | 39 | (95%) | 21 | (78%) |
| | | W-Small difference | 5 | (28%) | 0 | (0%) | 2 | (5%) | 2 | (7%) |
| | | W-Reversed category | 13 | (72%) | 11 | (100%) | 0 | (0%) | 4 | (15%) |
| | Continuous Variables | W-Case-based | 1 | (4%) | 0 | (0%) | 7 | (64%) | 0 | (0%) |
| | | W-Small difference | 0 | (0%) | 2 | (18%) | 0 | (0%) | 2 | (12%) |
| | | W-Reversed correlation | 23 | (96%) | 9 | (82%) | 4 | (36%) | 14 | (88%) |
| Not Enough Information | Categorical Variables | D-Different attribute values | 6 | (100%) | 0 | | 8 | (80%) | 0 | |
| | | D-No correlation | 0 | (0%) | 0 | | 2 | (20%) | 0 | |
| | Continuous Variables | D-Outside range | 1 | (6%) | 9 | (64%) | 2 | (13%) | 6 | (46%) |
| | | D-Different attribute values | 4 | (24%) | 0 | (0%) | 3 | (19%) | 0 | (0%) |
| | | D-No correlation | 12 | (70%) | 5 | (36%) | 11 | (68%) | 7 | (54%) |
| Correct Answer | Categorical Variables | C-Correlation-based | 17 | (11%) | 20 | (11%) | 18 | (10%) | 0 | (0%) |
| | | C-Case-based | 133 | (89%) | 159 | (89%) | 157 | (90%) | 186 | (100%) |
| | Continuous Variables | C-Correlation-based | 97 | (98%) | 118 | (100%) | 81 | (98%) | 99 | (100%) |
| | | C-Case-based | 2 | (2%) | 0 | (0%) | 2 | (2%) | 0 | (0%) |

Table 3.6: Frequencies and Percentages of Codes for Explanations

We perform the thematic analysis [26] from the text written by participants after each multiple-choice question to know why they selected an option. The text is a response to "Can you please explain why you selected this option?". We followed Nowell et al. [164] who gave a step-by-step approach for doing trustworthy thematic analysis. Three authors were involved in the qualitative analysis. The first author identified and documented the themes and the codes. Through multiple discussion meetings, two other authors critically analysed the codes and verified them. Finally, we decided on the final codes as in Table 3.5.

Every participant did the same 10 questions so we have 30 (participants) $\times$ 10 (questions) is 300 (texts) for a condition. Given that we have two treatment conditions and two datasets, we analysed a total of 1,200 texts and each text is assigned to one code or more than one code depending on the number of explanations in that response.

Each code is classified as one of: (1) a correct answer (C); (2) a wrong answer (W); or (3) "not enough information" (D). The final analysis includes 1,112 texts after removing

88 texts due to poor quality. We found the following observations, which suggest future improvements.

**Use text labels instead of numbers to present categorical variables.**. A categorical variable can be shown in numbers or text labels. In Table 3.6, the majority of wrong codes in the HR domain is *W-Linear assumption* (78% and 95%) because most explanations using categorical variables are written in numbers. There were no *linear assumption* codes in the income dataset since all explanations used text labels.

**When the labels of categorical features indicate ordinal data, visualise counterfactuals help to reduce the error "linear assumption", making it easier for people to interpret the highest or lowest values**. According to Table 3.6 (HR dataset), 95% (39) of wrong responses happened due to *linear assumption* in the example-based condition; however, we found only 78% (21) of *linear assumption* in the visualisation-based condition. For instance, in a question where the job level is a categorical variable and is not correlated with the confidence score, a participant in the example-based condition mentioned: *"Those with a higher job level had a higher confidence rating"*. In contrast, another participant in the visualisation-based condition could identify the highest confidence value at job level 2 without mentioning the linear trend: *"The AI predicted job level 2 has the highest chance of staying"*.

**It is hard for people to interpret the example-based explanations when the differences between counterfactual outputs and categorical attributes are minimal**. According to Table 3.6 (wrong answer, income dataset, categorical variables), we observe 28% (5) of *W-Small difference* codes in the example-based condition. For example, in a question where *Manager* occupation has the highest confidence score, some participants mistakenly selected *Skilled Specialty* as the highest even though this occupation is the second highest. In this case, the difference in confidence values between *Manager* and *Skilled Specialty* is only 2% (93% and 91%). This small difference made 5 participants choose a wrong answer in the example-based condition.

**Using visualisation-based explanations makes it easier to understand correlations; however, many participants were not willing to extrapolate the correlation beyond the lowest and highest values**. In Table 3.6 (Not Enough Information), we have fewer codes

of *D-No correlation* in visualisation-based explanations. However, we record a higher number of codes of *D-Outside range* in this visualisation condition. This issue suggests that we should not expect participants to extrapolate the correlation, and all counterfactual points should be shown in the explanations.

**Regardless of variables, if the counterfactual examples in the example-based explanations are not the same as the values in the question, many participants argued that they do not have enough information to decide** (*D-Different attribute values*). For example, a participant said: *"Because the position is different, lab tech versus sales rep, I feel that even though the AI chose the one with the highest confidence as the one with the lowest daily rate, I am not sure if the job description would change that confidence level"*. In this question, we provided the example-based explanation for *Sales Representative* job, but the question shows instances of *Lab Tech* job. Even though the daily rate increases linearly in all cases, some participants did not feel confident to apply this observation when we changed the instance values in the question. They applied *case-based reasoning* when interpreting the example-based explanation of a linear model rather than interpreting the linear correlation in this explanation. That is, they found the closest example in the counterfactual explanation presented and compared that example with the question. Similarly, we found an overall 8 codes of **W-Case-based** where participants applied case-based reasoning to do the task with example-based explanations of continuous values. A participant wrote: *"It really is a tough call but I chose employee 1 because the 400 range has the highest percentage of leaving"*. In this example, the participant saw that the daily rate of 400 has the highest confidence of leaving, therefore, they selected the value that is close to 400 rather than interpreting the linear correlation between the daily rate and the confidence score (lower daily rate indicates higher confidence of leaving). Specifically, the question has three daily rate options of 200, 201 and 247, they eventually selected 247 as the final answer, arguing that 247 is closest to 400 in the example-based explanation. In general, **it is clear that participants in the example-based condition used a 'case-based reasoning' approach to understanding the model**. This led participants to overlook the linear trend between the confidence score and the feature values. This finding suggests that we should be careful when using example-based explanations to interpret continuous

variables for models, except for cases when the underlying model is itself a case-based model. Using graphs to visualise continuous variables can mitigate this issue.

## 3.4 Limitations and Future Work

We follow the *uncertainty sampling* approach [132] to calculate the uncertainty. However, we have not considered *calibrated uncertainty* [21, 118], which represents the true uncertainty of the output [3]. Future work can explore how we can measure and explain the true uncertainty. Furthermore, the current human experiment was only conducted on a logistic regression model. Potential future work should consider more complex non-linear models.

Regarding the explanations, we can explore other XAI methods such as causal inference [173] to understand how inputs influence model confidence, where it is important to distinguish between correlation and causation. Moreover, we can extend our model by generating contrastive explanations based on the current counterfactual approach. We also acknowledge that the current explanations can be misleading, even though counterfactual explanations are often considered faithful because they are not post-hoc approximations. Exploring how users respond to incorrect explanations could yield valuable insights into human-AI interaction under uncertainty.

## 3.5 Conclusion

This chapter formalises counterfactual explanations of model confidence and studies two approaches: (1) example-based counterfactuals; and (2) visualisation-based counterfactuals. Through human-subject studies, we show that the counterfactual explanation of model confidence can help users improve their understanding and trust in the AI model. Finally, the qualitative analysis suggests directions for designing better counterfactual explanations.

[3] mentioned in Section 2.1

# Chapter 4

# Hypothesis-Driven Decision-Making Model

*Prior research on AI-assisted human decision-making has explored several different explainable AI (XAI) approaches. A recent paper has proposed a paradigm shift calling for hypothesis-driven XAI through a conceptual framework called evaluative AI that gives people evidence that supports or refutes hypotheses without necessarily giving a decision-aid recommendation. In this chapter, we describe and evaluate an approach for hypothesis-driven XAI based on the Weight of Evidence (WoE) framework, which generates both positive and negative evidence for a given hypothesis. Through human behavioural experiments, we show that our hypothesis-driven approach increases decision accuracy and reduces reliance compared to a recommendation-driven approach and an AI-explanation-only baseline, but with a small increase in under-reliance compared to the recommendation-driven approach. Further, we show that participants used our hypothesis-driven approach in a materially different way to the two baselines.*

## 4.1 Introduction

RESEARCH has shown that AI recommendations, even when accompanied with explanations, are not always helpful in supporting decision-making [16, 27, 64, 225]. The direct causes of this are *under-reliance* and *over-reliance* [215]. With under-reliance,

This chapter is based on the following published paper:
**[C2] (ECAI24 Main Track) [128]** Thao Le, Tim Miller, Liz Sonenberg, Ronal Singh. "Towards the New XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence". In *In 27th European Conference on Artificial Intelligence*, vol. 392, pp. 850-857. 2024.

decision-makers reject AI recommendations, even when they may be correct. Alternatively, decision-makers may overly rely on AI recommendations, hence be led to errors when the AI is incorrect. In either case, they tend to fixate on a particular hypothesis without sufficiently considering others [153]. Approaches such as *cognitive forcing*, based on ideas from human psychology, have been proposed to address limitations of the AI recommendation approach [27], with recent work indicating that withholding AI model the recommendations, at least for a short time, while still providing the user with an explanation of that recommendation, can be helpful [64]. Recently, Miller [153] proposed a so-called **hypothesis-driven** decision-making paradigm called **evaluative AI**. The main aim of this paradigm is to focus the decision loop on the human decision-maker, providing them with the right evidence to support their own intuitions, rather than focusing the decision loop on machine recommendations. This paradigm offers a promising direction in building better decision support in explainable AI (XAI) research by focusing on human decision-makers considering multiple possible hypotheses.

In this chapter, we describe and evaluate an approach for building a hypothesis-driven decision-making model that uses the *Weight of Evidence (WoE)* framework [148]. To the best of our knowledge, this is the first work that compares empirically and in a controlled manner the hypothesis-driven approach [153] with two other popular decision-making approaches (recommendation-driven and AI-explanation-only [64]). Our contributions are:

- The *Evidence-Informed Hypothesis-Driven Decision-Making* model, building on the *Weight of Evidence (WoE)* framework to the hypothesis-driven approach;
- Two human behavioural experiments comparing our *hypothesis-driven* approach with two common decision-aid approaches: (1) the standard model recommendation with explanation; and (2) a form of cognitive forcing that provides only AI explanations [64]. The results show that the hypothesis-driven approach improves decision accuracy and reduces over-reliance compared to standard recommendation-driven approaches, at the cost of a slight increase in under-reliance. Furthermore, the hypothesis-driven approach reduces under-reliance significantly compared to the AI-explanation-only approach. Our qualitative analysis identifies some limita-

tions and challenges in the three approaches and shows that participants used the hypothesis-driven approach in a materially different way than the recommendation-driven or AI-explanation-only conditions, with participants focusing more on the evidence than on their own background knowledge.

## 4.2 Methodology: Weight of Evidence (WoE)

We define the *evidence-informed hypothesis-driven decision-making* model by implementing the *evaluative AI* (hypothesis-driven) paradigm [153] using the WoE model. Specifically, given a classification problem, decision-makers explore evidence for and against each hypothesis (i.e. an output class). We allow decision-makers to interact with the model by repeatedly selecting a hypothesis for which they can then see the positive (or negative) evidence.

In a classification problem, a hypothesis $h \in Y$, where $Y$ is a set of possible hypotheses. Then, $Y_{-h} = Y \setminus \{h\}$ refers to all hypotheses other than $h$. For example, if a doctor asserts a set of hypotheses $Y = \{h_1, h_2, h_3\}$ where $h_1$ = *the patient has Covid*, $h_2$ = *the patient has Influenza* and $h_3$ = *the patient has pneumonia*, then $Y_{-h_1}$ = *the patient does not have Covid* which includes all possible hypotheses except having Covid, that is $Y_{-h_1} = \{h_2, h_3\}$.

We generate the *weight of evidence* for possible hypotheses by applying the Weight of Evidence (WoE) framework, which is a probabilistic approach for analysing variable importance, introduced in the context of explainability by Melis et al. [148] building on the approach of Good [70]. It provides a quantitative response to the question of why a model predicted output $h$ for a particular input $X$ in terms of how much each input feature $x_i$ provides in favour of, or against, $h$, relative to alternatives. Through Bayes rule, WoE can be understood as an adjustment to the prior log odds caused by observing the evidence.

For hypothesis $h$ and input feature $x_i$, weight of evidence, woe, is defined as follows:

$$\text{woe}(h \mid x_i) = \log \frac{P(x_i \mid h)}{P(x_i \mid Y_{-h})} = \log P(x_i \mid h) - \log P(x_i \mid Y_{-h}) \tag{4.1}$$

where $\log P(x_i \mid h)$ is the Gaussian log density for hypothesis $h$.

Based on the weight of evidence, we say the evidence supports or refutes a hypothesis:

- If $\text{woe}(h \mid x_i) > 0$, evidence $x_i$ supports hypothesis $h$
- If $\text{woe}(h \mid x_i) < 0$, evidence $x_i$ refutes hypothesis $h$
- If $\text{woe}(h \mid x_i) = 0$, evidence $x_i$ neither supports or refutes hypothesis $h$

Considering the vector input $X = [x_1, x_2, \ldots, x_n]$, the features can be independent or dependent on each other. We can apply Gaussian log density for both independent and dependent variables, depending on how we handle the covariance matrix $\Sigma$. The covariance matrix measures the relationship between two variables, indicating the direction (positive or negative) of how much each pair of features changes together and the strength of their relationship. For independent variables, the covariance matrix is diagonal, while for dependent variables, the covariance matrix is full.

**Independent Variables** For independent variables, the Gaussian log density is based on a univariate Gaussian distribution, which can be simplified as follows:

$$P(x_i|h) = \mathcal{N}(x_i; \mu_{i|h}, \sigma^2_{i|h}) = \frac{1}{\sqrt{2\pi\sigma^2_{i|h}}} \exp\left(-\frac{(x_i - \mu_{i|h})^2}{2\sigma^2_{i|h}}\right)$$
$$\log P(x_i|h) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2_{i|h}) - \frac{(x_i - \mu_{i|h})^2}{2\sigma^2_{i|h}} \tag{4.2}$$

where $\mu_{i,h}$ is the mean of $x_i$ for hypothesis $h$, and $\sigma^2_{i,h}$ is the variance of $x_i$ for hypothesis $h$.

**Dependent Variables** For dependent variables, the Gaussian log density is based on a multivariate Gaussian distribution. In this case, we compute the conditional distribution of $x_i$ given the remaining features $X_{-i}$ and the hypothesis $h$. We denote $-i$ as the set of features excluding the feature $i$. We calculate conditional mean, conditional variance and Gaussian log density as follows.

$$
\begin{aligned}
\mu_{i|-i,h} &= \mu_{i|h} + \Sigma_{i,-i|h}\Sigma^{-1}_{-i,-i|h}(X_{-i} - \mu_{-i|h}) \\
\sigma^2_{i|-i,h} &= \Sigma_{i,i|h} - \Sigma_{i,-i|h}\Sigma^{-1}_{-i,-i|h}\Sigma_{-i,i|h} \\
P(x_i|X_{-i},h) &= \mathcal{N}(x_i;\mu_{i|-i,h},\sigma^2_{i|-i,h}) = \frac{1}{\sqrt{2\pi\sigma^2_{i|-i,h}}}\exp\left(-\frac{(x_i-\mu_{i|-i,h})^2}{2\sigma^2_{i|-i,h}}\right) \\
\log P(x_i|X_{-i},h) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2_{i|-i,h}) - \frac{(x_i-\mu_{i|-i,h})^2}{2\sigma^2_{i|-i,h}}
\end{aligned} \tag{4.3}
$$

**Subset of features** When we consider a subset of features $S$ for hypothesis $h$, we can calculate the conditional distribution of $X_S$ given the remaining features $X_{-S}$ and the hypothesis $h$. We denote $-S$ as the set of features excluding $S$. We calculate the conditional mean, conditional variance matrix and Gaussian log density as follows.

$$
\begin{aligned}
\mu_{S|-S,h} &= \mu_{S|h} + \Sigma_{S,-S|h}\Sigma^{-1}_{-S,-S|h}(X_{-S|h} - \mu_{-S|h}) \\
\Sigma_{S|-S,h} &= \Sigma_{S,S|h} - \Sigma_{S,-S|h}\Sigma^{-1}_{-S,-S|h}\Sigma_{-S,S|h} \\
\log P(x_S|x_{-S},y) &= -\frac{|S|}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{S|x_{-S},y}| - \frac{1}{2}(x_S - \mu_{S|x_{-S},y})^\top\Sigma^{-1}_{S|x_{-S},y}(x_S - \mu_{S|x_{-S},y})
\end{aligned} \tag{4.4}
$$

**Mixture model** To compute $P(x_i|Y_{-h})$, we sum over all possible hypotheses $h' \in Y_{-h}$:

$$
P(x_i|Y_{-h}) = \sum_{k\in Y\setminus\{h\}} P(x_i|k)P(k) \tag{4.5}
$$

#### How decision-aid models can use WoE to make a decision

Using the weight of evidence for each feature $x_i$ as in Equation 4.1, a decision-aid model can make a prediction based on the total weight of evidence of a hypothesis $h$ by summing up the weight of evidence of this hypothesis based on each feature $x_i$. The total

weight of evidence is defined as follows.

$$\text{woe}(h) = \sum_{i=1}^{n} \text{woe}(h \mid x_i) \tag{4.6}$$

where $n$ is the number of features.

The decision-aid model will select the best hypothesis based on the maximum posterior, that is, $y = \arg\max_{h \in Y} P(h \mid X)$. If we have the same prior for all hypotheses, we can also use the total weight of evidence as another way to find the best hypothesis using Equation 4.1. Therefore, a decision-aid model can select the hypothesis with the maximum total weight of evidence as its prediction as follows (only apply to uniform priors).

$$y = \arg\max_{h \in Y} \text{woe}(h) \tag{4.7}$$

**How WoE can incorporate a human approach to making a decision**

To assist users with interpretability, Melis et al. [148] complement the display of the magnitude of the weight of the evidence with a notion of the significance level of the evidence, using a scale of seven categories: *decisive-against (**---**)*, *strong-against (**--**)*, *substantial-against (**-**)*, *not-significant (**N**)*, *substantial-in-favour (**+**)*, *strong-in-favour (**++**)*, *decisive-in-favour (**+++**)*. The details can be found in the rule-of-thumb guidelines here [4].

In addition to the weight of evidence of a feature, we suggest it is useful to distinguish the *importance* of a feature – with importance being domain-specific and determined by the domain expert using the model. Specifically, if a feature has a significant weight of evidence according to the WoE model, but that feature is not seen as important by the human decision-maker, then it is reasonable to anticipate the impact of that evidence on the decision would be reduced by the decision maker. For example, if a clinician looks at a dermatoscopic image and also is aware of some irrelevant but high weight of evidence features such as dense hair, they should ignore that evidence in making a prediction.

Formally, by considering the importance of the evidence, we re-define the total weight

of evidence from a human decision-making perspective as follows:

$$\text{woe}(h) = \sum_{i=1}^{n} \gamma_i \times \text{woe}(h \mid x_i) \tag{4.8}$$

where $\gamma_i$ is a parameter of feature $x_i$ that adjusts the weight of evidence based on importance, i.e., $\gamma_i > \gamma_j$ represents that feature $x_i$ is more important than feature $x_j$. $\gamma$ can be considered as the prior belief of the human decision-maker.

Then, for the skin cancer example just mentioned, in effect in Equation 4.8, the clinician has set $\gamma = 0$ for that feature.

## 4.3 Human Experiment Design

In this section, we describe the task implemented in the human behaviour experiment and the experiment design.

In selecting a decision-making task, we identified requirements similar to those used in other studies of how explanations can assist human decision-makers interacting with AI decision support, e.g. Vasconcelos et al. [213]: the task should not be too easy for humans to complete without a decision aid, but also, as we were using lay subjects from Prolific for this particular study, the task cannot require specialist knowledge.

We chose a version of the *housing price prediction* task studied previously in an XAI context [1, 44]. In this task, participants are provided with information about house features and other information which varies by experimental condition and are asked to choose whether the given house would have a sale price of *low*, *medium* or *high*. As noted by others [44], real estate evaluation is a domain where ML models have been developed to help people make better decisions, predicting house prices is a task that lay people may need to do in real life, so it is not unrealistic to expect they have sufficient day-to-day knowledge to make predictions and decide whether or not to rely on an AI model.

Experimenting with this task, we compared the *hypothesis-driven* approach with two state-of-the-art decision-making approaches using quantitative measures for *efficiency*, *performance* and *reliance* and qualitative analysis of *information use*. In the terminology of a recent review of XAI evaluation [124], the first two points of comparison are a form of

evaluation with respect to the decision task, and the latter two focus on users' perception and use of the AI system itself. Since we aim to measure whether the hypothesis-driven approach improves human decision-making, including reducing over- or under-reliance, this cannot be done without conducting human behaviour experiments.

### 4.3.1 Dataset and Model Implementation

To build our model, we used the Ames Housing Dataset [48] and the open source code on GitHub [201] for data pre-processing. The data after pre-processing has a total of 2616 instances and 28 features. We processed the dataset further by converting the house price into three output classes (*low price*, *medium price* and *high price*). We also balanced the dataset to ensure that the three classes had the same number of instances by using Near-Miss Undersampling. Finally, we had a total of 1920 instances with 640 instances for each class.

We selected six features for the human experiment in the house-price decision-making task by applying a Gradient Boosting Classification model over the data. Considering domain-specific decision-making about house prices, we propose there to be three important features (*quality of construction*, *house age* and *location*) and three unimportant features (*fireplaces*, *kitchen quality* and *central air conditioning*). We divided the dataset into 80% for the training set and 20% for the test set. Following Melis et al. [148], we use a Gaussian Naïve Bayes (GNB) classifier to obtain $P(x_i \mid h)$. This assumes that features are independent, but the model and implementation work for any probabilistic classifier. We chose this model because it is a simple discriminative classifier that aligns with previous work on evidence-based explanations [120, 177].

### 4.3.2 Experimental Conditions

All participants[1] were given the six house feature values plus other information, which varied by condition as set out below. Participants then chose whether the given house would have a price of *low*, *medium* or *high*. Using a *between-subject design*, participants

[1]We received ethics approval from our institution before conducting the human experiment (ID: 23208).

were randomly assigned to one of three conditions:

- (C1) *Recommendation-driven*: Participants see the AI prediction (i.e., either *low* or *medium* or *high*) *and* also the weight of evidence for that prediction;
- (C2) *AI-explanation-only*: Participants see the weight of evidence associated with the AI prediction, but the AI prediction itself is hidden;
- (C3) *Hypothesis-driven*: Participants see the weight of evidence for *all* hypotheses (*low*, *medium* and *high*), but the AI prediction itself is hidden.

Although participants in the *AI-explanation-only* and *hypothesis-driven* conditions did not see a recommendation, it was expected that the displayed information from the WoE framework would provide insight that participants could use to support their decision-making. We note a similar AI-explanation-only approach has been explored previously [64].

### 4.3.3 Research Questions and Hypotheses

Our overarching research questions were as follows:

- **RQ1**: *(Efficiency)* What form of AI assistance helps participants make faster decisions?
- **RQ2**: *(Performance)* What form of AI assistance helps participants make better decisions?
- **RQ3**: *(Reliance)* What form of AI assistance helps reduce over-reliance and under-reliance?
- **RQ4**: *(Information use)* How do people make decisions differently in *recommendation-driven*, *AI-explanation-only* and *hypothesis-driven* approach?

For **RQ1**, we evaluated the participants' speed in making a decision. We use the most common metric - *completion time* to measure the time taken on the task. The corresponding hypotheses for this question are:

- **H1a/b**: *(C3) Hypothesis-driven* approach will cost less time to finish the task than *(C1) Recommendation-driven* and *(C2) AI-explanation-only*.

For **RQ2**, we evaluated the quality of the decision. In the task, we asked the participants to assign the likelihood for each price range (low/medium/high) where 100 is the most likely and 0 is the least likely. The sum of three likelihoods must be equal to 100. We expect the participants to be confident when they make a correct prediction, and *not be confident when they make a wrong decision*. We apply *Brier score* as explained below to measure the task performance. The hypotheses for this question are:

- **H2a/b**: *(C3) Hypothesis-driven* approach will help participants make better decisions than *(C1) Recommendation-driven* and *(C2) AI-explanation-only*.

For **RQ3**, we investigated the participants' capability of appropriately calibrating their decision. Participants should follow the model's prediction when it is correct and should not use the model's prediction when it is wrong. We applied two measures *over-reliance* and *under-reliance* as shown below with the following hypotheses:

- **H3a**: *(C3) Hypothesis-driven* can reduce over-reliance compared to *(C1) Recommendation-driven*.

- **H3b**: *(C3) Hypothesis-driven* can reduce under-reliance compared to *(C2) AI-explanation-only*.

For **RQ4**, we looked into the text written by participants when they explained why they selected an option after each question to know how they used the provided information in each decision-making approach to make their decisions. Therefore, we can identify the limitations of each approach and the generated evidence that led the participants to make a wrong decision.

### 4.3.4 Measures

We took the following measures:

1. *Task Efficiency* (Completion time): The time participants take to complete each task.

2. *Task Performance* (Brier score): This metric quantifies the effectiveness of task performance in terms of accurate decision outcomes. The formula is:

$$\mathrm{BS}_p = \frac{1}{N}\sum_{i=1}^{N}(C_{p,i} - A_{p,i})^2 \tag{4.9}$$

where: $C_{p,i}$ is the likelihood level of a participant $p$ in question $i$, ranging from 0 to 1 (the likelihood level of a participant refers to how confident they are when answering the question); $A_{p,i}$ is the answer score of participant $p$ in question $i$, either 0 (wrong answer) or 1 (right answer); N is the number of questions for each participant. The best Brier score (i.e. equal to 0) for an individual task is when a participant answers the task correctly and gives it a 100% likelihood (or alternatively, a wrong answer but with 0% likelihood). Therefore, a participant has better task performance when they have a lower Brier score. The Brier score measures decision accuracy but awards a higher score for a correct answer when a participant is confident, and a lower penalty for an incorrect answer when not confident. This mitigates problems where participants guess answers (i.e. have low confidence in their answers). Brier score has potential for real-world applications, as it accounts for both the correctness of a decision and the confidence of the decision-maker.

We then measure the *over-reliance* and *under-reliance* [225]. Since study participants can only see the AI recommendation in C1 (*Recommendation-driven*), we measure *agreement*: whether participants have the same prediction or differ from the model's prediction in the other two conditions.

3. *Over-reliance*: the fraction of tasks where participants have the same decision as a model's prediction when it was wrong: $\Sigma_i(A_{p,i} = M_i = 0)/\Sigma_i(1 - M_i)$, where $A_{p,i}$ is as above and $M_i = 1$ if the model is correct and 0 otherwise.

4. *Under-reliance*: the fraction of tasks where participants have a different decision from a model's prediction when it was correct: $\Sigma_i(A_{p,i} \neq M_i = 1)/\Sigma_i M_i$.

### 4.3.5 Conduct

We conducted ***two*** separate human experiments in which participants were given the same task in the form of a question set, with the only difference being the way they answered the question.

In experiment 1, participants were asked to make a decision about the relative likelihood for each price range (*low*/*medium*/*high*) of given house instances. We answer **RQ1**, **RQ2** and **RQ3** by analysing the results of four measures mentioned above (completion time, Brier score, over-reliance and under-reliance).

In experiment 2, we recruited a new and smaller cohort and asked them to do the same tasks as in experiment 1, but in addition, we asked participants to explain their decisions using free text. We conduct this experiment separately from the quantitative data in experiment 1 because asking participants to explain their reasoning cognitively forces them to engage with the instance, interfering with their natural decision-making process, and therefore potentially affecting the quantitative results. We then performed a deductive analysis of their explanations to answer **RQ4**.

Each experiment was designed as a Qualtrics[2] survey and participants accessed the survey through Prolific[3]. The experiment required a maximum of 25 minutes to finish. There were 12 house instances given, equivalent to 12 questions. These 12 questions were evenly distributed into four question categories: (1) where the model gives *correct* predictions with *high uncertainty*, (2) where the model gives *correct* predictions with *low uncertainty*, (3) where the model gives *wrong* predictions with *high uncertainty* and (4) where the model gives *wrong* predictions with *low uncertainty*. Thus, there are three questions in each category. Study participants were *not* informed about in which category the question belongs. The questions are ordered randomly in the experiments.

The uncertainty is measured by the cross entropy as follows. We use the cross entropy because it is the most common measure of uncertainty in probabilistic models.

$$u(h) = -\sum_{h \in H} p(h) \log p(h) \tag{4.10}$$

[2]https://www.qualtrics.com
[3]https://www.prolific.com

where $u(h)$ is the uncertainty level of hypothesis $h$ given the probabilistic output is $p(h)$. To ensure there was a clear difference between high and low uncertainty, we selected instances with *low uncertainty* by choosing instances with entropy less than 0.3, and *high uncertainty* by choosing instances with entropy greater than 0.7. Participants did not know the certainty nor how many test instances were correct/incorrect. Each participant was paid a minimum of £4 for their time, plus a bonus of £2 if they could answer at least 9 out of 12 questions correctly. Participants were also given a plain language statement, and consent form and did a training phase with 3 example questions before answering the 12 test questions.

For **RQ4** the text is a response to "Can you please explain why you selected this option?". We analysed a total of 12 (questions) $\times$ 95 (participants) = 1140 responses. The final analysis includes 1,031 responses after removing 109 responses due to poor quality. Each response is assigned to at least one category (or code): *Using feature values* or *Using evidence*. We then perform a simple deductive analysis by reading each response and assigning the relevant codes. We explain each code as follows: (1) *Using Feature Values*: participants rely on the feature values and their background knowledge to make the final decision without using the model evidence; and (2) *Using Evidence*: Participants rely on the evidence provided by the model and possibly their background knowledge to make the final decision.

We chose these two codes based on the idea of *machine explanation* and *human intuition* [39, 40]. Specifically, *using evidence* refers to using the machine explanation and therefore, making use of the model evidence to support decision-making. On the other hand, *using feature values* is relevant to using people's intuitions of the task based on the input feature values. Therefore, using the qualitative analysis, we explore how people use the model evidence and their intuitions in the three decision-making approaches.

### 4.3.6 Participants

**Experiment 1** Using the power analysis for the F-test for one factor ANOVA and assuming the power of 0.8 and significant alpha of 0.05, we found that a sample size of 300 participants in three groups guarantees a small effect size of 0.2. In total, we recruited

$N = 302$ participants on Prolific, distributed into three conditions: 102 participants in C1, 99 participants in C2 and 101 participants in C3. Participants are selected from the United States, United Kingdom, and Australia and must be fluent in English. Gender-wise, 192 were women, 103 were men, 4 self-specified their gender and 3 declined to state their gender. Age-wise, 94 participants were between Ages 18 and 29, 91 were between Ages 30 and 39, 44 were between Ages 40 and 49, and 73 were over Age 50. All collected data were included in the analysis, as there was no evidence indicating that any participants provided poor-quality data.

**Experiment 2** We recruited $N = 95$ participants on Prolific, distributed into three conditions: 30 participants in C1, 34 participants in C2 and 31 participants in C3. Participants are selected from the United States, United Kingdom, and Australia and must be fluent in English. Gender-wise, 52 were women, 41 were men, and 2 declined to state the gender. Age-wise, 38 participants were between Ages 18 and 29, 37 were between Ages 30 and 39, 10 were between Ages 40 and 49, and 10 were over Age 50. Participants in the first study were *not* allowed to participate in the second.

## 4.4 Experiment Results

In this section, we show the results of two experiments. In the first experiment, we explore whether *hypothesis-driven* can improve task efficiency and task performance, and reduce reliance compared to *recommendation-driven* and *AI explanation only*. In the second experiment, we understand how participants used our hypothesis-driven approach differently compared to the other two baselines.

### 4.4.1 Experiment 1: Quantitative Results

We performed a Shapiro-Wilks test to check the data normality and we found that our data was not normally distributed ($p < 0.05$). Therefore, we apply the non-parametric Kruskal-Wallis test to analyse non-normally distributed data. We then perform post-hoc Mann-Witney U tests to do pairwise comparisons. The results are visualised in Figure 4.1

and Figure 4.2. The significant differences between the two conditions are highlighted in italic red in the figures where $p < 0.05$.

### Task efficiency

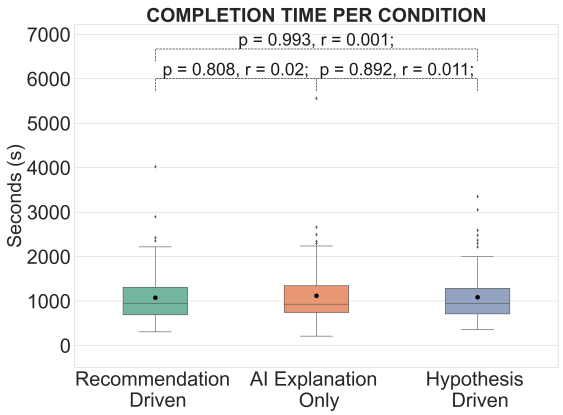



Figure 4.1: Completion time. Lower is better. Means represented as dots.

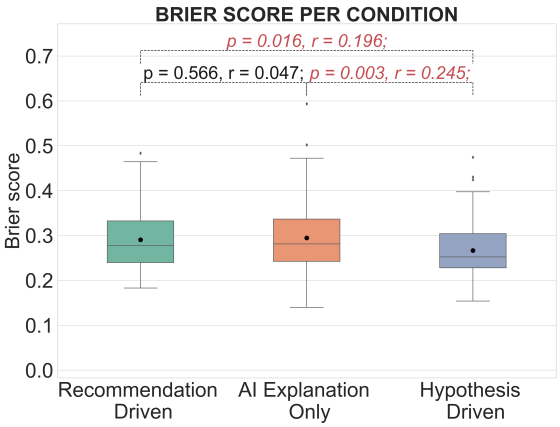



Figure 4.2: Brier score. Lower is better. Means represented as dots.

Figure 4.1 shows the completion time in three conditions. There is no statistically significant difference among these three conditions ($p \approx 0.9$). We reject **H1a/b**. ***This shows that hypothesis-driven does not take more time to complete the task than recommendation-driven and AI-explanation-only***.

### Task performance

We evaluate participants' decision-making performance by using the Brier score. A lower Brier score indicates better decision accuracy. As seen in Figure 4.2, ***participants in the hypothesis-driven condition*** ($M = 0.267$, $SD = 0.063$) ***have a lower Brier score than the other two approaches*** (C1: ($M = 0.290$, $SD = 0.071$), C2: ($M = 0.295$, $SD = 0.073$)). We **accept H2a/b** with small effect sizes. Therefore, hypothesis-driven helps participants be confident when they make a correct decision, and be less confident when they make a wrong decision.

### Over-reliance

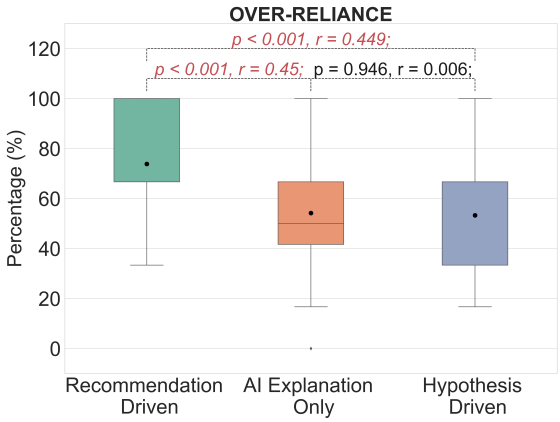



Figure 4.3: Over-reliance. Lower is better. Means represented as dots.

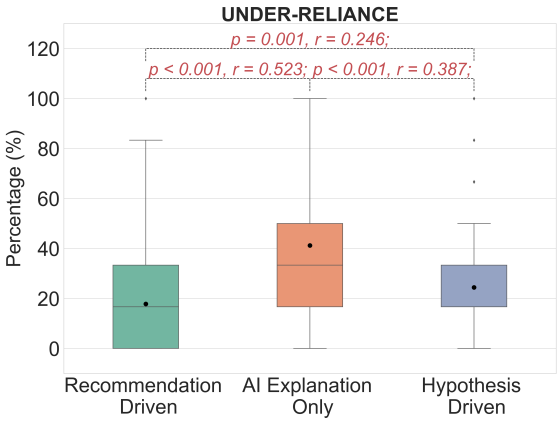
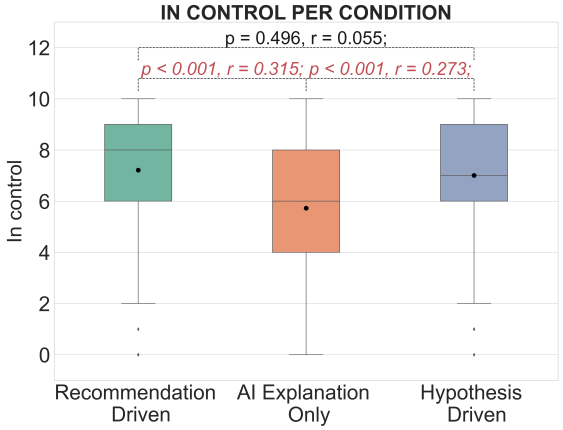



Figure 4.4: Under-reliance. Lower is better. Means represented as dots.

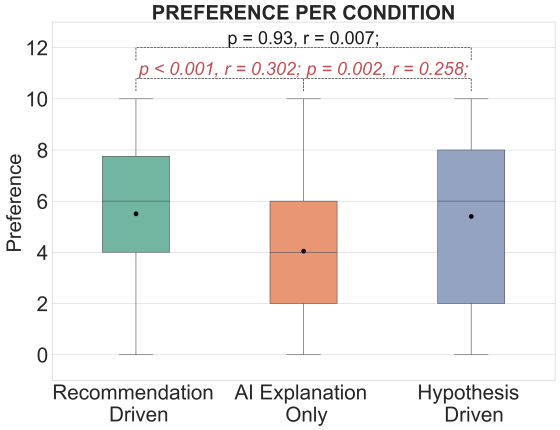
In Figure 4.3, ***hypothesis-driven*** ($M = 53.30$, $SD = 22.73$) ***reduced over-reliance significantly compared to recommendation-driven*** ($M = 73.86$, $SD = 20.91$) ($p = 1.5 \times 10^{-8}$, $r = 0.449$). We **accept H3a** with a medium effect size of 0.449. Moreover, *AI-explanation-only* ($M = 54.21$, $SD = 22.51$) also reduces over-reliance compared to the recommendation-driven approach ($p = 1.6 \times 10^{-8}$, $r = 0.450$).

### Under-reliance

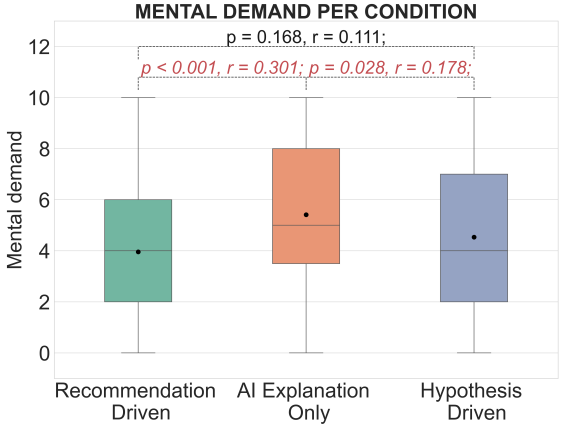
In Figure 4.4, ***hypothesis-driven*** ($M = 24.42$, $SD = 18.19$) ***significantly reduced under-reliance compared to AI-explanation-only*** ($M = 41.25$, $SD = 27.18$) ($p = 1.09 \times 10^{-6}$, $r = 0.387$). We **accept H3b** with a medium effect size of 0.387. This is not surprising because we expect that participants in the *AI-explanation-only* condition are the most likely to 'under-rely' due to being unable to compare evidence across hypotheses nor see a recommendation. Recommendation-driven ($M = 17.81$, $SD = 20.35$) has the least under-reliance value because participants were given recommendations.



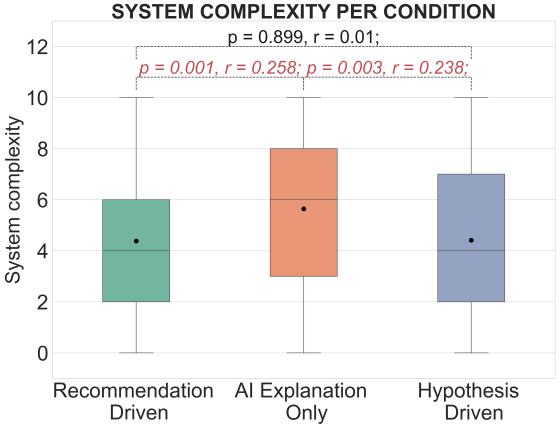

(a) In control (Higher is better)



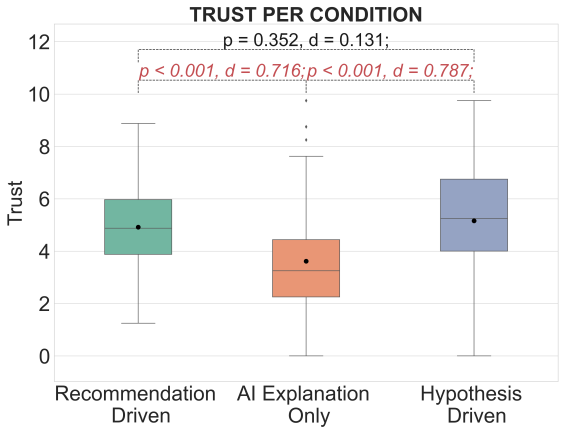

(b) Preference (Higher is better)




(c) Mental demand (Lower is better)




(d) System complexity (Lower is better)




(e) Trust (Higher is better)

Figure 4.5: Subjective Measures in Experiment 1. Means represented as dots.

### 4.4.2 Experiment 1: Subjective Questions

After doing the experiment 1, participants were asked to answer 12 subjective questions (SQs) as follows.

SQ1. **In control**: I feel in control of the decision-making process when using this decision aid. (0 = Disagree strongly; 10 = Agree strongly)

SQ2. **Preference**: I would like to use this decision aid frequently. (0 = Disagree strongly; 10 = Agree strongly)

SQ3. **Mental demand**: I found this task difficult. (0 = Disagree strongly; 10 = Agree strongly)

SQ4. **System complexity**: The decision aid was complex. (0 = Disagree strongly; 10 = Agree strongly)

SQ5. **Trust**: I am confident in the decision aid. I feel that it works well. (0 = Disagree strongly; 10 = Agree strongly)

SQ6. **Trust**: The decision aid is very predictable. (0 = Disagree strongly; 10 = Agree strongly)

SQ7. **Trust**: The decision aid is very reliable. I can count on it to be correct all the time. (0 = Disagree strongly; 10 = Agree strongly)

SQ8. **Trust**: I feel safe that when I rely on the decision aid I will get the right answers. (0 = Disagree strongly; 10 = Agree strongly)

SQ9. **Trust**: The decision aid is efficient in that it works very quickly. (0 = Disagree strongly; 10 = Agree strongly)

SQ10. **Trust**: I am wary of the decision aid. (0 = Disagree strongly; 10 = Agree strongly)

SQ11. **Trust**: The decision aid can perform the task better than a novice human user. (0 = Disagree strongly; 10 = Agree strongly)

SQ12. **Trust**: I like using the decision aid for decision-making. (0 = Disagree strongly; 10 = Agree strongly)

We evaluate SQ1-4 separately to measure 4 measures (**In control**, **Preference**, **Mental demand** and **System complexity**). We aggregate SQ5-12 to measure **Trust**. We based the trust questions on questions from [81].

In Figure 4.5, *AI-explanation-only* is significantly worse than the other conditions in all facets. Moreover, *recommendation-driven* and *hypothesis-driven* are quite similar and there is no statistical difference between these two conditions. The reason is that we conduct between-subject experiments so each participant has access to only one condition and they do not have another condition to compare to. If we run within-subject experiments to measure subjective questions in the future, participants can compare different decision-making approaches and evaluate which one they prefer the most. Factor analysis could also be used to consolidate the questions.

Notably, the *hypothesis-driven* condition did not result in the highest mental demand. This is because the *AI-explanation-only* condition was more difficult to interpret. In the *AI-explanation-only* condition, participants did not know which hypothesis the evidence was referring to, making it challenging to make sense of the information. Moreover, when participants relied on their prior knowledge, they could check their hypothesis and assess the evidence to confirm it, which is easier than trying to understand a machine recommendation when it differs.

### 4.4.3 Experiment 2: Qualitative Results

In Figure 4.6 and 4.7, we illustrate the number of times that participants used feature values and evidence to make their decisions based on the text analysis. The questions in this section (Q0-Q11) refer to the 12 house instances used in the experiment described in Section 4.3.5. Participant comments are attributed to the question number, not the participant ID.

***For the recommendation-driven approach, participants use the feature values to confirm whether the decision aid's prediction and explanation are reliable or not***. If participants think the feature values do not match the evidence explanation, they will go with

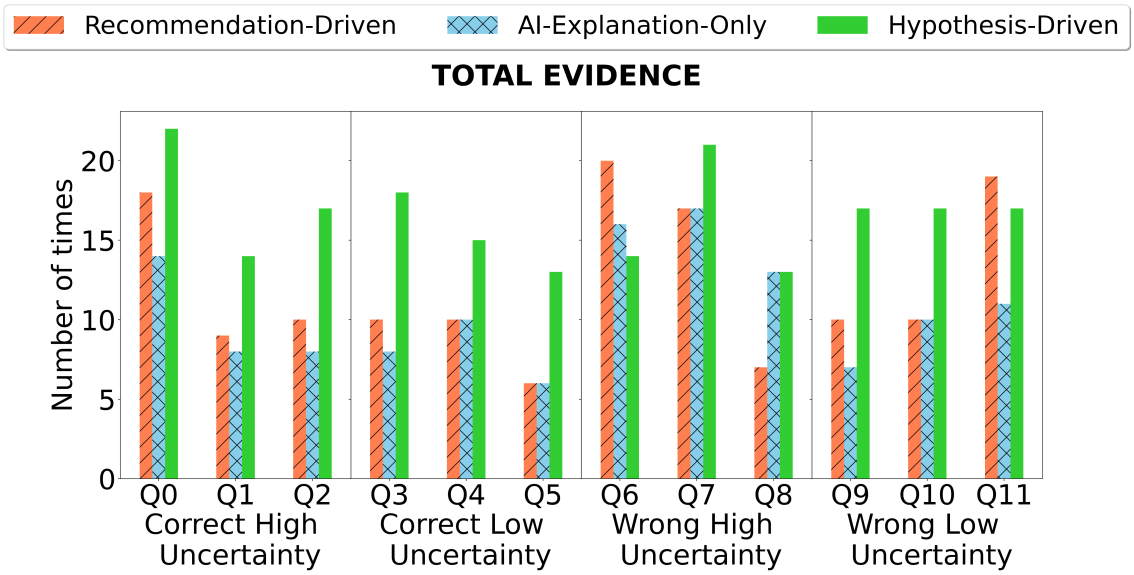



Figure 4.6: Frequency of using evidence to make a decision.

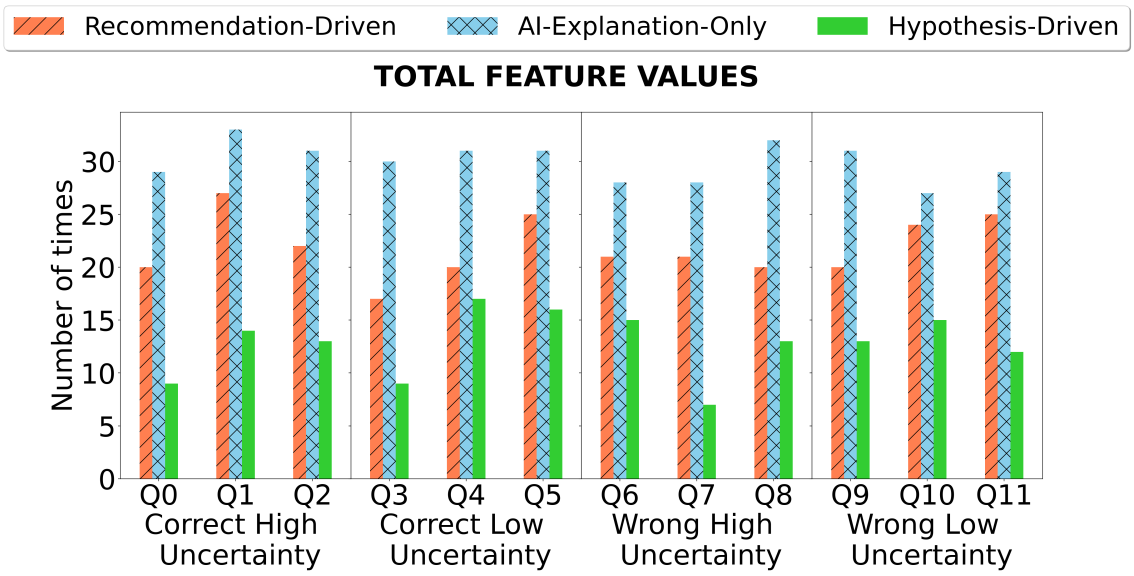



Figure 4.7: Frequency of using feature values to make a decision.

the feature values to make the final decision. Some examples that the study participants in the recommendation-driven condition go against the decision aid's prediction:

> *"Here, I believe the decision aid is mistaken. My rating would be medium because the house is very old which is overlooked by the model. Other features are all decent or above decent but the house age is an important feature."* – Q11

> *"The location of the property is low so I thought that would bring down the price"* – Q7

Ignoring evidence is, of course, a good strategy if the decision-maker believes that the evidence is wrong. However, ***recommendation-driven does not help participants to be aware of the high uncertainty among multiple predictions***. This limitation is mitigated by the hypothesis-driven approach.

***For the AI-explanation-only approach, participants often rely on the feature values and not on the evidence explanation to make a decision***. This is not surprising because participants can find it difficult to interpret the evidence without seeing the label that the evidence is referred to. We attribute this to the cognitive effort to link evidence to hypotheses, leading to participants ignoring evidence and relying on input feature values to make their decisions. This is a noteworthy limitation of *AI-explanation-only* as it makes people overlook the explanation if the link to the evidence is unclear. In the study by Gajos and Mamykina [64], the link from feature attributions to the task solution is more straightforward than in our study, which may explain the divergence of results.

***Participants more often use the evidence to make a decision in the hypothesis-driven approach*** than in recommendation-driven or AI explanation only. This shows participants took advantage of the model evidence. In the two baseline conditions, participants tended to ignore evidence seemingly due to the inability to interpret it, which means they will fail to take advantage of the underlying model. In Figure 4.6, there are only two exceptions at Q6 and Q11 where the evidence is not the most used in the hypothesis-driven condition.

We found that in hypothesis-driven, participants reported that it was difficult to make decisions for two main reasons:

- ***Uncertainty awareness***: This is where there are multiple hypotheses with similar strength evidence. Participants are aware of the uncertainty in the model solely based on the positive and negative evidence provided for all hypotheses. In this case, participants use the input feature values or choose the hypothesis that they think is slightly better than the others when making the final decision. Figure 4.8 shows an example where two hypotheses *low* and *medium* both have a positive and

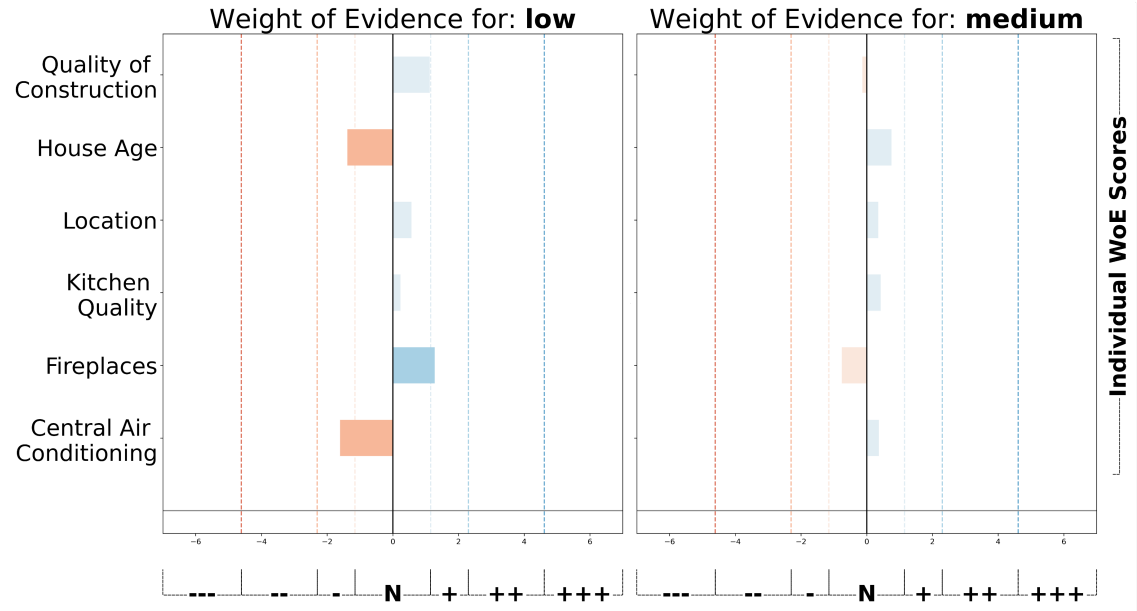



Figure 4.8: An example of ***uncertainty awareness*** (Q6).

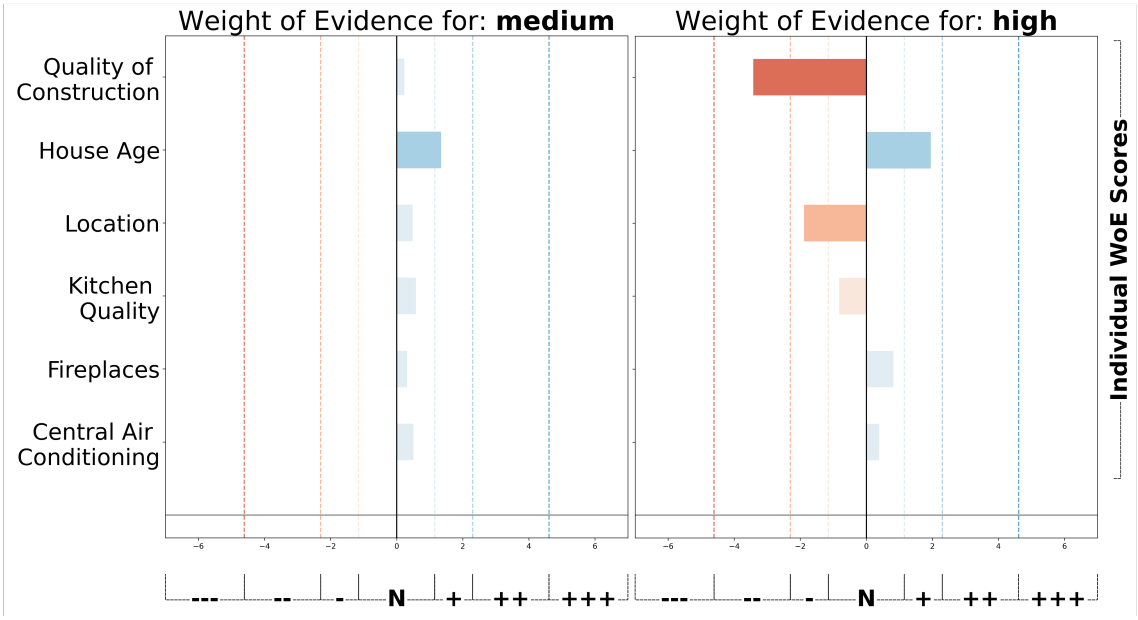



Figure 4.9: An example of ***deceptive evidence*** (Q9).

negative weight of evidence, especially in the top three important features. For instance, some participants explicitly explain their uncertainty in the text as follows.

*"I was choosing between high and medium. Quality of construction, age and lo-*

> *cation are the most important features. When it was high, these were all positive. Kitchen quality and fireplaces were negative, but these are not as important.*" – Q0
>
> "*The amount of negative or positive evidence for low or medium is the same, including the three more important factors. Both medium or low could be viable but medium has less variance and is overall more balanced.*" – Q6
>
> "*The house is clearly not in a high bracket, but it is somewhat difficult to decide between low and medium. There are stronger indicators in low, going both ways, while medium has largely insignificant indicators. Low has a significant negative weighting for house age and this pushed me towards medium.*" – Q6

- ***Deceptive evidence***: When the evidence was strongest for an incorrect option. In this case, many participants just follow the evidence and make the wrong decision. Figure 4.9 illustrates an example of Q9 where we have all positive evidence in hypothesis *medium*, but strong negative evidence in hypothesis *high*. Therefore, all participants choose hypothesis *medium*, but hypothesis *high* is the ground truth. Future work will need to address the challenge of building trustworthy evidence.

In summary, the qualitative analysis showed that participants took advantage of the decision aid more in the hypothesis-driven condition than in recommendation-driven and explanation-only conditions. Further, we also found that participants recognised model uncertainty in the hypothesis-driven condition. However, there still remains a limitation of having deceptive evidence.

## 4.5 Discussions

### 4.5.1 Strengths and weaknesses of our hypothesis-driven approach

First, participants using the hypothesis-driven approach required a similar time to complete the task compared to the recommendation-driven approach. Participants in the

hypothesis-driven condition also made higher quality decisions than *recommendation-driven* and *AI explanation only* based on the Brier score. The results indicated that the *hypothesis-driven* gave study participants a more complete picture of the underlying decision aid than the other two approaches, helping them to make use of the AI models when they are right and be less confident when the models are wrong.

Moreover, hypothesis-driven reduced over-reliance significantly compared to the standard AI recommendation. Similarly, hypothesis-driven also reduced under-reliance compared to AI explanation only. Importantly, the positive result for under-reliance using the recommendation-driven is not cancelled out by the poor over-reliance result, compared to the hypothesis-driven. The primary aim indicates potential for the use of uncertainty/confidence [21, 127] and conformal prediction [200] to direct decision makers' attention towards a set of hypotheses that it is confident about.

Using the qualitative analysis, hypothesis-driven helped participants take advantage of the decision support tool's evidence, and also recognise the uncertainty underlying the model. Using the strength of evidence, participants are aware of the uncertainty between multiple hypotheses. Therefore, they made an attempt to gauge the model uncertainty by calibrating the weight of evidence depending on whether the feature is important or not. Also, they could make use of the input feature values and choose the hypothesis that they perceive most likely matches with those values.

On the other hand, *recommendation-driven* and *AI explanation only* do not support this. We found that in recommendation-driven, people could use feature values to confirm the validity of the decision aid's prediction. However, they are not aware of the uncertainty among different hypotheses. In AI explanation only, people often ignore using the evidence and solely focus on using the feature values to make a decision because interpreting the evidence with this approach can be a lot more mentally demanding.

### 4.5.2 Study limitations

There are also some limitations to the study. First, we ran the experiment on one dataset (Ames Housing), which limits generalisability. In addition, since we follow the labels from this particular dataset, there is no single ground-truth for the price of a house. The

price can vary depending on many factors. Therefore, the experimental participants' tasks are somewhat subjective. Further, this task has only three output classes, so only three hypotheses, and we anticipate the results would be more interesting when we consider more hypotheses. The current form of explanation can be further improved by incorporating contrastive explanations. At present, we present multiple WoEs for different hypotheses, but these cannot be considered as contrastive explanations.

Regarding the experimental design, there are multiple forms of cognitive forcing, and we tested just one. We could have added another experimental condition where we asked participants to first think about their decision before showing the AI recommendation. This condition may have close performance to the hypothesis-driven approach. Finally, the human experiment is currently conducted with laypeople while experts likely interact with the decision-aiding tool differently from laypeople [60]. It is also important to note that we currently use the output labels of the model as the hypotheses, while domain experts may have more and different hypotheses.

## 4.6 Conclusion

In this chapter, we show that the hypothesis-driven approach using Weight of Evidence (WoE) can significantly reduce reliance and improve decision-making quality compared to two other prevalent decision-making approaches (recommendation-driven and AI explanation only). Furthermore, hypothesis-driven helps participants to be aware of the uncertainty among multiple options. Nevertheless, there still remains a challenge of study participants relying on the wrong (or misleading) evidence. Therefore, future work can address this challenge by exploring different approaches for presenting trustworthy evidence. More generally, potential future work is to consider the uncertainty/confidence in the generated evidence.

# Chapter 5
# Visual Evaluative AI

*This chapter presents **Visual Evaluative AI**, a decision-aid model that provides positive and negative evidence from image data for a given hypothesis. Different from chapter 4, extracting features from images is more complex than from well-structured tabular data, as it requires sophisticated techniques like convolutional neural networks, whereas tabular data provides predefined and explicit features. This challenge is not addressed in the original Weight of Evidence (WoE) framework. Particularly, our Visual Evaluative AI model finds high-level human concepts in an image and generates the WoE for each hypothesis in the decision-making process. We computationally demonstrate the effectiveness of Visual Evaluative AI on different concept-based explanation approaches. This model is further applied and evaluated in the skin cancer domain by building a web-based application that allows users to upload a dermatoscopic image, select a hypothesis and analyse their decisions by evaluating the provided evidence. Finally, we conduct a user study to understand the differences between the recommendation-driven approach and the hypothesis-driven approach and how they can impact human decision-making in supporting skin cancer diagnosis.*

## 5.1 Introduction

A common decision support approach called *recommendation-driven* provides either or both the AI recommendation and the explanation for the given recommenda-

This chapter is based on the following pre-print article under review:
**[P1] [130]** Thao Le, Tim Miller, Ruihan Zhang, Liz Sonenberg, Ronal Singh. "Visual Evaluative AI: A Hypothesis-Driven Tool with Concept-Based Explanations and Weight of Evidence."

tion [16, 205, 225]. However, this approach is yet to be effective because it limits the control of human decision-makers, which can cause *algorithm aversion* [52] where people do not trust the AI; or worse, *over-reliance* on the AI system [215]. Miller [153] proposes a paradigm shift called **hypothesis-driven** using a conceptual framework **evaluative AI**. The new framework aims to provide evidence for or against a given hypothesis. Rather than telling the decision-makers what to do, hypothesis-driven allows them to have more control of the decision-making process by incorporating their hypotheses and promoting uncertainty awareness [126, 128].

The original Weight of Evidence (WoE) framework in Chapter 4 had been implemented only for tabular data. Therefore, we extend the WoE in this chapter by applying concept-based explanations [234, 238] to extract human-understandable concepts from images and put these concepts into the WoE model. Additionally, despite existing methods like Grad-CAM [194] can provide contrastive explanations of various output classes, they did not use concepts that humans can understand in their explanation process. For that reason, we build and evaluate a *Visual Evaluative AI* model by combining concept-based explanations for image data and the Weight of Evidence (WoE) framework. This model offers hypothesis-driven decision-making by generating evidence for possible hypotheses of an image. In particular, we extract human concepts from images (e.g., *reddish structures*, *irregular pigmentation* in dermatoscopic images) by using concept-based explanation models (e.g., *Invertible Concept-based Explanation (ICE) [238]* and *Post-hoc Concept Bottleneck Model (PCBM) [234]*). The WoE model is then applied to show how much each concept contributes to a given hypothesis (e.g., *melanoma*, *benign keratosis*), which can be used as evidence to support human decision-making.

Our contributions are as follows:

- The *Visual Evaluative AI* model that provides positive and negative evidence for a given hypothesis for image datasets by combining concept-based explanations (Invertible Concept-based Explanation (ICE) [238] and Post-hoc Concept Bottleneck Model (PCBM) [234]) and the Weight of Evidence (WoE) framework. The combined models are called *ICE+WoE* and *PCBM+WoE*. We also provide public access

to *Visual Evaluative AI* as a Python package so other researchers can use our tool [1].

- We computationally evaluate *ICE+WoE* and *PCBM+WoE* and show that *ICE+WoE* and *PCBM+WoE* achieve comparable performance to the original CNN models with significantly fewer concepts.

- We conduct a human behavioural experiment to investigate the differences between the two approaches (recommendation-driven and hypothesis-driven) in the field of skin cancer diagnosis. Study participants are individuals with a background in the skin cancer field (e.g., PhD students, postdoctoral researchers, medical doctors and melanographers). The results show that both approaches have pros and cons, and the hypothesis-driven approach is more preferred by experienced diagnosers. Based on this study, we also propose suggestions for the design of the decision-making approach.

## 5.2 Methodology

In this section, we introduce our evidence generation model by combining a concept-based explanation model (e.g., Invertible Concept-based Explanation (ICE) [238], Post-hoc Concept Bottleneck Model (PCBM) [234]) and the Weight of Evidence (WoE) model [148]. In our experiments, we use *Invertible Concept-based Explanation (ICE)* [238] as an example of unsupervised concept learning, and *Post-hoc Concept Bottleneck Model (PCBM)* [234] as an example of supervised concept learning. Combining them together, we propose two models to generate the evidence-based explanations called *ICE+WoE* and *PCBM+WoE*. We provide a more detailed description of the concept-based explanation models and the WoE model below.

### 5.2.1 Concept-based Explanations

We divide concept-based explanations into two categories: (1) supervised learning concepts (concepts are labelled on each image in the training dataset) and (2) unsupervised

[1] https://github.com/thaole25/EvaluativeAI

learning concepts (not having concept labels in the training dataset). Supervised concept learning requires labelled concepts in the training set, or the concepts can be transferred using another labelled dataset [234]. Unsupervised learning concept methods do not require the concepts to be labelled during the training process. This method is helpful when labelling concepts can be laborious, require expertise, or are not always available. Moreover, unsupervised learning can give users more agency as they can find a new concept that has not been labelled but is still used by a machine learning model. In our application, the evidence will be referred to a concept (or feature) found in the image. Each concept will have a positive/negative quantitative value that shows how much it contributes to the given hypothesis.

The main difference between ICE+WoE and PCBM+WoE is that the concepts generated by ICE+WoE do not have labels returned by the model. In other words, ICE+WoE identifies only *important concepts* for the classifier but does not assign a name for them. The unlabelled concepts can then be assigned labels by a domain expert. On the other hand, PCBM+WoE provides a concept name for each concept, which is learned from the concept bank. It is important to note that the labelled concepts are not always reliable and still require validation by a domain expert. Furthermore, the number of concepts is fixed based on the concept bank in PCBM+WoE, while the number of concepts can be chosen by the user in ICE+WoE.

### 5.2.2 ICE+WoE and PCBM+WoE

We will now explain the implementation of the combined models of concept-based explanations and the Weight of Evidence (WoE) model, depending on the concept learning method (unsupervised or supervised). Figure 5.1 shows an overview of the unsupervised concept learning model, and how it can be combined with Weight of Evidence (WoE) [148]. Figure 5.2 shows an overview of the supervised concept learning model using transferred concept bank, and how it can be combined with Weight of Evidence (WoE) [148].

Formally, let $f : \mathcal{X} \rightarrow \mathbb{R}^d$ be the pre-trained backbone model (e.g. ResNet, ResNeXt) where $\mathcal{X}$ is the input space and $d$ is the size of the embedding space. The dimension $d$ of

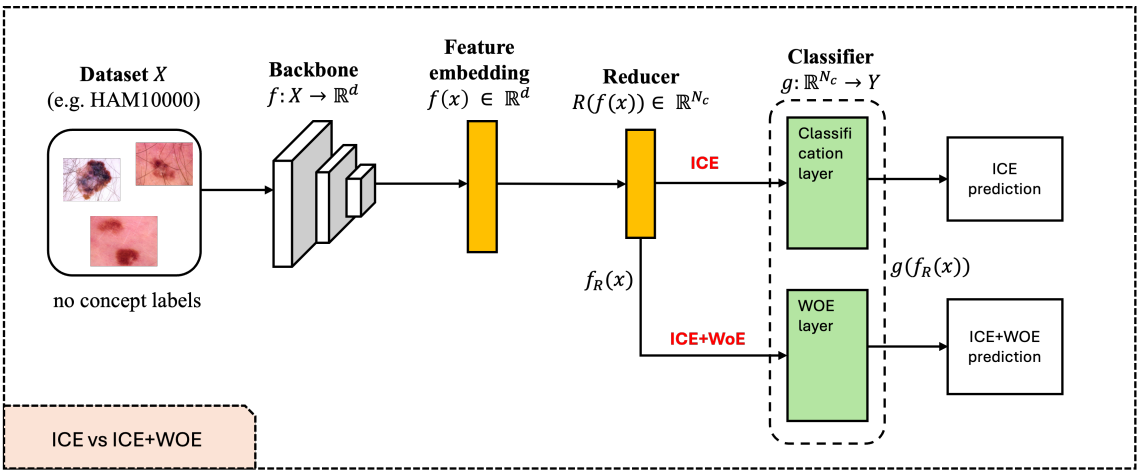



Figure 5.1: Unsupervised Concept Learning Model

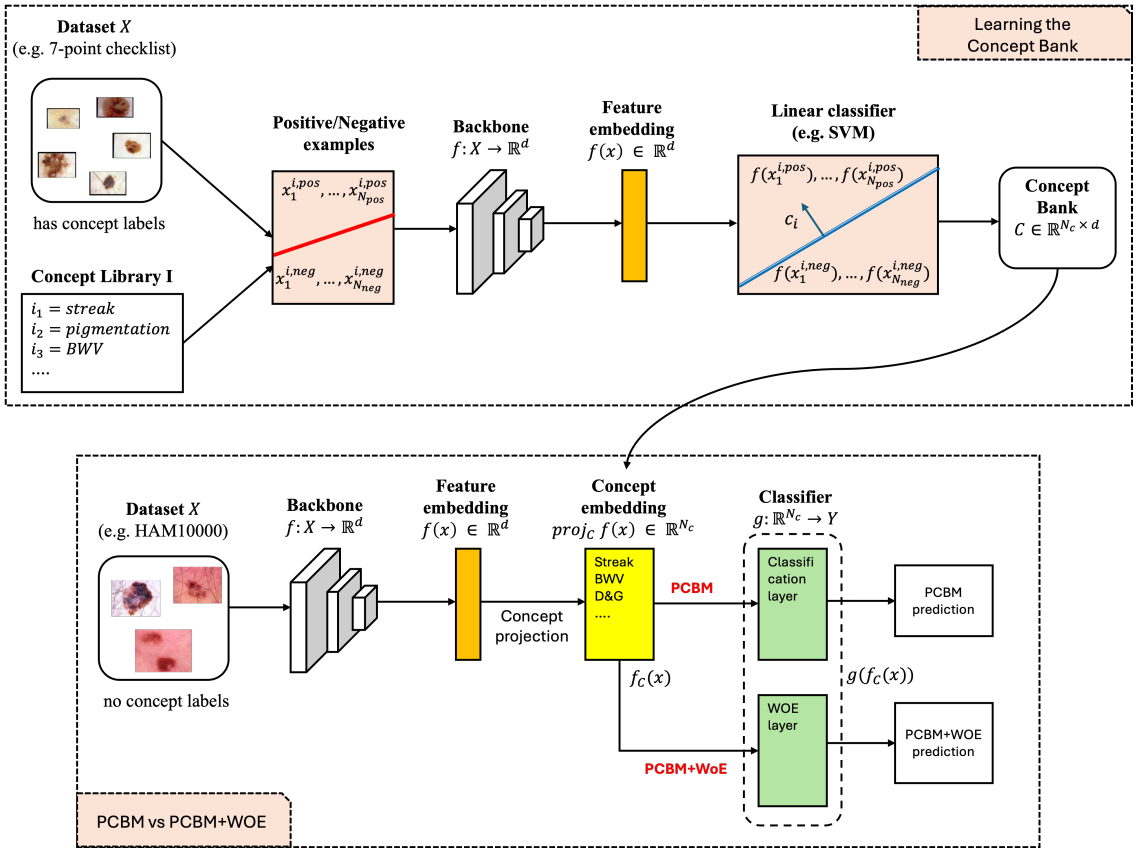



Figure 5.2: Supervised Concept Learning Model

the feature embedding space is often large (e.g., 2048 features). We then aim to reduce the dimensionality by using different techniques, either a reducer (for unsupervised learn-

ing) or a concept bank (for supervised learning). Next, we get a set of selected concepts and put them into a classifier layer $g : \mathbb{R}^{N_c} \rightarrow Y$ where $Y$ is the set of output classes, $N_c$ is the number of concepts. To implement the combined models ICE+WoE and PCBM+WoE, we replace the classifier layer of ICE and PCBM with the WoE model.

The following sections will describe in detail unsupervised concept learning (i.e., ICE [238]) and supervised concept learning (i.e., PCBM [234]) based on their main difference after the feature embedding layer. For unsupervised concept learning, we use a reducer to reduce the dimensionality of the feature embedding space. The learned concepts may not be related to the concepts used in the domain; for example, they could be invalid dermatological concepts or unknown to domain experts. In contrast, the supervised concept learning technique learns the concept bank and transfers it to create a concept embedding layer. This technique requires labelling concepts on the training images, which is expensive.

#### Invertible Concept-based Explanation (ICE)

ICE [238] is an unsupervised concept-based explanation approach that applies a reducer $R$ such as Non-negative Matrix Factorization (NMF) to reduce the dimensionality of the feature embedding space. NMF is a matrix factorization method that factorizes the feature matrix into two or more non-negative matrices [166]. We can also apply other dimensionality reduction methods such as PCA (Principal Component Analysis), which is a linear transformation method that projects the feature matrix into a lower-dimensional space. Eventually, we get a set of concepts $f_R(x) \in \mathbb{R}^{N_c}$ at the reducer layer.

#### Post-hoc Concept Bottleneck Model (PCBM)

PCBM [234] is a supervised concept-based explanation approach that learns the concept bank $\mathcal{C}$ and transfers it to create a concept embedding layer. We learn the concept bank by training a linear SVM to separate *positive concept examples* (contain the concept) and *negative concept examples* (do not contain the concept) in the embeddings based on CAV (Concept Activation Vector) approach [104]. Importantly, the dataset used to learn the

concept bank can be different from the dataset used in the task prediction. Therefore, when the training dataset does not have concept labels, we can annotate concepts by using another dataset that has concept labels and is in the same domain.

Formally, the concept library is defined as $I = \{i_1, i_2, ..., i_{N_c}\}$ where $N_c$ denotes the number of concepts. The concept library can be selected by domain experts or learned from the data [67]. For each concept $i$, there are a set of feature embeddings for positive examples $P_i = \{f(x_1^{i,pos}), ..., f(x_{N_{pos}}^{i,pos})\}$ and for negative examples $N_i = \{f(x_1^{i,neg}), ..., f(x_{N_{neg}}^{i,neg})\}$ where $N_{pos}$ and $N_{neg}$ are the number of positive and negative examples, respectively. Next, a linear SVM is trained using $P_i$ and $N_i$ to learn CAV (the normal to the SVM's decision boundary) for concept $i$, denoted as $c_i$. Finally, we get the concept matrix $\mathcal{C} \in \mathbb{R}^{N_c \times d}$ at the concept embedding layer.

Let $g : \mathbb{R}^{N_c} \to Y$ be the classifier. To learn the PCBM, we minimise the loss function:

$$\min_{g} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\mathcal{L}(g(f_{\mathcal{C}}(x)), y)] + \frac{\lambda}{N_c K} \Omega(g) \tag{5.1}$$

where $f_{\mathcal{C}}(x)$ is the projection onto the concept subspace, $\mathcal{L}(\hat{y}, y)$ is a loss function such as cross-entropy loss, $\Omega(g)$ is a complexity measure to regularize the model, and $\lambda$ is the regularization strength. In PCBM, a linear classifier such as a stochastic gradient descent model is implemented in this layer.

### Weight of Evidence

We replace the original classifier layer in ICE and PCBM with the WoE model. WoE is used to measure the weight of evidence for each concept, whereas using plain concepts does not provide this information. Similar to the WoE model in Section 4.2, we now calculate the weight of evidence for each concept. For hypothesis $h$ and concept $c_i$, which is equivalent to feature $x_i$ in 4.2, the weight of evidence woe is defined as follows.

$$\text{woe}(h \mid c_i) = \log \frac{P(c_i \mid h)}{P(c_i \mid Y_{-h})} = \log P(c_i \mid h) - \log P(c_i \mid Y_{-h}) \tag{5.2}$$

In the implementation of Visual Evaluative AI, we use the WoE *without* independence assumption.

## 5.3 Implementation

In this section, we show examples of concept-based explanations when applying *Visual Evaluative AI* on a skin cancer dataset (i.e., HAM10000 dataset [204]).

### 5.3.1 Basic Concepts in Skin Cancer Diagnosis

Dermatologists usually diagnose skin cancer by following ABCD rule [160] or 7-point checklist criteria [10, 100]. Comparing these two criteria, the 7-point checklist gives higher sensitivity, which is the accuracy of correctly identifying malignant lesions [9]. In this section, we provide an overview of the basic concepts in skin cancer diagnosis by focusing on the 7-point checklist criteria.

Following the terminology in Table 2 (page 18) [107] and [25], the seven concepts used in the 7-point checklist are: (1) atypical pigment network, (2) blue-white veil, (3) atypical vascular pattern, (4) irregular streaks, (5) irregular pigmentation, (6) irregular dots/globules, and (7) regression structures. This is a scoring system that assigns a score to each criterion, and the total score is used to classify the lesion as benign or malignant. Details of the scoring are described in Table 5.1.

| | **Criteria** | **7-point score** |
|---|---|---|
| Major criteria | Atypical pigment network | 2 |
| | Blue-white veil | 2 |
| | Atypical vascular pattern | 2 |
| Minor criteria | Irregular streaks | 1 |
| | Irregular pigmentation | 1 |
| | Irregular dots/globules | 1 |
| | Regression structures | 1 |

Table 5.1: 7-point checklist criteria

Based on the 7-point checklist above, Kawahara et al. [100] provide a 7-point criteria evaluation database called *Derm7pt* dataset. Using this dataset, we can extract 12 concepts included in Table 5.2. Each concept can indicate whether the lesion is benign or malignant.

Despite there being two common diagnostic outputs (benign and malignant), there

| Concept Name | Description |
|---|---|
| *Atypical Pigment Network* | concept activation for melanoma |
| *Typical Pigment Network* | concept activation for benign |
| *Blue Whitish Veil* | concept activation for melanoma |
| *Irregular Vascular Structures* | concept activation for melanoma |
| *Regular Vascular Structures* | concept activation for benign |
| *Irregular Pigmentation* | concept activation for melanoma |
| *Regular Pigmentation* | concept activation for benign |
| *Irregular Streaks* | concept activation for melanoma |
| *Regular Streaks* | concept activation for benign |
| *Regression Structures* | concept activation for melanoma |
| *Regular Dots and Globules* | concept activation for benign |
| *Irregular Dots and Globules* | concept activation for melanoma |

Table 5.2: 12 concepts used in the supervised method [231]

are seven classes of skin lesions that can be diagnosed based on the HAM10000 dataset [204], as shown in Table 5.3. Each class has different characteristics and is classified into either benign or malignant. It is worth noting that there is no clear answer for *actinic keratoses* [2]. This class can be considered as a pre-cancerous lesion, which can be classified as either benign or malignant. In Australia, actinic keratoses are common and usually treated as benign. However, in other countries, they can be considered as malignant. In this thesis, I follow the authors of the HAM10000 dataset, which classify actinic keratoses as malignant [205].

| Diagnosis | Lesion Type |
|---|---|
| Benign | (BKL) Benign keratosis-like lesions |
| | (DF) Dermatofibroma |
| | (NV) Melanocytic nevi |
| | (VASC) Vascular lesions |
| Malignant | (AKIEC) Actinic keratoses |
| | (BCC) Basal cell carcinoma |
| | (MEL) dermatofibroma |

Table 5.3: Seven output classes

[2] Based on a conversation with Prof. Peter Soyer

### 5.3.2 Dataset and Model Implementation

We use the **HAM10000 dataset** [204] to train all models (original CNN backbones, ICE, ICE+WoE, PCBM and PCBM+WoE). This dataset has a total of 10015 dermatoscopic images and seven output classes: Actinic keratoses (AKIEC), basal cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV) and vascular lesion (VASC). Among these seven classes, actinic keratoses (AKIEC), basal cell carcinoma (BCC), and melanoma (MEL) are malignant, while benign keratosis (BKL), dermatofibroma (DF), melanocytic nevi (NV) and vascular lesion (VASC) are benign. We choose the HAM10000 dataset instead of the 7-point checklist dataset [100] (2000 images) because HAM10000 is a larger dataset with more samples, which can help achieve more accurate classifiers. HAM10000 is also more generalised, as it was collected from multiple institutes, whereas the 7-point checklist dataset was collected from a single source. Lastly, HAM10000 is more well-known, with many more existing works and baseline models using this dataset.

We balance the dataset by applying Weighted Random Sampler [3] and data augmentation. Finally, each class has 1000 samples that were used for the training process, making a total of 7000 samples for seven classes. The test set is selected as a fraction of the original dataset (without augmentation). As in the original HAM10000, class DF has the lowest number of samples (i.e., 75 samples). Therefore, we choose 20 samples in each class for the test set, which represents 26% of class DF. We then have a total of 140 samples (20 samples $\times$ 7 classes) for the test set to evaluate the model performance.

Since images in the HAM10000 dataset do not have the concept labels, to get the concept labels for the PCBM model, we train Concept Activation Vectors (CAVs) [104] on the **7-point checklist dataset** [100] to obtain the concept library $I$. Followed the previous work [231, 234], we have 12 concepts: *Atypical Pigment Network*, *Typical Pigment Network*, *Blue Whitish Veil*, *Irregular Vascular Structures*, *Regular Vascular Structures*, *Irregular Pigmentation*, *Regular Pigmentation*, *Irregular Streaks*, *Regular Streaks*, *Regression Structures*, *Irregular Dots and Globules* and *Regular Dots and Globules*. The PCBM model then used the trained CAVs based on these 12 concepts and applied that to extract the concept. In more

[3] https://pytorch.org/docs/stable/_modules/torch/utils/data/sampler.html

| Concept Name | Positive Samples | Negative Samples |
|---|---|---|
| Atypical Pigment Network | 230 | 781 |
| Typical Pigment Network | 381 | 630 |
| Blue Whitish Veil | 195 | 816 |
| Irregular Vascular Structures | 71 | 940 |
| Regular Vascular Structures | 117 | 894 |
| Irregular Pigmentation | 305 | 706 |
| Regular Pigmentation | 118 | 893 |
| Irregular Streaks | 251 | 760 |
| Regular Streaks | 107 | 904 |
| Regression Structures | 253 | 758 |
| Irregular Dots and Globules | 448 | 563 |
| Regular Dots and Globules | 334 | 677 |

Table 5.4: The number of positive and negative samples for each concept in the concept bank.

detail, we show the number of positive and negative samples for each concept based on the 7-point checklist dataset in Table 5.4, which is used to learn the concept embeddings in the PCBM model. The number of positive samples refers to how many images contain the concept, while the number of negative samples refers to how many images do not contain the concept. In our final chosen model, for each concept, we select 50 positive samples (contain the concept) and 50 negative samples (do not contain the concept). The learning rate was set to 0.01 and ridge regression was used at the classifier layer of PCBM.

### 5.3.3 Concept-based Explanations

Figure 5.3: Reddish structures

We iteratively try different numbers of concepts and work closely with a domain expert to refine and annotate the concepts. The domain expert is an academic dermatolo-

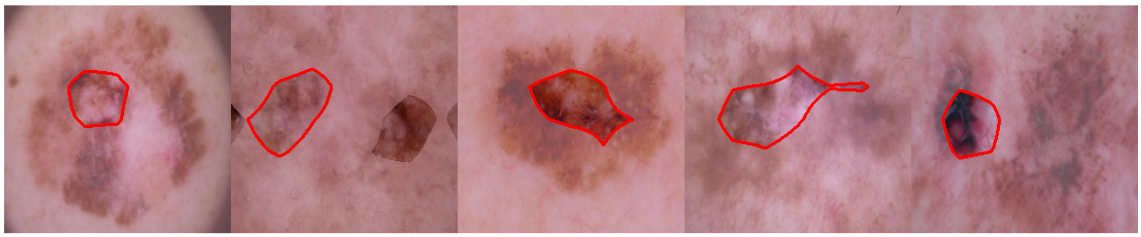


Figure 5.4: Irregular pigmentation

Figure 5.5: Irregular dots and globules

Figure 5.6: Whitish veils

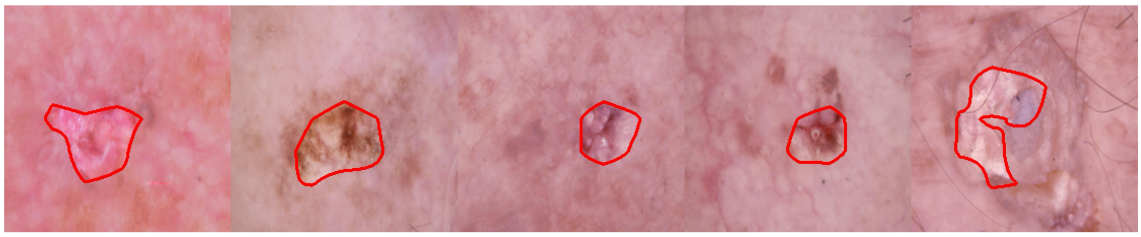


Figure 5.7: Irregular pigmentation

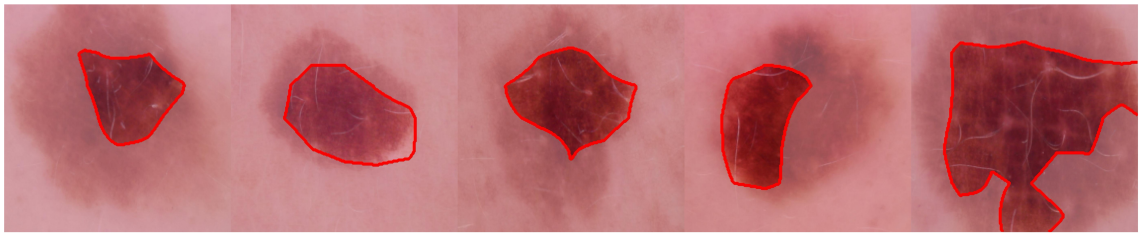


Figure 5.8: Dark irregular pigmentation

Figure 5.9: Lines (Hair)

gist with more than 40 years of experience in the field. We decided to use the ICE+WoE model to generate concept-based explanations for our human study because currently the labelled concepts found by the PCBM+WoE model are often wrong and considered unreliable by the expert. One possible reason could be the Out-of-Distribution (OOD) problem, as the concept bank is trained on the 7-point checklist dataset [100] and then applied to the HAM10000 dataset [204]. Furthermore, ICE+WoE also achieves higher performance than PCBM+WoE in our computational experiments in Section 5.4. Therefore, it is important to note that the concepts in our human experiment are not labelled by the model, but are labelled by the expert.

Figures 5.3 to 5.9 show seven concept explanations found by ICE [238] in the HAM10000 dataset [204]. Each concept is represented by five examples (instances) in the training set that segment areas of interest, specifically, the red polygon outlines. These outlines are segmented using the ICE method, not hand-drawn. For instance, in Figure 5.3, the segmentations detected by the model are reddish structures. These examples are selected by getting the best feature importance, which can be estimated using the method in TCAV [104]. In this case, we select five best examples that have the highest feature importance. We also ensure that the lesions are different from each other among the examples. The examples are arranged in ascending order of feature importance scores, from left to right. For instance, although we could not find a concept name that matches all five examples in Figure 5.9, we chose to annotate it as *lines (hair)* because the three rightmost examples, i.e., the ones with the highest scores, focus on the hair in the images.

Specifically, the found concepts are (1) Reddish structures, (2) Irregular pigmentation, (3) Irregular dots and globules, (4) Whitish veils, (5) Irregular pigmentation, (6) Dark irregular pigmentation and (7) Lines (Hair). These concepts are identified as important

by the classifier to make a decision. We then use these seven concepts to generate the evidence by applying the WoE [148] and conduct a human experiment in Section 5.5. Study participants can decide whether to use the evidence generated by the model to make the final decision. For example, the model detects *lines* (Figure 5.9) as an important concept, though it is a *confounding feature* and should be ignored in making the diagnosis.

## 5.4 Computational Experiments

In this section, we evaluate the computational performance of the combined models ICE+WoE and PCBM+WoE on the skin cancer dataset (HAM10000) [204]. We compare them in terms of accuracy and investigate the impact of the number of concepts on the performance of ICE+WoE. The model configurations are explained in Section 5.3.

### 5.4.1 Experiment Design

We apply different CNN backbones (Resnet50, ResneXt50 and Resnext152) to train the original CNN models, ICE, ICE+WoE, PCBM and PCBM+WoE. Based on each CNN backbone, we compare the original backbone model with ICE, ICE+WoE, PCBM and PCBM+WoE. We use the F1-score metric to evaluate the performance of the models. The F1-score is calculated as the harmonic mean of precision and recall, which is defined as:

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

where True Positive (TP) is the number of correctly predicted malignant cases; False Positive (FP) is the number of incorrectly predicted malignant cases; True Negative (TN) is the number of correctly predicted benign cases; False Negative (FN) is the number of

incorrectly predicted benign cases.

### 5.4.2 Computational Results

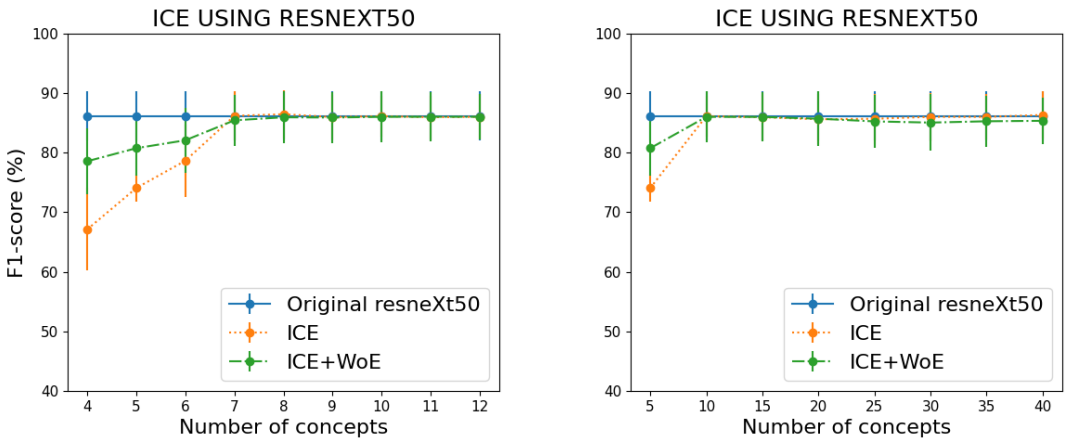



Figure 5.10: F1-score of ICE, ICE+WoE and the original ResneXt50 over different number of concepts. The left figure shows the performance of ICE and ICE+WoE with a small number of concepts (4-12), while the right figure shows the performance of ICE and ICE+WoE with a larger range of number of concepts (5-100).

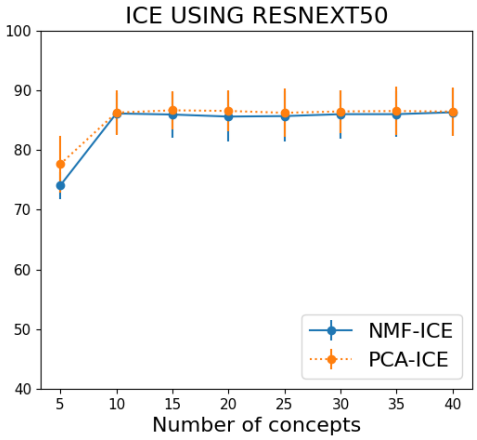



Figure 5.11: Comparing different reducers NMF and PCA.

| CNN Backbone | Model | Precision ↑ | Recall ↑ | F1-Score ↑ |
|---|---|---|---|---|
| Resnet50 | Backbone | $83.08 \pm 5.98$ | $85.33 \pm 6.20$ | $\mathbf{84.04 \pm 5.01}$ |
| | ICE(7) | $73.34 \pm 8.69$ | $87.50 \pm 10.04$ | $78.99 \pm 4.91$ |
| | ICE(7)+WoE | $80.13 \pm 5.44$ | $82.00 \pm 6.81$ | $80.85 \pm 4.55$ |
| | PCBM(12) | $73.93 \pm 8.94$ | $82.08 \pm 12.67$ | $76.58 \pm 6.31$ |
| | PCBM(12)+WoE | $80.73 \pm 5.21$ | $84.25 \pm 3.35$ | $82.32 \pm 2.98$ |
| ResneXt50 | Backbone | $85.46 \pm 4.63$ | $87.25 \pm 6.31$ | $86.20 \pm 4.18$ |
| | ICE(7) | $84.23 \pm 5.49$ | $88.58 \pm 5.41$ | $\mathbf{86.20 \pm 4.11}$ |
| | ICE(7)+WoE | $84.73 \pm 5.00$ | $86.33 \pm 4.76$ | $85.45 \pm 4.25$ |
| | PCBM(12) | $78.93 \pm 8.28$ | $83.17 \pm 14.43$ | $79.83 \pm 8.28$ |
| | PCBM(12)+WoE | $84.48 \pm 4.86$ | $85.50 \pm 3.98$ | $84.92 \pm 3.64$ |
| Resnet152 | Backbone | $84.49 \pm 6.48$ | $86.08 \pm 5.70$ | $\mathbf{84.96 \pm 3.09}$ |
| | ICE(7) | $78.30 \pm 8.11$ | $87.42 \pm 7.48$ | $82.10 \pm 4.37$ |
| | ICE(7)+WoE | $81.21 \pm 4.90$ | $85.08 \pm 5.14$ | $83.01 \pm 4.13$ |
| | PCBM(12) | $76.49 \pm 7.75$ | $87.08 \pm 5.15$ | $81.09 \pm 4.21$ |
| | PCBM(12)+WoE | $82.97 \pm 5.37$ | $84.83 \pm 4.04$ | $83.73 \pm 2.99$ |

Table 5.5: Performance for the original CNN model, ICE, ICE+WoE, PCBM and PCBM+WoE. The ICE model uses an NMF (non-negative matrix factorization) reducer. ICE(7) represents the ICE model with 7 different concepts. PCBM(12) is the PCBM model with 12 labelled concepts. *mean* $\pm$ *standard deviation* of the performance are reported over 20 random seeds. Winners are indicated in bold.

**ICE+WoE and PCBM+WoE achieved comparable performance to the original CNN models**

Table 5.5 reports the performance of ICE(7), ICE(7)+WoE, PCBM(12) and PCBM(12)+WoE using three different CNN backbone models (Resnet50, Resnet152 [79] and ResneXt50 [230]). We select 12 concepts for PCBM based on previous work [231, 234]. For ICE, we run experiments with a number of concepts ranging from 5 to 40. As shown in Figure 5.10, performance peaks at 7 concepts. Therefore, the final comparison in this table is made between ICE(7) and PCBM(12).

The results show that ICE(7)+WoE and PCBM(12)+WoE achieve comparable performance to the original CNN models. Particularly, with ResneXt50, the F1-score of ICE(7)+WoE and PCBM(12)+WoE are $85.45 \pm 4.25$ and $84.92 \pm 3.64$, respectively, while the original ResneXt50 has an F1-score of $86.20 \pm 4.18$. Therefore, ICE(7)+WoE (using 7 features) and PCBM(12)+WoE (using 12 features) have similar performance compared to the original

ResneXt50 with 2048 features.

Similar to the findings in [238], when we compare the performance using different reducers as in Figure 5.11, NMF and PCA (principal component analysis), PCA provided the best performance but could be less interpretable compared to NMF.

**Having more concepts did not lead to better accuracy**

Figure 5.10 shows the performance of the original ResneXt50, ICE and ICE+WoE over different numbers of concepts from 5 concepts to 40 concepts. Two figures from the left show the performance of ICE using the NMF reducer. When there are 5 concepts, ICE(5)+WoE ($80.77 \pm 4.56$) has a significantly higher F1-score than ICE(5) ($74.08 \pm 2.26$) ($p = 8.48 \times 10^{-}7 < 0.001$, $d = 1.857$). Since we have 2048 features at the classifier layer of ResneXt50, ResneXt50 outperforms ICE(5)+WoE and ICE(5) significantly ($p < 0.001$). But the performance of both ICE+WoE and ICE match the performance of the original ResneXt50 when we have at least 7 concepts. Particularly, with *as few as 7 concepts*, ICE and ICE+WoE achieve similar performance to the original ResneXt50 using 2048 features. The performance of ICE and ICE+WoE also stopped improving at 7 concepts with a backbone of ResneXt50. The reason is that when we apply a reducer in ICE (e.g. NMF), some important concepts are detected at first. Then after we increase the number of concepts, some noisy concepts are detected, which could lead to a slight drop in the performance. Eventually, all important concepts are found and match the performance of the original CNN model.

In summary, the results show that with a few number of concepts (i.e., 7 concepts), we can achieve comparable performance compared to the original CNN models. Therefore, this indicates the accuracy of the evidence being generated, which is potentially useful to the decision-makers. Importantly, despite the concept-based models (ICE(7), ICE(7)+WoE, PCBM(12) and PCBM(12)+WoE) being slightly less accurate than the CNN backbones, it would also be much easier for users to interpret and evaluate the evidence by not showing too many concepts.

### 5.4.3 Ablation Studies

#### Concept Bank Evaluation

| Model | Learning Rate | Number of Samples (Positive or Negative) | Test Accuracy |
|---|---|---|---|
| resneXt50 | 0.001 | 25 | $0.63 \pm 0.08$ |
| resneXt50 | 0.001 | 50 | $0.64 \pm 0.06$ |
| resneXt50 | 0.001 | 75 | $0.65 \pm 0.05$ |
| resneXt50 | 0.001 | 100 | $0.65 \pm 0.05$ |
| resneXt50 | 0.01 | 25 | $0.65 \pm 0.07$ |
| resneXt50 | 0.01 | 50 | $0.67 \pm 0.06$ |
| resneXt50 | 0.01 | 75 | $0.69 \pm 0.06$ |
| resneXt50 | 0.01 | 100 | $0.70 \pm 0.05$ |
| resneXt50 | 0.1 | 25 | $0.66 \pm 0.08$ |
| resneXt50 | 0.1 | 50 | $0.68 \pm 0.06$ |
| resneXt50 | 0.1 | 75 | $0.71 \pm 0.07$ |
| resneXt50 | 0.1 | 100 | $0.72 \pm 0.07$ |

Table 5.6: The performance of the concept bank using different learning rates and number of samples (the number for each positive or negative sample).

We evaluate the performance of the concept bank using the 7-point checklist dataset [100] as shown in Table 5.6 with different learning rates and the number of samples. With the same learning rate, the performance is very similar despite an increase in the number of learned samples. It is important to note that the concept *Irregular Vascular Structures* only has 71 positive samples (Table 5.4). Therefore, we chose a maximum of 100 samples in this ablation study to ensure that we do not rely on many augmented samples to learn the concept bank.

#### Unsupervised learning (ICE) and Supervised learning (PCBM)

In Table 5.7, we compare the performance between ICE(12) and PCBM(12) using the same classification layer (ridge) and the same number of concepts (12 concepts). The results show that ICE(12)+Ridge outperforms PCBM(12)+Ridge in all three CNN backbones. The reason is that the concept bank in PCBM is learned from the 7-point checklist dataset,

| CNN Backbone | Model | Precision ↑ | Recall ↑ | F1-Score ↑ |
|---|---|---|---|---|
| Resnet50 | ICE(12)+Ridge | $81.94 \pm 4.76$ | $85.50 \pm 6.40$ | $\mathbf{83.64 \pm 5.25}$ |
| | PCBM(12)+Ridge | $73.93 \pm 8.94$ | $82.08 \pm 12.67$ | $76.58 \pm 6.31$ |
| ResneXt50 | ICE(12)+Ridge | $86.08 \pm 4.73$ | $87.50 \pm 4.91$ | $\mathbf{86.70 \pm 4.01}$ |
| | PCBM(12)+Ridge | $78.93 \pm 8.28$ | $83.17 \pm 14.43$ | $79.83 \pm 8.28$ |
| Resnet152 | ICE(12)+Ridge | $82.49 \pm 5.13$ | $86.92 \pm 3.64$ | $\mathbf{84.53 \pm 3.25}$ |
| | PCBM(12)+Ridge | $76.49 \pm 7.75$ | $87.08 \pm 5.15$ | $81.09 \pm 4.21$ |

Table 5.7: A comparison between the unsupervised learning model (ICE) and supervised learning model (PCBM). Both models use 12 concepts and have the same classification layer (ridge). *mean* ± *standard deviation* of the performance are reported over 20 random seeds. Winners are indicated in bold.

while ICE specifically focuses on the HAM10000 dataset and learns the concepts that are most important for the classifier. Therefore, the concepts learned by ICE are more accurate and relevant to the HAM10000 dataset.

**Different Classifier Layers for ICE**

| CNN Backbone | Model | Precision ↑ | Recall ↑ | F1-Score ↑ |
|---|---|---|---|---|
| Resnet50 | ICE(7) | $73.34 \pm 8.69$ | $87.50 \pm 10.04$ | $78.99 \pm 4.91$ |
| | ICE(7)+Ridge | $80.05 \pm 6.19$ | $85.42 \pm 9.63$ | $\mathbf{82.08 \pm 4.56}$ |
| | ICE(7)+GNB | $80.13 \pm 5.44$ | $82.00 \pm 6.81$ | $80.85 \pm 4.55$ |
| | ICE(7)+WoE | $80.13 \pm 5.44$ | $82.00 \pm 6.81$ | $80.85 \pm 4.55$ |
| ResneXt50 | ICE(7) | $84.23 \pm 5.49$ | $88.58 \pm 5.41$ | $86.20 \pm 4.11$ |
| | ICE(7)+Ridge | $84.85 \pm 5.06$ | $88.00 \pm 5.58$ | $\mathbf{86.24 \pm 3.86}$ |
| | ICE(7)+GNB | $84.73 \pm 5.00$ | $86.33 \pm 4.76$ | $85.45 \pm 4.25$ |
| | ICE(7)+WoE | $84.73 \pm 5.00$ | $86.33 \pm 4.76$ | $85.45 \pm 4.25$ |
| Resnet152 | ICE(7) | $78.30 \pm 8.11$ | $87.42 \pm 7.48$ | $82.10 \pm 4.37$ |
| | ICE(7)+Ridge | $80.99 \pm 5.58$ | $87.33 \pm 6.68$ | $\mathbf{83.82 \pm 4.34}$ |
| | ICE(7)+GNB | $81.21 \pm 4.90$ | $85.08 \pm 5.14$ | $83.01 \pm 4.13$ |
| | ICE(7)+WoE | $81.21 \pm 4.90$ | $85.08 \pm 5.14$ | $83.01 \pm 4.13$ |

Table 5.8: Different classification layers in ICE. *mean* ± *standard deviation* of the performance are reported over 20 random seeds. Winners are indicated in bold.

Since ICE uses the weights of the original CNN backbone models, we conduct a

further ablation experiment by replacing the classifier layer of ICE with the Gaussian Naive Bayes (ICE+GNB) to compare with ICE+WoE. The results in Table 5.5 show that ICE(7)+WoE has similar performance to ICE(7)+GNB. The reason is that our implementation using WoE and GNB are both Naive Bayes methods so they use similar loss functions when learning the concept scores.

In [238], the original ICE estimates weights by applying the method from TCAV [104]. As shown in Table 5.8, ICE(7) refers to this original ICE without using any classifier layer to learn the weights and having 7 concepts after the reducer. We conduct an ablation test by comparing the performance of the original ICE with the performance of using classifiers such as Ridge, GNB and WoE, as shown in the flow 5.1. The results show that ICE(7)+Ridge is slightly better than others in terms of classification accuracy.

## 5.5 Human Experiment

In this section, we conduct a human experiment to evaluate the effectiveness of the recommendation-driven and the hypothesis-driven approach in skin cancer diagnosis. We aim to investigate the impact of these two approaches on the decision-making process, accuracy and user satisfaction.

### 5.5.1 Study Design

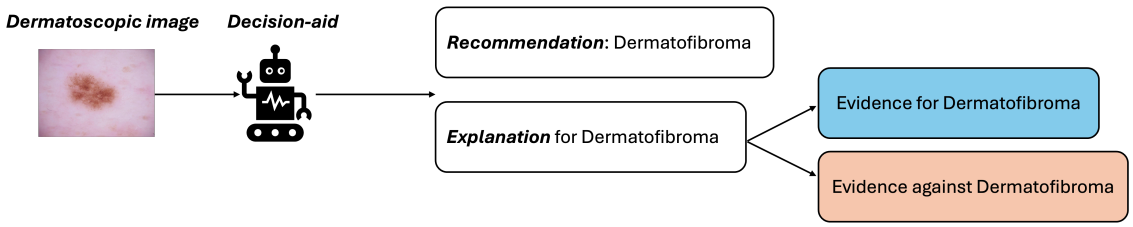



Figure 5.12: The recommendation-driven flow

The goal of this experiment is to understand the differences in terms of decision accuracy, decision time and user satisfaction between the recommendation-driven and hypothesis-driven approaches in supporting skin cancer diagnosis. We test only these

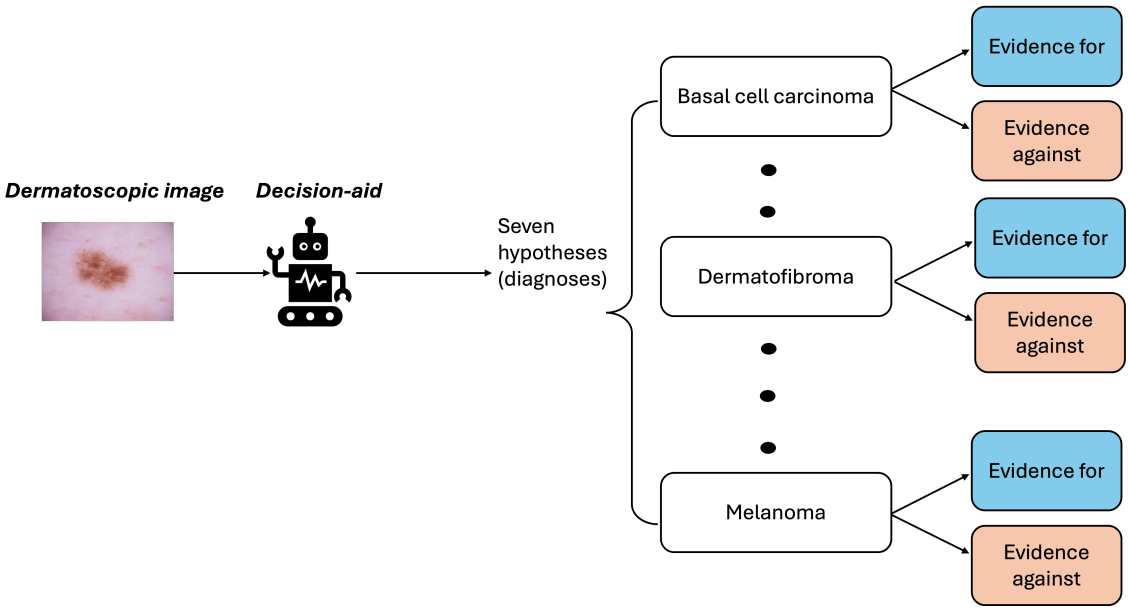



Figure 5.13: The hypothesis-driven flow

two conditions because the study follows a within-subject design that requires participants to complete all conditions. Adding another condition would make the study too long for participants and could lead to fatigue. Moreover, these two conditions are the main approaches in decision support systems that we aim to compare. The study is conducted on a web application called *EvaSkan* (Evaluative Skin Cancer), where participants are asked to make skin cancer diagnoses using the two approaches. An example EvaSkan interface is shown in Figure 5.14. To see the full protocol of our experiment and the web interfaces being used, please refer to the supplementary document C.

There are three phases in the study. In phase 1, we collect the demographic information of the participants including their roles, background, years of experience, and whether they are familiar with AI and skin cancer diagnosis. In phase 2, we conduct a within-subject design study. Study participants use two web interfaces for two conditions (*recommendation-driven* and *hypothesis-driven*) and perform skin cancer diagnosis tasks on a web page. In the *recommendation-driven* condition, participants are given the AI prediction for the skin cancer diagnosis and the explanation of that prediction. In the *hypothesis-driven* condition, participants are given explanations for and against all possible hypotheses. The web page records the interaction log of participants.

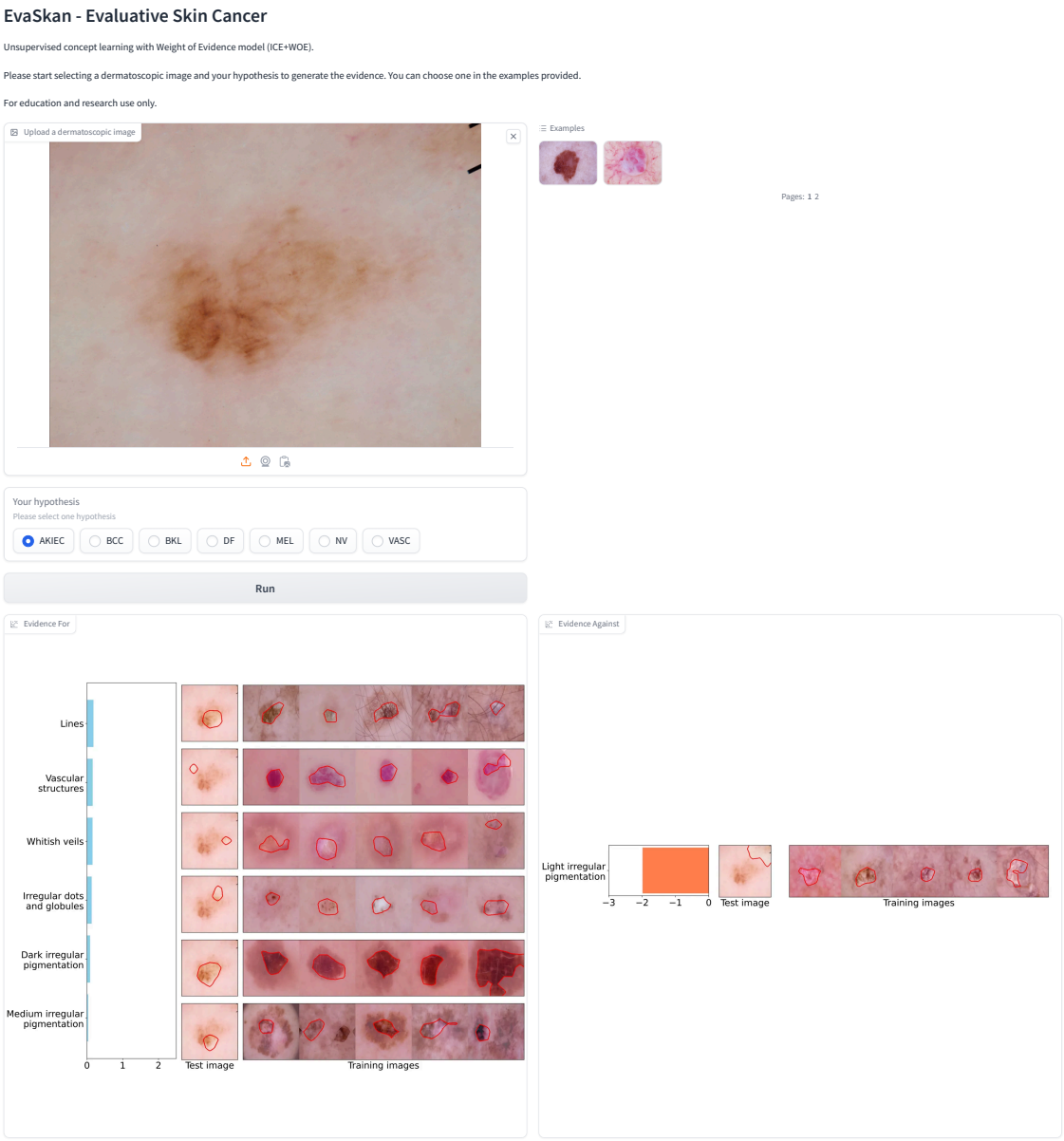



Figure 5.14: A screenshot of the EvaSkan web application

There are a total of sixteen different tasks (questions) for two conditions, eight of which are in one condition, and eight in the other. Sixteen questions are uniformly distributed into four categories: (1) where the model gives correct predictions with high uncertainty, (2) where the model gives correct predictions with low uncertainty, (3) where the model gives wrong predictions with high uncertainty and (4) where the model gives wrong predictions with low uncertainty.

Conditions and tasks are randomly counterbalanced. Specifically, the order of the conditions is randomised and the order of the tasks within a condition is randomised. Further, out of sixteen images, we also randomly select images for each condition to minimise selection bias of using the same set of images in each condition. At the end of phase 2, we ask them to evaluate their preferences in these two conditions using bipolar scale questions as follows:

1. In control: Scale these conditions based on how much you are in control of the decision-making process.
2. Decision-making: Scale these conditions based on how helpful it is to you to make the diagnosis.
3. Ease of use: Scale these conditions based on how easy it is to use.
4. Error detection: Scale these conditions based on how easy it is to spot mistakes in the decision aid.

Finally, in Phase 3, we conducted a semi-structured interview by asking participants to reflect on how they made the diagnoses in Phase 2 using a think-aloud protocol and open questions about the design of our decision aids (DAs). The questions are included in the supplementary document C. The study was pre-registered [4] and received ethics approval (ID: 23208) before data collection. This study requires a maximum of one hour to finish.

### 5.5.2 Study Hypotheses

Our research hypotheses are as follows:

- **H1**: The hypothesis-driven approach will take more time to make decisions than the recommendation-driven approach.
- **H2**: The hypothesis-driven approach will help study participants make more accurate decisions than the recommendation-driven approach.

[4] https://osf.io/d9csz

- **H3**: Study participants will be more satisfied with the hypothesis-driven approach than the recommendation-driven approach. More specifically,
  - **H3a**: Participants feel they have more control of the decision-making process when using the hypothesis-driven approach compared to the recommendation-driven approach.
  - **H3b**: Participants feel the hypothesis-driven approach is more helpful in making a diagnosis than the recommendation-driven approach.
  - **H3c**: Participants find the hypothesis-driven approach is easier to use than the recommendation-driven approach.
  - **H3d**: Participants find it is easier to spot mistakes in the decision-aid when using the hypothesis-driven approach compared to the recommendation-driven approach.

### 5.5.3 Participants

We recruit individuals with a background in skin cancer through our professional networks. Participants receive 25 AUD upon completing the study. There are a total of **14** participants whose details are summarised in Table 5.9. Gender-wise, there are **7** females and **7** males.

Some participants have used AI decision support tools in research settings, such as *Canfield Dermoscopy Explained Intelligence (DEXI)*, *Canfield's Dermagraphix*, *FotoFinder MoleAnalyzer* and *Lesion Change Detection model*. None of them have used AI tools in clinical settings. We classify participants into either *experienced* or *inexperienced* as in Table 5.9 (*Experience in Skin Cancer Diagnosis*). Experienced participants are individuals who have received clinical training (e.g., resident doctors, senior house officer, principal house officer, senior melanographer [5]). Inexperienced participants include PhD students and postdoctoral researchers working in the skin cancer field, but are not specifically trained to become clinicians.

[5] https://www.careers.health.qld.gov.au/medical-careers/career-structure

| ID | Role | Years of Experience | Research/Work Field | Experience in Skin Cancer Diagnosis |
|---|---|---|---|---|
| P0 | PhD Student & Senior Research Technician | 5 | Dermatology | No |
| P1 | PhD Student | 2 | Melanoma Detection | No |
| P2 | PhD Student & Research Assistant | 5 | Melanoma Detection | No |
| P3 | Resident Doctor | 2 | Cutaneous Phenotyping | Yes |
| P4 | Melanographer | 1 | Melanography | No |
| P5 | Resident Doctor | 2 | Dermatology | Yes |
| P6 | Melanographer | 10 | Dermatology | Yes |
| P7 | Postdoc | 12 | AI Implementation in Skin Cancer | No |
| P8 | Postdoc | 2 | Biostatistics | No |
| P9 | Principal House Officer | 1 | Dermatology | Yes |
| P10 | Resident Doctor | 3 | Melanoma Prognosis | Yes |
| P11 | Senior House Officer | 1 | Dermatology | Yes |
| P12 | Principal House Officer | 2 | Dermatology | Yes |
| P13 | Senior House Officer | 3.5 | Melanoma | Yes |

Table 5.9: Study participant's details. *Year of Experience* refers to the years they have spent in that role.

### 5.5.4 Experiment Variables

This study has two independent variables (two *conditions*): recommendation-driven and hypothesis-driven. To compare between these two conditions, we took the following measures:

1. *Time spent on each task (Instance time)*: We measure the time participants spend on each diagnosis task in Phase 2 (one out of sixteen tasks in total). Then, we also calculate the total time spent on all tasks in one condition (*Total time*);

2. *Brier score* We measure the effectiveness of task performance by evaluating both the

confidence of the participant and the correctness of the answer. The formula is:

$$\mathrm{BS}_p = \frac{1}{N}\sum_{i=1}^{N}(C_{p,i} - A_{p,i})^2 \tag{5.3}$$

where: $C_{p,i}$ is the confidence level of participant $p$ in question $i$, ranging from 0 to 1; $A_{p,i}$ is the answer score of participant $p$ in question $i$, either 0 (wrong answer) or 1 (right answer). $N$ is the number of questions, which is $N = 8$ in one condition;

3. *Selected hypotheses*: In the hypothesis-driven condition, we record which hypotheses are being checked. We then calculate the *percentage of selected hypotheses* by dividing the number of selected hypotheses by the total number of hypotheses, which is 7. This measure can indicate whether study participants used their prior knowledge in the decision-making process;

4. *Self-reported bipolar scales*: We ask participants to evaluate their preferences in these two conditions in terms of *in control*, *decision-making*, *ease of use* and *error detection* using bipolar scale questions as in the study design.

### 5.5.5 Quantitative Results

We now show the performance and self-reported bipolar scales of participants when interacting with two conditions, recommendation-driven and hypothesis-driven.

#### Performance

Participants' performances are summarised in Table 5.10. Regarding the time spent to complete the task, the hypothesis-driven condition takes more time than the recommendation-driven condition for all participants significantly. Moreover, experienced participants take more time to evaluate each instance (task) in both interfaces, suggesting that they are more careful in making decisions. Moreover, for experienced participants, the distribution of Brier scores is tighter in the hypothesis-driven, with most participants achieving scores close to 0. This result implies that experienced participants performed better with the hypothesis-driven, whereas inexperienced participants had better performance with

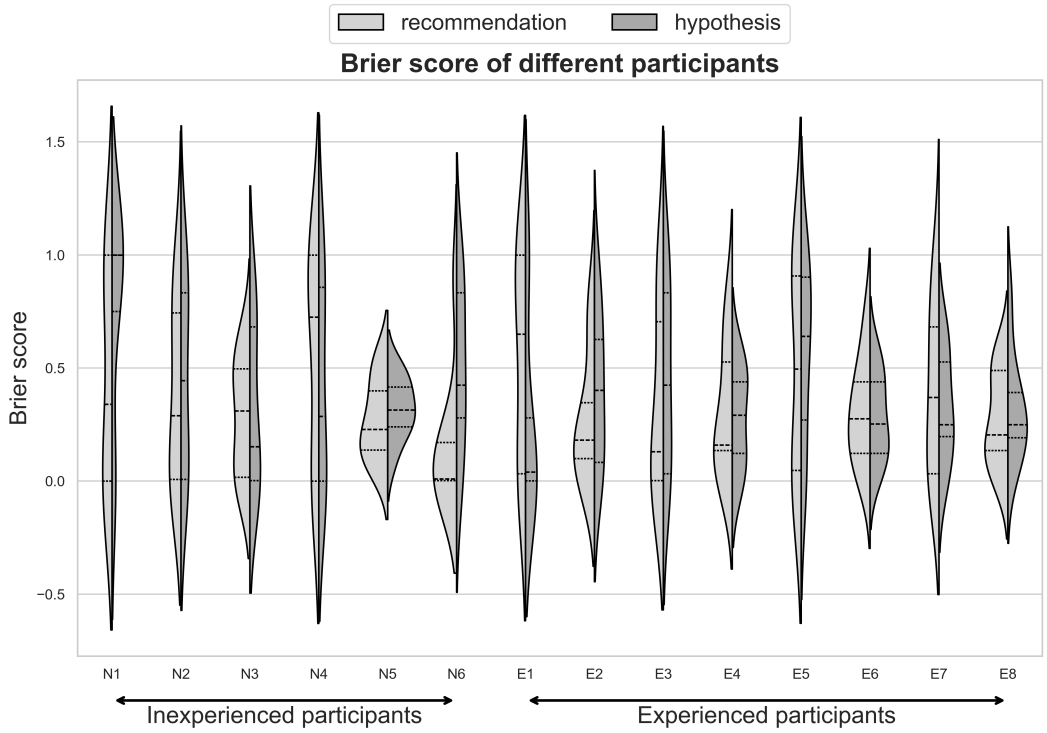



Figure 5.15: Brier score for all participants. The IDs here are different from the IDs in Table 5.9 to protect participants' privacy. Participants are separated into *Experienced participants* and *Inexperienced participants* based on the classification in *Experience in Skin Cancer Diagnosis* in Table 5.9.

| | | R | H | Wilcoxon t-test |
|---|---|---|---|---|
| All | Total time (s) ↓ | **449.78 ± 301.95** | 580.68 ± 310.91 | $\mathbf{p = 0.05, r = 0.53}$ |
| | Instance time (s) ↓ | **56.22 ± 45.59** | 72.59 ± 48.71 | $\mathbf{p < 0.001, r = 0.37}$ |
| | Brier score ↓ | **0.36 ± 0.36** | 0.40 ± 0.36 | $p = 0.35, r = 0.09$ |
| Experienced | Total time (s) ↓ | **503.21 ± 377.06** | 618.11 ± 369.43 | $p = 0.31, r = 0.36$ |
| | Instance time (s) ↓ | **62.90 ± 53.96** | 77.26 ± 55.05 | $\mathbf{p = 0.005, r = 0.35}$ |
| | Brier score ↓ | 0.37 ± 0.35 | **0.36 ± 0.32** | $p = 0.94, r = 0.01$ |
| Inexperienced | Total time (s) ↓ | **378.54 ± 165.38** | 530.78 ± 234.57 | $p = 0.09, r = 0.68$ |
| | Instance time (s) ↓ | **47.32 ± 29.38** | 66.35 ± 38.37 | $\mathbf{p = 0.006, r = 0.40}$ |
| | Brier score ↓ | **0.35 ± 0.38** | 0.46 ± 0.40 | $p = 0.16, r = 0.20$ |

Table 5.10: Performance of participants in terms of time required to complete all tasks and the Brier score. *R*: Recommendation-driven, *H*: Hypothesis-driven. Winners/significances are in bold.

the recommendation-driven interface. Note that there is no significant difference in the Brier score between the two conditions for all participants. However, the time required to complete all tasks is very close to the significance level ($p = 0.05$) between the two conditions for all participants. Based on the result, we can accept **H1** (time) and reject **H2** (accuracy).

### Subjective Bipolar Scales

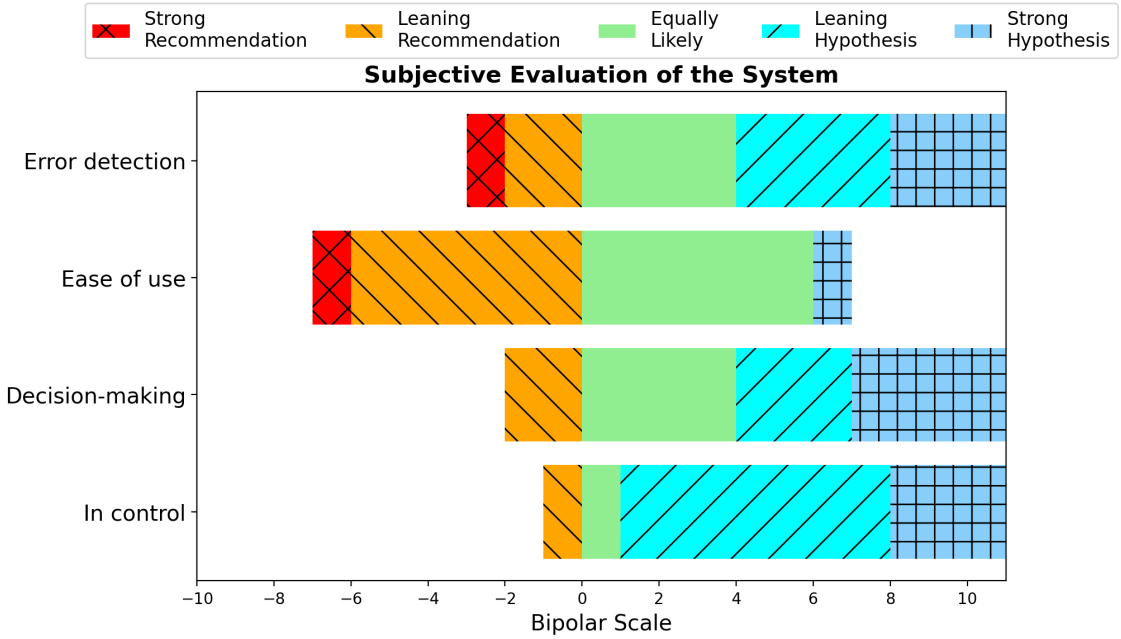



Figure 5.16: Bipolar scale counts of the approach's subject metrics.

| Metric | Mean ± std | One sample t-test |
|---|---|---|
| In control | 2.71 ± 1.82 | $\mathbf{p < 0.001, d = 1.49}$ |
| Decision-making | 1.57 ± 2.53 | $\mathbf{p = 0.037, d = 0.62}$ |
| Ease of use | -1.07 ± 2.27 | $p = 0.100, d = 0.472$ |
| Error detection | 0.93 ± 2.79 | $p = 0.234, d = 0.333$ |

Table 5.11: Results of Subjective Bipolar Scales (-5 = Recommendation-driven is the best; 0 = equally likely, 5 = Hypothesis-driven is the best). Significances are in bold.

From Figure 5.16 and Table 5.11, participants show preferences in the hypothesis-driven interface in terms of *in control*, *decision-making* and *error detection*. However, the

two interfaces have no significant difference in *ease of use*, with a slight preference for the recommendation-driven interface. This result is consistent with the quantitative results in Table 5.10 where the recommendation-driven interface is faster to complete. We can accept **H3a** (in control) and **H3b** (helpful in decision-making), but reject **H3c** (ease of use) and **H3d** (error detection).

### Selected Hypotheses

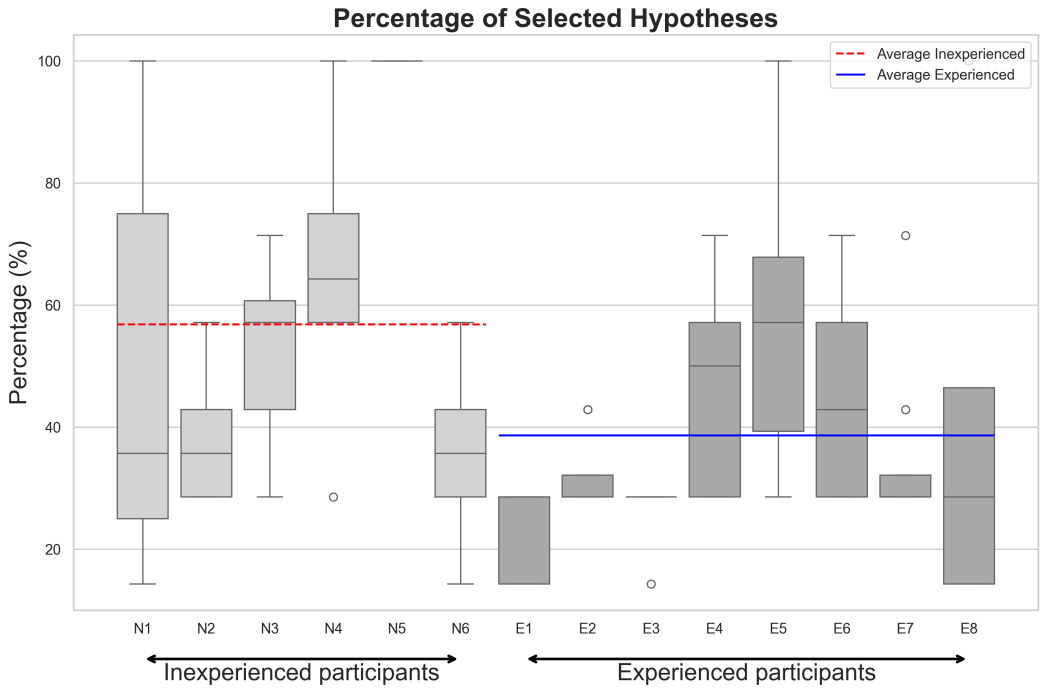



Figure 5.17: Percentage of selected hypotheses by participants.

Figure 5.17 shows the percentage of selected hypotheses by participants in the hypothesis-driven condition. Experienced participants tend to select fewer hypotheses than inexperienced participants when checking the evidence. This result implies that experienced participants apply their thoughts about the diagnosis and only check the evidence for the hypotheses they think are relevant. Inexperienced participants, on the other hand, tend to check more hypotheses as they do not have the base knowledge to decide the possible hypotheses by themselves.

### 5.5.6 Qualitative Results

In this section, we discuss the interview results after participants experience the two interfaces. The results show that they are consistent with previous literature. More importantly, the qualitative results provide insights into the differences between the recommendation-driven and the hypothesis-driven interfaces, which have not been addressed before.

#### Perceived Accuracy and Reliability of the Decision Support

Most participants commented that it was difficult to say about the accuracy because they did not know about the ground truth of the test lesions. A participant said that their decision would be very different without the AI information, so it would be helpful in the future to evaluate whether there is a difference between having a decision support and without having a decision support.

> "*I am pretty sure having this information would be a great help for me as a beginner.*" – inexperienced user

> "*A suggested diagnosis can distract me from making my own decision [...] I tried to look at the image first and then look at the diagnosis provided by the AI.*" – inexperienced user

Regarding the AI can help participants make better or worse decisions, a participant said ***AI could make them doubt their initial decision if the AI contradicts their decision. However, it would be helpful if they were not sure and wanted a second opinion from the AI.*** Should the AI disagrees with the participant's opinions, they would go with their own diagnosis.

> "*It really depends on the case. If you see a clinical image and if you are very sure that it's [a diagnosis], I think the AI is going to make you doubt your initial thought if the AI contradicts your decision. But I think it would be helpful in case you are not sure and you would want a second opinion from an AI system.*" – inexperienced user

**User Concerns**

A participant expressed concern regarding the ***dataset*** used to train the AI, noting that the training data performs extremely well within a controlled environment. The real-world data can be variable and they would be keen to know how AI would behave outside of the training data. Moreover, HAM10000 data do not have a category between nevus and melanoma. They look similar but very separate. This is the current failure of the training dataset that needs to be addressed by differentiating this middle ground in the future.

Regarding the ***described features (concepts)***, features described are high-level, which can be difficult for beginners and junior doctors. A participant suggested it would be more helpful if there were more descriptions of the labels (e.g., what *reddish structures*, *pigmentations* mean). Users also suggested that the concepts (which are represented as segmentations) should focus on the lesion, and have the ability to ignore the irrelevant background (e.g., rulers, dark corners, etc.), which are confounding features. But in fact, this shows that explainable AI is helpful in terms of helping participants take into consideration if they should trust the AI or not.

> "*Sometimes the AI picked up something from the background rather than the lesion itself, so I think it would give a false recommendation.*" – inexperienced user

> "*When you see that the model is picking up things outside of the lesion of interest, then you know that the diagnosis is going to be skewed by that. [...] But I think that helps you as a clinician [...]. And you're thinking about whether should I change my diagnosis to be more in line with the model or should I stick with my own.*" – experienced user

Regarding the ***test images***, some of them are straightforward but some others are more challenging. This shows the limitation of the current study as doctors would need more information such as patient history, clinical location, age, sex, other images, etc. before they can make a diagnosis rather than relying on a single dermoscopic image.

> "*The pictures you selected [...] Some of them were more quite straightforward but some of them were more challenging, particularly the pigmented lesions were often*

*at the borderline. The recommendation-driven struggles with those.*" – experienced user

**Perceived Evidence Quality**

Overall, the provided evidence is quite good. Participants used both the concepts on the images and the weight of evidence of each concept. They often check the segmentations that represent features first to see if they can rely on them (how much they agree with them). Then, they use either the weight of evidence provided to make their diagnosis or go with their own knowledge of the dermoscopy. If they think the evidence were reliable, they would calibrate their decision to be closer to the decision aid.

"*[..] realise that the areas that it was highlighting weren't necessarily the area of interest. So I sort of go more with my own understanding of dermoscopy and think what would I say if I didn't have the aid.*" – inexperienced user

"*If it pulled out a good region of interest and it has what I would assume is the correct amount of weight put into it. Then that would sway more than if you're looking at something that you don't think is relevant.*" – experienced user

Sometimes, the segmentation on the test lesion and the training lesions are not consistent in finding the same area on the skin. For instance, in one example, the segmentation on the test lesion shows the whole lesion, including some external skin. But in the example training lesions, it just points at a specific area within the larger lesion. In another case, the evidence overlaps in multiple features - highlighting the same area but representing different features.

**Use of Additional Self-Sourced Evidence in Decision-Making**

Some participants tried to look at the original lesion first before checking the recommendation and evidence provided by the AI. Participants with experience in the skin cancer field would apply their knowledge to validate whether they should trust the decision aid's provided information. Furthermore, they also found other evidence that the model

failed to detect in the explanation. However, an experienced participant said that they did not use external evidence often in the hypothesis model. The self-sourced evidence only influences probably 15% of their decision-making, and they mainly rely on the evidence being provided, especially weights of evidence. Overall, the two interfaces have the advantage of pointing out important aspects of the dermoscopy that the doctors might have missed.

**Pros and Cons of the Recommendation-driven Interface**

The recommendation-driven has the main advantage of ***requiring a shorter time*** to do the diagnosis task and simpler to follow the information on the interface. Furthermore, the recommendation-based approach provides the most likely diagnosis, which is assumedly closer to the ground truth by inexperienced participants. Therefore, beginners in the field prefer the recommendation-driven because it is ***more streamlined*** and they only need to confirm if they want to follow the recommendation or not.

> "*If you look at lots of lesions, like hundreds of lesions. I think the benefit of the recommended is it sort of directs you [...] It's directing you to the most likely and showing you the evidence. So be a lot quicker.*" – experienced user

However, a major disadvantage of the recommendation-driven is that it ***can bias the clinicians***. Experienced participants suggested that users should be required to make their own decisions before looking at the AI recommendation. Furthermore, when the recommendation contradicts the user's own diagnosis, despite being very confident with their initial diagnosis, users can doubt their opinions and eventually a wrong AI recommendation can lead to a wrong direction.

> "*Having a decision-aid's recommendation would be a disadvantage because it may influence your own clinical assessment in a negative way.*" – inexperienced user

> "*I'm a huge fan of the recommendation one because as soon as I open the slide, it's just got a diagnosis there. So it's sort of pre-primed you to think what it is.*" – experienced user

Some experienced participants actually found that the recommendation-based approach easier to make the decision compared to the hypothesis-driven approach because they only have to compare the AI recommendation and their own diagnosis, which requires processing less information and ***clearer than the hypothesis-driven modality***.

> "*I assume it's pulling from a probability-weighted diagnosis [...], so it gives you the most amount of contrast between yourself and the computer, which I like.*" – experienced user

> "*I think it [the recommendation-driven] was clearer than the hypothesis-driven modality in that I found the compare and contrast a bit more finicky to use, whereas the recommendation put forward what the AI thought was the best but you could still use your own decision making when it came to.*" – experienced user

**Pros and Cons of the Hypothesis-driven Interface**

Some participants felt more confident in using the hypothesis-driven because it has ***more options*** for the users to choose from and they can compare and contrast between different hypotheses. It lets the users check where the AI is looking for each diagnosis. This design is especially important when users disagree with the model's recommendation and want to see the alternative options with their corresponding evidence. However, it ***requires users to have a good base of knowledge*** for all seven diagnoses to make comparisons between different hypotheses.

> "*I think the hypothesis-driven interface works better because you can compare between different diagnoses and how the AI looked for in both diagnoses.*" – inexperienced user

> "*I think the hypothesis-driven was a lot more helpful because a lot of that stems from where the recommended viewing parts of the lesion are.*" – experienced user

When using this interface, participants spent a substantial amount of time evaluating the lesion before checking the evidence. This can ***reduce the bias*** of relying on the AI model.

> *"The advantage is that you're still relying on your own initial clinical knowledge first because you're selecting what you think the possible lesions are."* – experienced user

However, ***having too much information*** can be a disadvantage, especially for participants who have a few years of experience in this field. When there is a lot of information, participants can ***feel uncertain*** which diagnosis they should choose, they eventually go with their own clinical assessment to make the final decision.

> *"If you have too much information displayed at once, then it becomes hard to pick and choose. But maybe that's for the best if our initial diagnosis is showing as not great evidence, [...] maybe I should reevaluate my own choice."* – experienced user

> *"It's depending on the experience of the user, whether they are resident or consultant, it's very easy to just click through all of the links and experience decision fatigue as a result"* – experienced user

Regarding which model is easier to use, it really depends on the setting. If we had individuals who do not have much experience in diagnosis, they would opt for the recommendation model. But for participants who have more experience, would prefer the hypothesis-driven. An experienced participant in the field commented that if they used the aid in a clinical setting, they would be more likely to use the hypothesis model because it allows them to put forward their hypothesis first and avoid bias.

#### Suggested Improvements for the Decision Support

Participants have suggested some improvements for the decision aid to help them make better decisions. The first suggestion is about ***improving feature descriptions***. An inexperienced user said adding more detailed descriptions for the features' labels would be useful for them. For example, adding explanations to describe what *reddish structures* mean, etc. and users can mouse over the feature's label to see more details.

Secondly, users suggested to add ***more supporting information*** in the aid. Hypothesis-driven could be improved by providing the rank of the hypotheses according to the AI model, such as clearly ordering the hypotheses based on the most likely to the least likely.

Or specifically providing a probability distribution of all possible diagnoses can be very helpful. However, we should do that after the hypotheses have been selected to avoid biasing the initial decision of the user. Furthermore, when giving a recommendation, we should also supply the probability that the model has given to that diagnosis. This can help users to be aware of the certainty of the AI model. Overall, adding a certainty level for each diagnosis for both interfaces is an important improvement.

> *"I didn't like that I had to click through all of them to see all of the evidence for each one. [...] someone maybe with less experience would have to click through all of them and that might be more time-consuming"* – inexperienced user

Another suggestion is that the decision aid needs to assess the ***chance of malignancy***. So adding information about malignant versus benign, or showing the presence or absence of features that would suggest malignancy is an important thing. Moreover, adding information such as having more images of the surrounding area and the ***history of the lesion*** can be very helpful. Another idea is to provide ***case-based explanations*** for each diagnosis. For example, if the model thinks a lesion is a melanoma, then the model should provide other lesions from other cases that look similar to the current case, but with the same diagnosis of melanoma. Regarding the decision-making workflow, the recommendation-driven could be improved further by implementing the ***human-first workflow***.

> *"I like the idea that you look at it and you make up your own mind and it tells you the recommendation afterwards. Whereas if it tells you what it thinks it is before you even look at it, I think there's cognitive guidance happening there."* – experienced user

## 5.6 Discussions

In this section, we summarise the findings from the experiments and discuss the implications of the results. We also discuss the validity and limitations of the study and suggest future work.

### 5.6.1 The Two Sides of the Coin: Recommendation-driven and Hypothesis-driven

| | Recommendation-driven | Hypothesis-driven |
|---|---|---|
| Pros | • *Shorter time* to do the task.<br>• *More streamlined for beginners.* | • Can *compare and contrast* between different hypotheses.<br>• *Reduce the bias* of relying on the AI recommendation. |
| Cons | • *Wrong AI* can lead to a wrong direction. | • Users *need to have a base of knowledge*.<br>• *Wrong evidence* can lead to a wrong direction. |

Table 5.12: Summary of the two interfaces.

Table 5.12 summarises the pros and cons of the two interfaces based on the qualitative results. Recommendation-driven is preferred by beginners who do not have much experience in the field. It is easier to use and faster to make a decision because they only need to confirm if they want to follow the recommendation or not. However, it can bias the clinicians and make them doubt their own diagnosis.

Hypothesis-driven is more suitable for experienced participants who have knowledge in the field. It allows users to compare and contrast between different hypotheses and check where the AI is looking for each diagnosis. However, it requires users to have a good base of knowledge for all possible hypotheses to be able to evaluate the provided evidence. It can also be overwhelming for users who do not have much experience in the domain. In addition, the evidence provided by the AI can be wrong, which can lead to a wrong decision.

Moreover, the order of AI in relation to the human decision-maker is also important. We can either put the AI first or the human first, both of which have their own advantages and disadvantages [89]. The AI-first approach can be used as a triage tool to reduce the assessment time, but it can face regulatory challenges and over-reliance on AI. The human-first approach can retain the current clinical workflow and use AI as a second opinion, which can increase the sensitivity of the diagnosis. However, it can be time-

consuming, with a potential increase in consultation time when there are disagreements between the AI and the human. In our study, recommendation-driven implemented the AI first, while hypothesis-driven could be improved by putting human opinions first before providing evidence from the AI.

The question now is *how should we design the decision-making interface in practice?* The answer is that we can combine the advantages of both interfaces. We can start by allowing the user to put forward their hypothesis and evidence first to avoid automation bias [33]. Then, we can provide the evidence found by the AI model to help users make a comparison. Furthermore, they will be shown a ranking of hypotheses based on the level of uncertainty. This allows users to be aware of the AI's recommendation (i.e., the hypothesis with the least uncertainty), as well as the supporting and opposing evidence for all possible hypotheses. Alternatively, we can use conformal prediction [200] to present multiple hypotheses within a given confidence bound. Uncertainty information can also reduce the number of possible hypotheses when there are too many to choose from, helping users focus on the most likely ones and avoid decision fatigue.

### 5.6.2 How Should the Evidence Be Presented?

The evidence should be detailed enough for users to understand the AI's decision-making process. Each feature (or concept) should be clearly and precisely shown where the AI is looking at the image. We also need to give the weight of evidence for each feature. It is not only about how much a feature contributes to the hypothesis, but also whether the feature is important, or relevant to the hypothesis. Particularly, even if we have strong evidence for a hypothesis, but the evidence is not relevant to it, its weight should be calibrated or even ignored when making the final decision.

### 5.6.3 Threats to Validity

We will identify threats to the validity of the human study, including both internal and external validity.

### Threats to Internal Validity

A threat to internal validity for a long human experiment (i.e., required approximately one hour to finish) is the **maturation**. Participants could become tired over time and lose concentration when doing later tasks. **Instrumentation** is the next threat that needs to be considered. We use labels from the HAM10000 dataset [204] to measure the performance of participants. But it is important to note that there is no single *ground-truth*. Different experts in the field can still have different opinions on the labels. Therefore, we will need a better approach to measuring the correctness of diagnoses rather than solely relying on these labels. Moreover, our pool of participants is relatively small (14 participants), the number of tasks is limited (16 tasks), trained on a single dataset (HAM10000), and limited number of conditions as we did not consider *no AI* condition in the human experiment. All these factors can affect the significance of the results.

### Threats to External Validity

The first threat to external validity is **sample characteristics**. We focus on using our network to invite experienced diagnosers to participate. This may not generalise to a broader population. Secondly, regarding **ecological validity**, laboratory experiments very often do not reflect the real world. For example, although some images can be straightforward, in most cases, doctors would need more information such as the history of the patient, age, regional images, etc. before they could make the diagnosis. So having a single dermatoscopic image can be challenging for participants and cause high uncertainty in the diagnosis. Moreover, a concern has been raised that AI can perform extremely well in a controlled environment, but it can behave very differently in clinical settings. That is why none of our experienced participants have used AI to support skin cancer diagnosis in a clinical setting. Overall, at this stage, there is currently no formal accredited AI system available, and we use AI only for research and internal testing purposes.

### 5.6.4 Future Work

The presentation of evidence can be improved further. We can achieve this by considering different sampling strategies to find image instances to describe the concept. We seek a sampling strategy that makes the concept easier to understand for users.

Moreover, we can design a condition that combines the two interfaces (recommendation-driven and hypothesis-driven) by providing both the AI recommendation and allowing users to choose their own hypotheses. But importantly, users should be required to put forward their own thoughts before looking at the AI recommendation. The decision-aid can also allow *argumentation* between the decision-maker and the AI by identifying differences between the user's hypothesis and the AI's recommendation, and the user's evidence and the AI's evidence.

It is also important to note that our participants are not experts. They have background knowledge in the field of skin cancer but are still far from being experts. In our current user study, we categorised them into *experienced* (those who have received clinical training) and *inexperienced* (those who have not received clinical training) participants. Future work should consider recruiting experts in the field, which could provide further insights into how domain expertise might influence reliance on AI.

## 5.7 Conclusion

In this chapter, we introduce **Visual Evaluative AI** [6], a tool for hypothesis-driven decision support for image data. This tool can highlight the high-level concepts in an image and provide positive and negative evidence for all possible hypotheses. Our tool is further applied and evaluated in the skin cancer domain with a web-based application called ***EvaSKan*** that offers skin cancer diagnosis support. By conducting a human study and interviewing participants experienced in the skin cancer field, we compare the hypothesis-driven approach with the recommendation-driven approach. We found that these two approaches have their own pros and cons, but can be combined to provide a better decision-support tool in the future.

[6] https://github.com/thaole25/EvaluativeAI

# Chapter 6
# Conclusion

In the final chapter, I will discuss the overarching motivations, and revisit the research questions and their contributions. I will then identify the limitations of the research and suggest future work. Finally, I will conclude the thesis with the summary remarks.

## 6.1 Research Contributions

In this section, I address the research questions proposed in Chapter 1. Table 6.1 summarises the contributions produced by addressing these research questions.

### 6.1.1 Explaining the Uncertainty

Presenting uncertainty has been applied as a way to promote users' trust and understanding when interacting with AI models [222, 239]. However, users might want to know why the model is uncertain, which can be helpful in deciding if they should trust this uncertainty measure or not. To the best of my knowledge, there is still limited research on explaining the uncertainty. Therefore, **RQ1** seeks to address this challenge.

Explaining the uncertainty of the AI prediction is a promising research direction and may lead to promoting appropriate trust in AI models [196, 202]. Seuß [196] suggest some possible ways to explain the uncertainty that is appropriate for humans, such as *concrete explanation* and *counterfactual explanation*. In concrete explanation, the source of the uncertainty is explicitly indicated (e.g. "The prediction is salary $> 50,000$ with 15% of confidence because the person has a blue-collar job"). In counterfactual explanation, it shows the minimal changes in the input to have a different output (e.g. "The prediction

| Research Question | Contribution |
|---|---|
| **RQ1 (Chap 3):** How can we explain model uncertainty? | • Formalising counterfactual explanation of confidence scores. |
| **RQ2 (Chap 3):** Can explaining model uncertainty improve user trust and understanding in the machine learning model? | • Conducting two user studies to investigate whether explanations of model uncertainty can help users better understand and trust the model;<br>• Identifying limitations of example-based explanations and visualisation-based explanations using qualitative analysis. |
| **RQ3 (Chap 4):** How can we design an effective evidence-based decision-support model? | • Proposing evidence-informed hypothesis-driven decision-making model based on the Evaluative AI framework [153] and the Weight of Evidence (WoE) framework [148];<br>• Conducting two user studies to evaluate whether the hypothesis-driven approach can improve decision quality and reduce over-reliance on the AI model;<br>• Identifying limitations and challenges of the three decision-support approaches, namely (1) AI-recommendation, (2) AI-explanation-only and (3) hypothesis-driven approach. |
| **RQ4 (Chap 5):** Based on the new decision-support paradigm, how can we build a decision-aid tool for image datasets? | • Extending the Weight of Evidence (WoE) framework to apply to image datasets;<br>• Building a decision-aid library that offers hypothesis-driven decision-support by providing evidence for and against a given hypothesis. |
| **RQ5 (Chap 5):** How do different decision-support approaches impact human decision-making in skin cancer diagnosis? | • Applying the aforementioned decision-support library to a case study in skin cancer diagnosis;<br>• Conducting a user study with experienced people in dermatology to evaluate the impact of the decision-support approaches (recommendation-driven and hypothesis-driven) on decision quality and user satisfaction. |

Table 6.1: Research Questions and Contributions

| | **Original CF Model [189]** | **Proposed CF Model** |
|---|---|---|
| Goal | Search for CF inputs of another class. | Search for CF inputs of the same class but with a different confidence score. |
| Question | Why does the model predict this employee will **leave** instead of will stay in this company? | The model predicts that this employee will leave. Why is the model **70% confident** instead of 40% confident or less? |
| Answer | You could have got a prediction of stay instead if Age had taken the value of 45 rather than **25**. | You could have got a confidence score of 40% instead if Daily Rate had taken the value 400 rather than **300**. |

Table 6.2: Differences between the original CF model and the proposed CF model. **Bold text** indicates the factual input/class, underline text indicates the CF input/class.

is salary $> 50,000$ with 15% of confidence. If the person had a white-collar job, the confidence would be 80%"). My work is in line with the counterfactual explanation approach. There are also some other approaches to solving counterfactuals for tabular [101, 159], image [38, 51, 71], text [88, 186] and time series data [50]. However, none of these is for explaining model confidence. Furthermore, van der Waa et al. [211] propose a framework called *Interpretable Confidence Measures (ICM)* which provides predictable and explainable confidence measures based on case-based reasoning [11]. This approach did not address counterfactual explanations of model confidence.

Some recent works have proposed to explain the uncertainty. For example, Antoran et al. [7] propose Counterfactual Latent Uncertainty Explanations (CLUE), to learn which input features are responsible for the model's uncertainty. Their model for finding counterfactual examples is similar to mine. However, I go further by running more comprehensive user studies to measure the impact of the explanations on users' trust, understanding and satisfaction. This corresponds to **RQ2**. I also consider different ways to present the counterfactual explanations and compare them in terms of their strengths and limitations.

Chapter 3 demonstrates the contributions stem from addressing **RQ1** and **RQ2**. First, I formalise the counterfactual explanation of confidence scores followed by the counter-

factual model proposed by Russell [189]. Specifically, the proposed CF model shows how the confidence score would change if the input features were different, but the output prediction class remains the same. Table 6.2 shows the differences between the original CF model and my proposed CF model.

Further, I present counterfactual explanations of confidence scores in two different ways: (1) example-based explanations and (2) visualisation-based explanations. The former shows different examples with different confidence scores by modifying an input feature and presenting them in a table. The latter visualises how changing a feature affects the confidence score. To address **RQ2**, I then conduct two user studies to evaluate the effectiveness of these explanations in improving users' trust, understanding and satisfaction. The results indicate that having explanations of model uncertainty can improve users' trust and understanding, compared to a baseline of no explanations. However, there is no significant difference between the two types of explanations in terms of user trust, understanding and satisfaction.

Moreover, this work identifies the strengths and limitations of these two types of explanations. More specifically, using visualisation-based explanations makes it easier to understand the correlations between input features and confidence scores. Regarding example-based explanations, study participants often apply *case-based reasoning* by finding the closest example in the counterfactual explanations to the test example, rather than interpreting the linear correlation between the input feature and the confidence scores. The findings suggest that we should use example-based explanations to present case-based variables, and visualisation-based explanations to present continuous variables.

### 6.1.2 Designing the Evidence-based Decision-Support Model

A traditional decision-support approach is to provide recommendations to users based on the AI model's predictions. However, this approach only allows users to either accept or reject the recommendation, without providing any other alternatives. This can lead to *under-reliance* and *over-reliance* on the AI model [215]. To address this challenge, **RQ3** aims to design a new decision-support approach that lets users make informed decisions by allowing them to make their own decisions based on the evidence provided by the AI

model, built on the *Evaluative AI* framework [153]. Figure 6.1 describes the difference between the traditional recommendation-driven approach and the new hypothesis-driven approach.

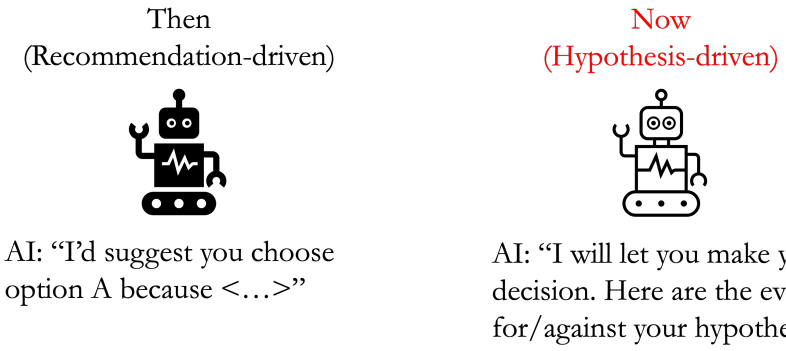



Figure 6.1: Comparison between the recommendation-driven approach and the hypothesis-driven approach

In Chapter 4, I propose an evidence-informed hypothesis-driven decision-making model based on the Evaluative AI framework [153] and the Weight of Evidence (WoE) framework [148]. This model allows users to select a hypothesis and see the corresponding positive and negative evidence for that hypothesis. I further conduct user studies to evaluate the effectiveness of the (1) hypothesis-driven approach when comparing with two baselines: (2) AI recommendation and (3) AI-explanation-only. Using qualitative data analysis, I explore how people make decisions differently when using these three approaches. The findings suggest that in the recommendation-driven, users tend to use input feature values to decide if they should trust the recommendation and its explanation or not. In the AI-explanation-only, they mostly rely on the input feature values and ignore the explanation. The reason is that the AI prediction (recommendation) is hidden in this approach. Therefore, as they can only see the explanation without knowing which recommendation it is referring to, it is difficult for them to understand the explanation. In the hypothesis-driven approach, users can see the evidence for and against all hypotheses.

Moreover, I identify the strengths and limitations of the hypothesis-driven approach. Regarding the strengths, the hypothesis-driven approach allows similar completion time, improves decision quality and reduces over-reliance on the AI model, compared to the

AI-recommendation approach. Hypothesis-driven can also reduce under-reliance when compared to the AI-explanation-only approach. It is important to note that when comparing with the recommendation-driven, even though the hypothesis-driven can increase under-reliance, it is more important to reduce over-reliance. The reason is that over-reliance can lead to *automation bias* [215], which happens when people have unwarranted trust in the AI [87]. This can result in harmful outcomes, especially in high-stake domains like healthcare. Under-reliance, while reducing AI benefits, still allows human decision-makers to remain in control and cautious, mitigating the risk of automation bias. Further, the hypothesis-driven approach allows users to be aware of the uncertainty among different hypotheses, which is not addressed in the other two baseline approaches. Users can then discount or completely ignore the weights of evidence depending on whether the feature is important or not among hypotheses with high uncertainty.

### 6.1.3 Visual Evaluative AI - A Case Study in Skin Cancer Diagnosis

The Weight of Evidence (WoE) framework [148] in Chapter 4 had been implemented only for tabular data, but now I introduce a method for dealing with image data. Extracting features from images is more complex than processing pre-defined features in tabular data. This requires deep learning techniques such as convolutional neural networks (CNNs). Therefore, referring to **RQ4** and in Chapter 5, I extend the WoE framework by applying concept-based explanations [234, 238] to extract *high-level concepts*. These concepts represent the features of the image. These concepts (e.g., beak, leg, wing, etc.) are then put into the WoE framework to calculate the evidence for and against a given hypothesis (e.g., pelican, gull). By combining the WoE framework with the concept-based explanations, I build and publish a decision-aid library called *Visual Evaluative AI (VisE)*. This library allows users to check the positive and negative evidence for all possible hypotheses of the image.

More specifically, I apply two concept-based explanation methods to generate evidence: (1) unsupervised (e.g., Invertible Concept-based Explanations (ICE) [238]) and (2) supervised (e.g., Post-hoc Concept Bottleneck Model (PCBM) [234]). When combined with the Weight of Evidence (WoE) framework, I have two different versions of the Visual

Evaluative AI library: (1) ICE+WoE and (2) PCBM+WoE. The experiment's findings indicate that both ICE+WoE and PCBM+WoE with as few as 7 concepts can achieve similar performance to the original CNN models with 2048 features in the final CNN layer.

To understand the impact of different decision-support approaches, I apply the Visual Evaluative AI library to a case study in supporting skin cancer diagnosis. The web application built on this is called *Evaluative Skin Cancer (EvaSkan)*.

This refers to **RQ5** in Chapter 5. Particularly, I conduct a user study with people who have a background in dermatology to evaluate the impact of the two decision-support approaches (recommendation-driven and hypothesis-driven) on diagnosis quality and user experience. The results show that experienced users prefer the hypothesis-driven, while inexperienced users prefer the recommendation-driven. Both approaches have their pros and cons. The recommendation-driven can be easier to use, but the hypothesis-driven is more informative. Through subjective reports, the hypothesis-driven approach is preferred in terms of allowing users to have more control over the decision-making process, more helpful in making the diagnosis and easier to spot errors in the information provided by the decision-aid.

### 6.1.4 Experimental Domains

| | Domain | Dataset Type | Source of Subjects | Number of Subjects |
|---|---|---|---|---|
| Chap 3 (Explaining the Uncertainty) | Income prediction [143, 217, 228] & HR (resignation prediction) [99, 199] | Tabular data | Crowdsourcing (Amazon MTurk) | 180 (90 in each domain) |
| Chap 4 (Evidence-Based Decision-Support Model) | Housing price prediction [43, 179] | Tabular data | Crowdsourcing (Prolific) | 397 (302 in Experiment 1, 95 in Experiment 2) |
| Chap 5 (Visual Evaluative AI) | Skin cancer diagnosis [17, 37, 205] | Image data | Professional network | 14 |

Table 6.3: Summary of human experiments

Table 6.3 provides a summary of domains used in human experiments. In Chapter 3 and 4, I use domains (income, resignation, housing price predictions) that are familiar to laypeople. Research has shown that participants provide high-quality answers [54] on Prolific and much more attentional engagement [2] than on Amazon MTurk. Therefore, I moved to Prolific for the second experiment in Chapter 4.

In Chapter 5, I use a domain that requires domain knowledge (dermatology) to incorporate human prior knowledge into the decision-making process. I recruited participants from my professional network to ensure that they have the required background. This can lead to selection bias and the selected pool of participants may not be representative of all skin cancer experts. The pool of participants is also small, which can affect the significance of the results.

## 6.2 Limitations

This section discusses the limitations of the proposed explainable models (counterfactual models, evidence-informed hypothesis-driven decision-making model and Visual Evaluative AI library) and the human studies conducted.

### 6.2.1 Proposed Explainable Models

In Chapter 3, the main drawback is that this is one form of counterfactual explanation that was built on the counterfactual model proposed by Russell [189]. A more comprehensive evaluation of different counterfactual approaches is needed, especially to address the five deficits of counterfactual explanations in [102]. In Chapter 4 and 5, a limitation lies in providing trustworthy evidence. Since the evidence is generated by the AI model, it is important to note that the evidence is not always reliable. Challenges remain in presenting the evidence in a way that is easy to understand and trustworthy to users. In Chapter 5, the concept-based explanations are not perfect and can be noisy, which are still required to be validated by domain experts.

### 6.2.2 Human Studies

First, regarding the ***internal validity***, there are *instrumentation* threats. In all user studies, I use publicly available datasets such as the income dataset from UCI Machine Learning Repository [55] and the IBM HR Analytics Employee Attrition Performance dataset published in Kaggle [172] in Chapter 3; the Ames Housing [48] in Chapter 4; and the HAM10000 dataset [204]. As in the dataset, we have labels to evaluate the performance in experimental tasks. However, these labels are not the *ground truth*. Different people, including domain experts, can label the same data differently. Therefore, the experimental tasks can be subjective.

When considering the *external validity*, first, the scale regarding the number of tasks, the number of participants and the number of datasets used is limited. Particularly, all human studies have from ten to sixteen tasks. In the skin cancer diagnosis experiment (Chapter 5), the number of participants is fourteen. I also only use either one or two datasets in each study. This can limit the generalisability of the findings. Second, I recruited participants in Chapter 5 from my network to get people who have backgrounds in dermatology. This can lead to both *selection bias* and the selected pool of participants may not be representative of the general population. Third, the studies are artificial and conducted in a controlled and laboratory environment so it may not reflect the real-world scenarios.

## 6.3 Future Work

In this section, I propose some future work to address the limitations.

### 6.3.1 Combining Recommendation-driven and Hypothesis-driven

In Chapter 4, I compare the recommendation-driven and hypothesis-driven approaches. These two approaches both have pros and cons. Therefore, I propose to combine these two approaches to leverage the strengths of each. In this combined approach, users can first put forward their initial thoughts without being provided any aid from the AI. Fol-

lowing this, they will be provided with the rank of hypotheses based on the level of uncertainty. Based on this, users can be aware of the AI recommendation (i.e., the hypothesis with the least uncertainty), as well as the evidence for and against all possible hypotheses. This can help users to make better decisions by considering both the AI recommendation and the evidence. Further, since they can provide their initial thoughts, this can help to reduce the bias on the AI model [33].

Moreover, the decision-aid can allow *argumentation* between the decision-maker and the AI. In this approach, users can provide their thoughts, including the *evidence* for their decisions. Then, the decision-support approach will provide comparisons between the user's evidence and the AI's evidence, and between the user's hypothesis and the AI's hypothesis. Moreover, users can adjust the AI's evidence based on their domain knowledge. This can be helpful for users when they can compare their decisions with the AI model's decisions and contribute to improving the AI model when they believe the model is wrong. This approach is based on the theory of sensemaking ***Data/Frame Theory*** [112], where the process is iteratively done by combining System 1 (fast, intuitive) and System 2 (slow, rational) thinking [97]. People make their decisions based on their intuition and then adjust their decisions when carefully examining the evidence being provided by the AI. In this case, people can construct a *frame* based on the *data* and can question the frame to construct a new one when new information arises.

Some recent works have used argumentation-based approaches to better design AI decision support. For example, [45] introduced *devil's avocate* to challenge the AI recommendation or the majority opinion within a group. Alternatively, the devil's advocate can be used to present counter-arguments against the user's decision [144]. Moreover, [145] proposed the idea of *Deliberative AI*, which allows the human user and the AI to deliberate conflicting evidence and arguments.

### 6.3.2 Human Experiment Design

In Chapter 4 and 5, we should control the cognitive load across experimental conditions by including a cognitive load measurement such as NASA-TLX scales [78] or reaction-time analysis. Since the hypothesis-driven condition can require more cognitive effort

than the traditional recommendation-driven condition, we should seek to identify if users perform worse with the hypothesis-driven approach, and if so, whether it is due to cognitive load or the approach itself. Future work can consider using these cognitive load measures when evaluating different decision-support approaches. Another aspect to improve the human experiments is to assess whether trust calibration and decision quality improve *over time*. Current experiments only compare the decision-support approaches in a short-term setting, without considering the long-term effects of repeated exposure.

### 6.3.3 Generalisability

To improve the generalisability of the findings, I propose to conduct more user studies with different datasets and domains. For example, in Chapter 5, I only use the HAM10000 dataset for the skin cancer domain. The Visual Evaluative AI tool can be applied to other image datasets as well, which can help to strengthen the findings further and explore how different domains can impact human decision-making. Moreover, I propose to conduct user studies with different groups of people. In Chapter 5, I recruited participants using my professional network, which is a threat to external validity. Since the number of participants in this study is currently 14, the pool of participants should be larger to ensure the generalisability of the results.

## 6.4 Summary Remarks

This thesis has addressed the challenges of improving explainable decision-support models. Overall, I have proposed a counterfactual explanation model of confidence scores, an evidence-informed hypothesis-driven decision-making model and a Visual Evaluative AI library. The results indicate that the evidence-informed models have the potential to reduce over-reliance on the AI model and help users make better decisions. However, challenges remain in providing better and more trustworthy evidence and reducing both over-reliance and under-reliance on the AI model. Therefore, this thesis hopes to inspire future research in addressing these challenges and developing more effective and reliable decision-support approaches.

# Appendix A
# Explaining the Uncertainty

## A.1 Human Experiment

In this section, I will provide example questions used in the human experiment. There are three conditions in the experiment: (1) Control, (2) Treatment (Example-Based) and (3) Treatment (Visualisation-Based). Details of the experiment structure are shown in Table 3.2 in Chapter 3.

### A.1.1 Phase 2: Task Prediction

These are the example questions designed based on the income dataset [55]. Figure A.1, A.2, A.3, A.4, A.5 and A.6 show the training and question phases in each condition.

### A.1.2 Phase 3: 10-point explanation satisfaction rating scale

In phase 3, there are 8 rating scale questions to evaluate users' satisfaction as follows (1 = Disagree strongly; 10 = Agree strongly):

1. From the explanation, I *understand* how the confidence score changes.
2. This explanation of how the confidence score changes is *satisfying*.
3. This explanation of how the confidence score changes has *sufficient detail*.
4. This explanation of how the confidence score changes seems *complete*.
5. This explanation of how the confidence score changes *tells me how to use* it.

6. This explanation of how the confidence score changes is *useful to my goals*.
7. This explanation of the confidence score shows me how *accurate* the AI prediction is.
8. This explanation lets me judge when I should *trust and not trust* the AI algorithm.

### A.1.3 Phase 4: 10-point trust rating scale

In Phase 4, there are 8 rating scale questions to evaluate users' trust as follows (1 = Disagree strongly; 10 = Agree strongly):

1. I am *confident* in the AI model. I feel that it works well.
2. The outputs (prediction and confidence score) of the AI model are very *predictable*.
3. The AI model is very *reliable*. I can count on it to be correct all the time.
4. I feel *safe* that when I rely on the AI model, I will get the right answers.
5. The AI model is *efficient* in that it works very quickly.
6. I am *wary* of the AI model.
7. The AI model can *perform the task better* than a novice human user.
8. I like using the AI model for *decision-making*.

**Training: Understanding the Task**

In the following task, you will see the attributes (information) of an anonymous employee, such as marital status, education, occupation, age, etc. The Artificial Intelligence model (AI model) uses these attribute values and gives a prediction that this employee either has an income *equal to or greater than $50,000* or *less than $50,000*.

The AI model also provides a **confidence score** of the income prediction, which defines how confident the AI model is in its own prediction. The **confidence score** is ranged from 0 to 100. The higher the score, the more confident the model prediction is.

You will be given information about some employees, such as marital status, education, occupation, age, etc. You will also be given the output of the AI model prediction (>= $50,000 or < $50,000) but the confidence score will not be shown. **Your task is to decide for which employee the AI model will predict a higher confidence score.**

***You will be scored based on your answer. A correct answer will give you 1 point, a wrong answer will reduce 2 points. If you select "I don't have enough information to decide", you will receive 0 points for that question.***

***The final compensation will be calculated based on your final score: a score of 0 (or less than 0) will receive $7 USD and for each additional score, you will receive a bonus of $0.2 USD***

*After you finish the task prediction, we will ask you to evaluate your trust and satisfaction using sliders (track bars).* ***You will not be scored when evaluating your trust and satisfaction.***

Figure A.1: (C1) Control condition: Training phase

| Attribute | Employee 1 | Employee 2 | Employee 3 |
|---|---|---|---|
| *Marital Status* | Married | Married | Married |
| *Number of years of education* | **12** | **15** | **14** |
| *Occupation* | Service | Service | Service |
| *Age* | 64 | 64 | 64 |
| *Any capital gain* | No | No | No |
| *Working hours per week* | 40 | 40 | 40 |
| *Education* | Bachelors | Bachelors | Bachelors |
| **AI model prediction** | **Lower than $50,000** | | |

*Q4.3.*

For which employee in the above table the AI model predicts with the **highest confidence score**?

- ○ Employee 1
- ○ Employee 2
- ○ Employee 3
- ○ I don't have enough information to decide

*Q4.4.*

Can you please explain why you selected this option? (Please write a brief sentence in the text box)

Figure A.2: (C1) Control condition: Question phase

**Training: Understanding the Task**

In the following task, you will see the attributes (information) of an anonymous employee such as marital status, education, occupation, age, etc. The Artificial Intelligence model (AI model) uses these attribute values and gives a prediction that this employee either has an income *equal to or greater than $50,000* or *less than $50,000*.

The AI model also provides a **confidence score** of the income prediction, which defines how confident the AI model is in its own prediction. The **confidence score** is ranged from 0 to 100. The higher the score, the more confident the model prediction is.

You will be given an **explanation table** that helps you to understand how the confidence score is calculated. In this explanation table, you will be presented with person's details, the prediction whether their income is less than or equal to $50K or whether it is greater than $50K, and the confidence score. You will also be presented with several alternative values for some of the person's details, such as a change in their marital status or education. These do not change the prediction of their income, but they do change the AI model's confidence. The confidence level for each alternative is also given in the table. ***You should look at the value changes of the person's details and see how that correlates with the confidence score.***

**Your task is to decide for which employee the AI model will predict a higher confidence score.**

Here is an example of the **explanation table**

| Attribute | Original Value | Alternative 1 | Alternative 2 | Alternative 3 | Alternative 4 |
|---|---|---|---|---|---|
| *Marital Status* | Never Married | - | **Divorced/Widowed** | **Married** | - |
| *Number of years of education* | 4 | **3** | **10** | - | - |
| *Occupation* | Job Service | - | - | - | **Blue-Collar** |
| *Age* | 21 | - | - | **35** | - |
| *Any capital gain* | No | - | **-** | **Yes** | - |
| *Working hours per week* | 48 | - | - | - | **37** |
| *Education* | No High School | - | **Graduate** | - | - |
| **Confidence score** | **99.1%** | **99.3%** | **91.6%** | **40.6%** | **99.5%** |
| **AI model prediction** | **Lower than $50,000** | | | | |

The above table shows the attributes of an employee with a confidence score of the AI model. The AI model predicts that the income of this employee is **lower than $50,000.** When we change the values of employee's attributes as in columns **Alternative 1, 2, 3 and 4**, the confidence score changes but the AI model still predicts that the income of this employee is **lower than $50,000**.

**Note that in Alternative Columns, notation (-) means the value is unchanged from the original value, we only highlight the values that changed.**

***You will be scored based on your answer. A correct answer will give you 1 point, a wrong answer will reduce 2 points. If you select "I don't have enough information to decide", you will receive 0 points for that question.***

***The final compensation will be calculated based on your final score: a score of 0 (or less than 0) will receive the standard base rate of $7 USD and for each additional score, you will receive a bonus of $0.2 USD***

*After you finish the task prediction, we will ask you to evaluate your trust and satisfaction using sliders (track bars).* ***You will not be scored when evaluating your trust and satisfaction.***

Figure A.3: (C2) Example-based condition: Training phase

*Q7.3.* **See the explanation table below**

| Attribute | Alternative 1 | Alternative 2 | Original Value | Alternative 3 | Alternative 4 |
|---|---|---|---|---|---|
| *Marital Status* | - | - | Married | - | - |
| *Number of years of education* | **11** | **10** | **9** | **8** | **7** |
| *Occupation* | - | - | Service | - | - |
| *Age* | - | - | 63 | - | - |
| *Any capital gain* | - | - | No | - | - |
| *Working hours per week* | - | - | 12 | - | - |
| *Education* | - | - | High School | - | - |
| **Confidence score** | **39.9%** | **49.4%** | **57.8%** | **65.2%** | **71.5%** |
| **AI model prediction** | **Lower than $50,000** | | | | |

In the following table, for which employee the AI model predicts with the **highest confidence score**?

| Attribute | Employee 1 | Employee 2 | Employee 3 |
|---|---|---|---|
| *Marital Status* | Married | Married | Married |
| *Number of years of education* | **12** | **15** | **14** |
| *Occupation* | Service | Service | Service |
| *Age* | 64 | 64 | 64 |
| *Any capital gain* | No | No | No |
| *Working hours per week* | 40 | 40 | 40 |
| *Education* | Bachelors | Bachelors | Bachelors |
| **AI model prediction** | **Lower than $50,000** | | |

- ○ Employee 1
- ○ Employee 2
- ○ Employee 3
- ○ I don't have enough information to decide

*Q7.4.*
Can you please explain why you selected this option? (Please write a brief sentence in the text box)

Figure A.4: (C2) Example-based condition: Question phase

**Training: Understanding the Task**

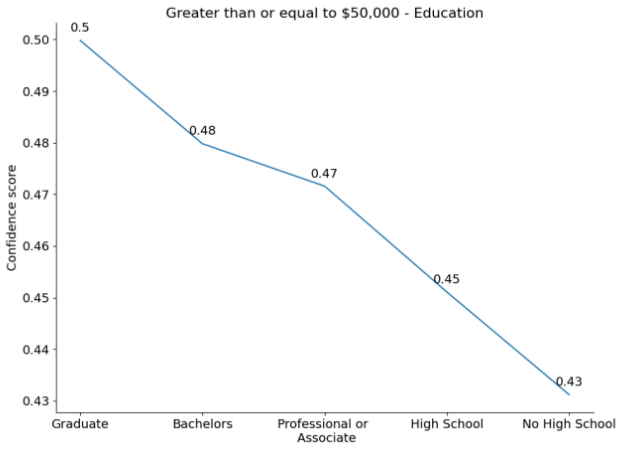
In the following task, you will see the attributes (information) of an anonymous employee such as marital status, education, occupation, age, etc. The Artificial Intelligence model (AI model) uses these attribute values and gives a prediction that this employee either has an income *equal to or greater than $50,000* or *less than $50,000*.

The AI model also provides a **confidence score** of the income prediction, which defines how confident the AI model is in its own prediction. The **confidence score** is ranged from 0 to 1. The higher the score, the more confident the model prediction is.

You will be given an **explanation graph** that helps you to understand how an employee's information (e.g. Education) changes can change the confidence score. The changes do not change the prediction of their income (*equal to or greater than $50,000* or *less than $50,000*), but they do change the AI model's confidence. ***You should look at the value changes of the person's details and see how that correlates with the confidence score.***

**Your task is to decide for which employee the AI model will predict a higher confidence score.**

Here is an example of the **explanation graph**




The above graph shows the changes in the Education of an employee with changes in the confidence score of the AI model. The AI model predicts that the income of this employee is **greater than or equal to $50,000.** When we change the values of the employee's education as in the horizontal line, the confidence score changes as presented in the blue line but the AI model still predicts that the income of this employee is **greater than or equal to $50,000**.

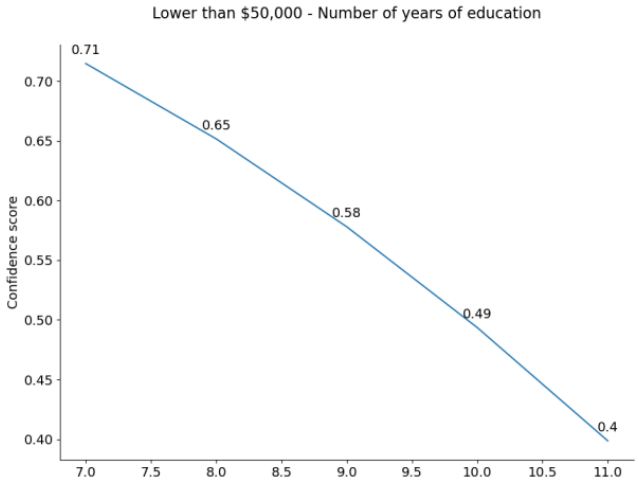
***You will be scored based on your answer. A correct answer will give you 1 point, a wrong answer will reduce 2 points. If you select "I don't have enough information to decide", you will receive 0 points for that question.***

***The final compensation will be calculated based on your final score: a score of 0 (or less than 0) will receive the standard base rate of $7 USD and for each additional score, you will receive a bonus of $0.2 USD***

*After you finish the task prediction, we will ask you to evaluate your trust and satisfaction using sliders (track bars).* ***You will not be scored when evaluating your trust and satisfaction.***

Figure A.5: (C3) Visualisation-based condition: Training phase

*Q10.3.* **See the explanation graph below**




In the following table, for which employee the AI model predicts with the **highest confidence score**?

| Attribute | Employee 1 | Employee 2 | Employee 3 |
|---|---|---|---|
| *Marital Status* | Married | Married | Married |
| *Number of years of education* | **12** | **15** | **14** |
| *Occupation* | Service | Service | Service |
| *Age* | 64 | 64 | 64 |
| *Any capital gain* | No | No | No |
| *Working hours per week* | 40 | 40 | 40 |
| *Education* | Bachelors | Bachelors | Bachelors |
| **AI model prediction** | **Lower than $50,000** | | |

- ○ Employee 1
- ○ Employee 2
- ○ Employee 3
- ○ I don't have enough information to decide

*Q10.4.*
Can you please explain why you selected this option? (Please write a brief sentence in the text box)

Figure A.6: (C3) Visualisation-based condition: Question phase

# Appendix B

# Hypothesis-Driven Decision Making Model

## B.1 Statistics of Experiment 1

In this section, I show the statistics of four measures in Experiment 1 as in Table B.1, B.2, B.3 and B.4. In these four tables, *R* means Recommendation-driven, *O* means AI-explanation-only and *H* means Hypothesis-driven. The statistics include the count, mean, standard deviation, minimum, 25% (first quartile), 50% (median), 75% (third quartile) and maximum of the measures.

| condition | count | mean | std | min | 25% | 50% (median) | 75% | max |
|---|---|---|---|---|---|---|---|---|
| (C1) R | 102.000 | 17.920 | 9.342 | 5.100 | 11.492 | 15.767 | 21.792 | 67.017 |
| (C2) O | 99.000 | 18.619 | 11.258 | 3.500 | 12.375 | 15.467 | 22.417 | 92.650 |
| (C3) H | 101.000 | 18.087 | 9.255 | 5.967 | 11.850 | 15.733 | 21.450 | 55.833 |

Table B.1: Statistics of **completion time** per condition (in minutes).

| condition | count | mean | std | min | 25% | 50% (median) | 75% | max |
|---|---|---|---|---|---|---|---|---|
| (C1) R | 102.000 | 0.290 | 0.071 | 0.183 | 0.239 | 0.277 | 0.333 | 0.484 |
| (C2) O | 99.000 | 0.295 | 0.073 | 0.140 | 0.242 | 0.281 | 0.337 | 0.594 |
| (C3) H | 101.000 | 0.267 | 0.063 | 0.154 | 0.228 | 0.252 | 0.304 | 0.474 |

Table B.2: Statistics of **Brier score** per condition.

| condition | count | mean | std | min | 25% | 50% (median) | 75% | max |
|---|---|---|---|---|---|---|---|---|
| (C1) R | 102.000 | 73.856 | 20.913 | 33.333 | 66.667 | 66.667 | 100.000 | 100.000 |
| (C2) O | 99.000 | 54.209 | 22.505 | 0.000 | 41.667 | 50.000 | 66.667 | 100.000 |
| (C3) H | 101.000 | 53.300 | 22.733 | 16.667 | 33.333 | 66.667 | 66.667 | 100.000 |

Table B.3: Statistics of **over-reliance** (%) per condition.

| condition | count | mean | std | min | 25% | 50% (median) | 75% | max |
|---|---|---|---|---|---|---|---|---|
| (C1) R | 102.000 | 17.810 | 20.346 | 0.000 | 0.000 | 16.667 | 33.333 | 100.000 |
| (C2) O | 99.000 | 41.246 | 27.183 | 0.000 | 16.667 | 33.333 | 50.000 | 100.000 |
| (C3) H | 101.000 | 24.422 | 18.191 | 0.000 | 16.667 | 16.667 | 33.333 | 100.000 |

Table B.4: Statistics of **under-reliance** (%) per condition.

## B.2 Human Experiment

In this section, I will show example questions in my human experiments. There are two main phases in each experiment:

- Training phase: Participants were given three example questions.
- Test phase: After finishing the training phase, participants were given twelve test questions.

First, Figure B.1a, B.2a and B.3a show the first page of the **training phase** in three conditions. This page introduces the task, how to read the evidence, what information

will be given and how the participants will be scored. After that, three example questions will be shown, one of which is Figure B.1b or B.2b or B.3b depending on the condition.

Second, after finishing the training phase, there are twelve questions in the **test phase**. Here are the following tasks in this phase.

- **Experiment 1**: Assign the likelihood for each price range (Low/Medium/High) of the given house
- **Experiment 2**: In addition to the same tasks as in experiment 1, participants were asked to explain their decision using free text. Participants in Experiment 1 were not allowed to participate again in Experiment 2.

Example questions in Experiment 1 are shown in Figure B.4, B.5 and B.6. Experiment 2 will use the same set of questions, with an addition of having a free text after the likelihood slider to ask participants why they made such choices.

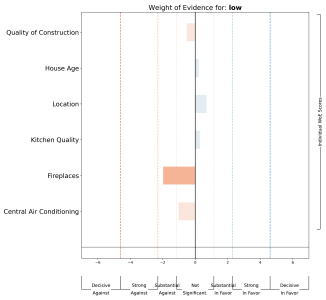
To compare between the three conditions ***(C1) Recommendation-driven***, ***(C2) AI-explanation-only*** and ***(C3) Hypothesis-driven***:

- In Figure B.4, participants can see the AI prediction (i.e. *low* price in this case) and the weight of evidence (the explanation) for that prediction.
- In Figure B.5, the AI prediction is hidden. Therefore, even though the participants can see the explanation, they do not know which class (low/medium/high) the evidence refers to.
- In Figure B.6, the house features selected are similar to the example question in Figure B.4 and B.5. I show participants the evidence for all hypotheses (*low*, *medium* and *high*). I do not give them the AI prediction. Specifically, I have supportive evidence in most features for hypothesis *low*. By contrast, strongly negative evidence refutes hypothesis *high*. The correct answer here is *low*.

In the following task prediction, you will see the features (evidence) of a house, including:

1. Quality of Construction: 6 out of 10 (10 is the best score)
2. House Age: 44 years
3. Location: 2 out of 4 (4 is the best score)
4. Kitchen Quality: Not good
5. Fireplaces: Available
6. Central Air Conditioning: Available

A decision aid uses these features (also called **evidence**) and gives a prediction whether this house will be in a price range of: **low, medium or high**.




The decision aid also provides an explanation, which is presented as **weight of evidence** for all features.
**How to read the evidence?** Looking at the above figure, we have the weight of evidence (WoE) for each feature. Positive weight of evidence (presented as blue colour) means the feature's value speaks in favour of the hypothesis that the value of the house is 'Low'. Negative weight of evidence (presented as red colour) means the feature's value speaks against the hypothesis that the value of the house is 'Low'. The weight of evidence is also measured as whether they are significant or not based on the horizontal axis.

There are three hypotheses for a house in this task: Low price, Medium price and High price.
For example, in this figure, House Age, Location and Kitchen Quality support the hypothesis 'Low'. Specifically, since the house age is 44 years old and the kitchen quality is 'not good', this implies that the house price should be low. However, their weight of evidence are not significant as shown on the horizontal axis. On the other hand, Quality of Construction, Fireplaces and Central Air Conditioning refute the hypothesis 'Low', which means they suggest this house might have a higher price.

**You will be given information about some houses. You will also be given the output of the decision aid's prediction (low or medium or high price) along with its evidence. Your task is to assign the likelihood for each price range of the given house.**
**You can make use of the evidence provided and use these information to make a final decision**

- **whether the evidence supports or refutes your prediction**
- **whether the weight of evidence is significant or not**
- **whether the feature with significant weight of evidence is important or not. In this task, Quality of Construction, House Age and Location are more important than Kitchen Quality, Fireplaces and Central Air Conditioning because the first three features are not easy to be fixed.**
- **However, because the decision aid can sometimes be wrong about the evidence, you may want to rely on your own intuition in some cases**

**You will be scored based on your answer. Your answer is correct when you assign the highest likelihood to the correct price range. A correct answer will give you 1 point.**
The final compensation will be calculated based on your final score: a score of 0 will receive a bonus of $0, and you will receive the standard base rate of 4 GBP. **You will receive a bonus of 2 GBP if you answer at least 9 out of 12 questions correctly**.

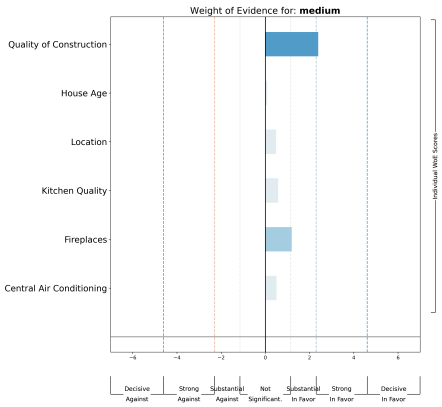
**Here we will show you three example questions and explain their answers.**
After you finish the task prediction, we will ask you to evaluate your trust and satisfaction using sliders (track bars). You will not be scored when evaluating your trust and satisfaction.

(a) Introduction page

House features:

1. Quality of Construction: 5 out of 10 (10 is the best score)
2. House Age: 58 years
3. Location: 2 out of 4 (4 is the best score)
4. Kitchen Quality: Not good
5. Fireplaces: 2
6. Central Air Conditioning: Available

The decision aid predicts that this house has a **medium** price range.
**Using the below evidence of this prediction, assign the likelihood for each option (Low, Medium, High) where 100 is the most likely, 0 is the least likely. Please total the choices to 100. You will not be able to continue unless you do so.**




Answer: We know that the decision aid's prediction is *medium*. Also, looking at the evidence provided for the *medium* price range, the evidence is positive in all features, which show that they support the hypothesis *medium* substantially. Therefore, we can set **medium** with the highest likelihood and **medium** is the correct answer.

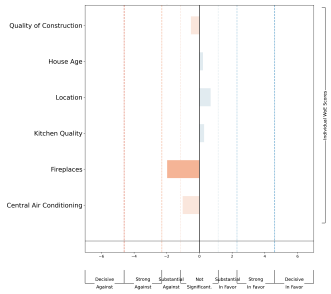

(b) An example question

Figure B.1: Training phase in **(C1) Recommendation-driven**

In the following task prediction, you will see the features (evidence) of a house, including:

1. Quality of Construction: 6 out of 10 (10 is the best score)
2. House Age: 44 years
3. Location: 2 out of 4 (4 is the best score)
4. Kitchen Quality: Not good
5. Fireplaces: Available
6. Central Air Conditioning: Available

A decision aid weights these features (also called **evidence**). You should look at the evidence provided and whether the evidence supports a particular hypothesis.




The decision aid provides **weight of evidence** for all features. Although the decision aid has a prediction, to avoid biasing your decision, we will not show you that prediction. Instead, we will show you the evidence that the decision aid used to make that prediction.
**How to read the evidence?** Looking at the above figure, we have the weight of evidence (WoE) for each feature. Positive weight of evidence (presented as blue colour) means the feature's value speaks in favour of the decision aid's prediction. Negative weight of evidence (presented as red colour) means the feature's value speaks against the decision aid's prediction. The weight of evidence is also measured as whether they are significant or not based on the horizontal axis.

There are three hypotheses for a house in this task: Low price, Medium price and High price.
For example, in this figure, House Age, Location and Kitchen Quality support **a hypothesis**. However, their weight of evidence are not significant as shown on the horizontal axis. On the other hand, Quality of Construction, Fireplaces and Central Air Conditioning are considered to work against the hypothesis.

**You will be given information about some houses. You will also be given a weighting of the evidence by the decision aid. The evidence shows the decision aid's estimate of the weight of features, so can be useful to help guide which features you pay attention to. Based on the provided evidence, your task is to assign the likelihood for each price range of the given house.**
**You can make use of the evidence provided and use these information to make a final decision**

- **whether the evidence supports or refutes your prediction**
- **whether the weight of evidence is significant or not**
- **whether the feature with significant weight of evidence is important or not. In this task, Quality of Construction, House Age and Location are more important than Kitchen Quality, Fireplaces and Central Air Conditioning because the first three features are not easy to be fixed.**
- **However, because the decision aid can sometimes be wrong about the evidence, you may want to rely on your own intuition in some cases**

**You will be scored based on your answer. Your answer is correct when you assign the highest likelihood to the correct price range. A correct answer will give you 1 point.**
The final compensation will be calculated based on your final score: a score of 0 will receive a bonus of $0, and you will receive the standard base rate of 4 GBP. **You will receive a bonus of 2 GBP if you answer at least 9 out of 12 questions correctly**.

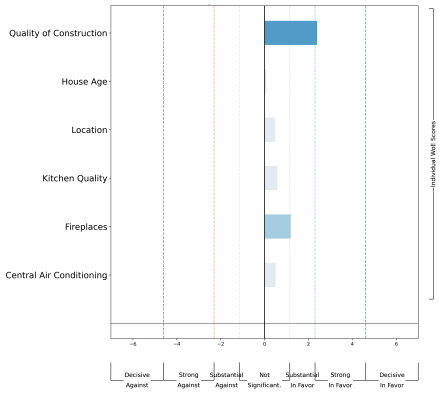
**Here we will show you three example questions and explain their answers.**
After you finish the task prediction, we will ask you to evaluate your trust and satisfaction using sliders (track bars). You will not be scored when evaluating your trust and satisfaction.

(a) Introduction page

House features:

1. Quality of Construction: 5 out of 10 (10 is the best score)
2. House Age: 58 years
3. Location: 2 out of 4 (4 is the best score)
4. Kitchen Quality: Not good
5. Fireplaces: 2
6. Central Air Conditioning: Available

**Using the below evidence of a hidden decision aid's prediction (low/medium/high), assign the likelihood for each option (Low, Medium, High) where 100 is the most likely, 0 is the least likely. Please total the choices to 100. You will not be able to continue unless you do so.**




Answer: We don't know about the decision aid's prediction, we can only see its evidence. Looking at the evidence provided, the evidence is positive in all features, which shows that they all support this prediction. Especially, the quality of construction and fireplaces significantly support this prediction. As in the house features provided, the quality of construction is average and we have 2 fireplaces. Therefore, we can argue that this prediction cannot be *high* because of the quality of construction. Also, we have a house age of 58 years, and the evidence of the hidden prediction says that the house age doesn't significantly support this prediction. So this prediction cannot be *low* because if it was low, we could have a stronger evidence in house age and also negative evidence in the number of fireplaces.

Therefore, we can set **medium** with the highest likelihood and this is the decision aid's prediction and also the correct answer.

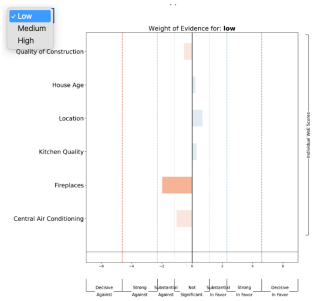

(b) An example question

Figure B.2: Training phase in **(C2) AI-explanation-only**

In the following task prediction, you will see the features (evidence) of a house, including:

1. Quality of Construction: 6 out of 10 (10 is the best score)
2. House Age: 44 years
3. Location: 2 out of 4 (4 is the best score)
4. Kitchen Quality: Not good
5. Fireplaces: Available
6. Central Air Conditioning: Available

A decision aid uses these features (also called **evidence**) to reason about the likelihood whether this house will be in a price range of: **low, medium or high**. You should look at the evidence provided and decide what should be the best price range for the given house.




**How to read the evidence?** Looking at the above figure, we have the weight of evidence (WoE) for each feature. Positive weight of evidence (presented as blue colour) means the feature's value speaks in favour of the hypothesis that the value of the house is 'Low'. Negative weight of evidence (presented as red colour) means the feature's value speaks against the hypothesis that the value of the house is 'Low'. The weight of evidence is also measured as whether they are significant or not based on the horizontal axis.

There are three hypotheses for a house in this task: *Low price, Medium price and High price.* You can use the dropdown list to change the hypothesis and see its corresponding evidence. For example, in this figure, when we choose hypothesis **low**, house age, location and kitchen quality support this hypothesis. Specifically, since the house age is 44 years old and the kitchen quality is 'not good', this implies that the house price should be low. However, their weight of evidence are not significant as shown on the horizontal axis. On the other hand, Quality of Construction, Fireplaces and Central Air Conditioning refute the hypothesis 'Low', which means they suggest this house might have a higher price based on these three features.

**You will be given information about some houses. The decision aid's prediction will not be given in this task. You will be able to view the evidence for each hypothesis. your task is to assign the likelihood for each price range of the given house.**
**You can make use of the evidence provided and use these information to make a final decision**

- **whether the evidence supports or refutes your prediction**
- **whether the weight of evidence is significant or not**
- **whether the feature with significant weight of evidence is important or not. In this task, Quality of Construction, House Age and Location are more important than Kitchen Quality, Fireplaces and Central Air Conditioning because the first three features are not easy to be fixed.**
- **However, because the decision aid can sometimes be wrong about the evidence, you may want to rely on your own intuition in some cases**

**You will be scored based on your answer. Your answer is correct when you assign the highest likelihood to the correct price range. A correct answer will give you 1 point.** The final compensation will be calculated based on your final score: a score of 0 will receive a bonus of $0, and you will receive the standard base rate of 4 GBP. **You will receive a bonus of 2 GBP if you answer at least 9 out of 12 questions correctly**.

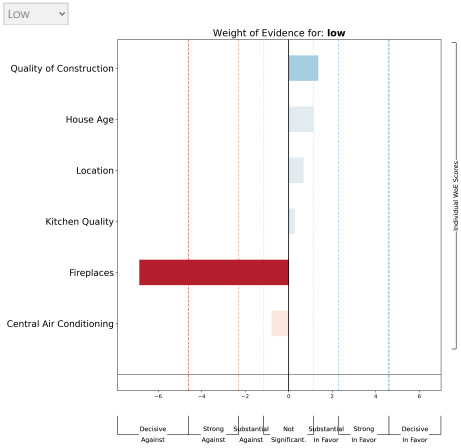
**Here we will show you three example questions and explain their answers.**
After you finish the task prediction, we will ask you to evaluate your trust and satisfaction using sliders (track bars). You will not be scored when evaluating your trust and satisfaction.

(a) Introduction page

House features:

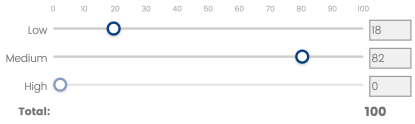
1. Quality of Construction: 5 out of 10 (10 is the best score)
2. House Age: 58 years
3. Location: 2 out of 4 (4 is the best score)
4. Kitchen Quality: Not good
5. Fireplaces: 2
6. Central Air Conditioning: Available

**Using the evidence for each hypothesis (low/medium/high), assign the likelihood for each option (Low, Medium, High) where 100 is the most likely, 0 is the least likely. Please total the choices to 100. You will not be able to continue unless you do so.** Please use the dropdown list to see the evidence for all hypotheses.




Answer: Looking at the evidence provided for three possible hypothese (low/medium/high), the evidence is positive in all features for the hypothesis *medium*.

Therefore, we can be confident when we set **medium** with the highest likelihood and this is the correct answer.




(b) An example question

Figure B.3: Training phase in **(C3) Hypothesis-driven**

House features:

1. Quality of Construction: 5 out of 10 (10 is the best score)
2. House Age: 53 years
3. Location: 1 out of 4 (4 is the best score)
4. Kitchen Quality: Not good
5. Fireplaces: 0
6. Central Air Conditioning: Available

The decision aid predicts that this house has a **low** price range. **Using the below evidence of this prediction, assign the likelihood for each option (Low, Medium, High) where 100 is the most likely, 0 is the least likely. Please total the choices to 100. You will not be able to continue unless you do so.**

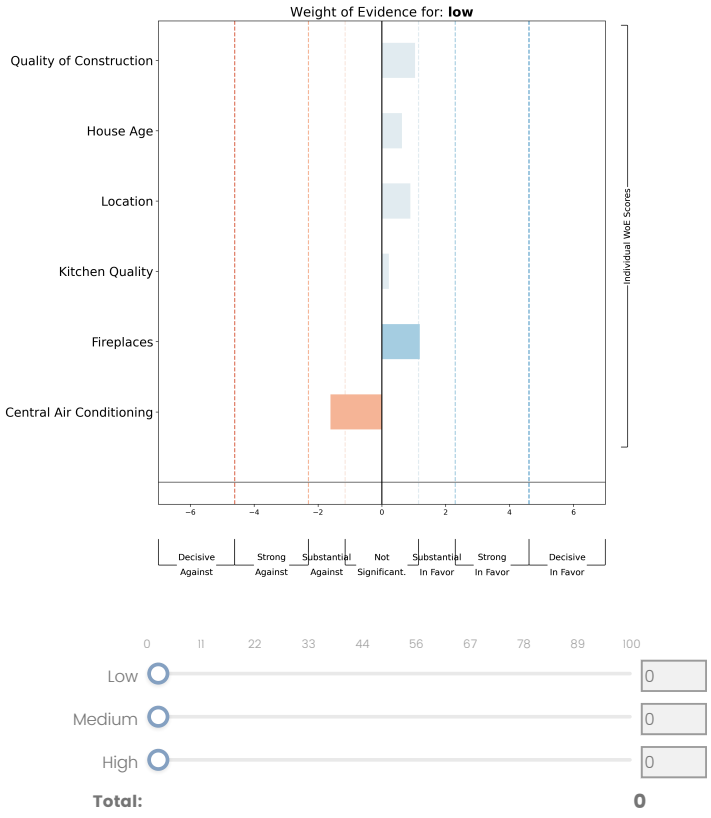



Figure B.4: A screenshot of a question in **(C1) Recommendation-driven**.

House features:

1. Quality of Construction: 5 out of 10 (10 is the best score)
2. House Age: 53 years
3. Location: 1 out of 4 (4 is the best score)
4. Kitchen Quality: Not good
5. Fireplaces: 0
6. Central Air Conditioning: Available

**Using the below evidence of a hidden decision aid's prediction (low/medium/high), assign the likelihood for each option (Low, Medium, High) where 100 is the most likely, 0 is the least likely. Please total the choices to 100. You will not be able to continue unless you do so.**

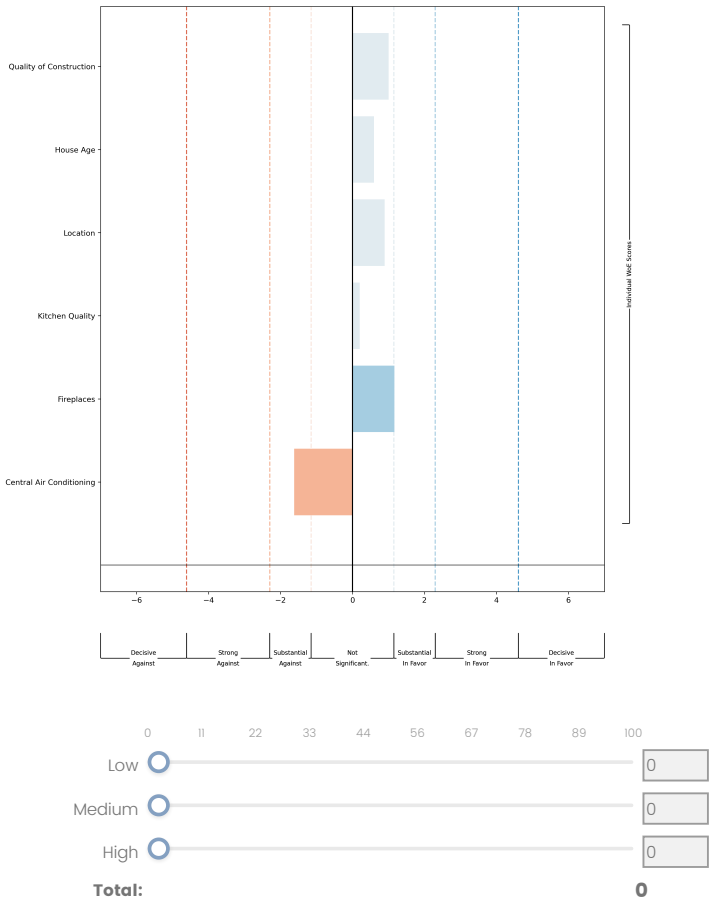



Figure B.5: A screenshot of a question in **(C2) AI-explanation-only**.

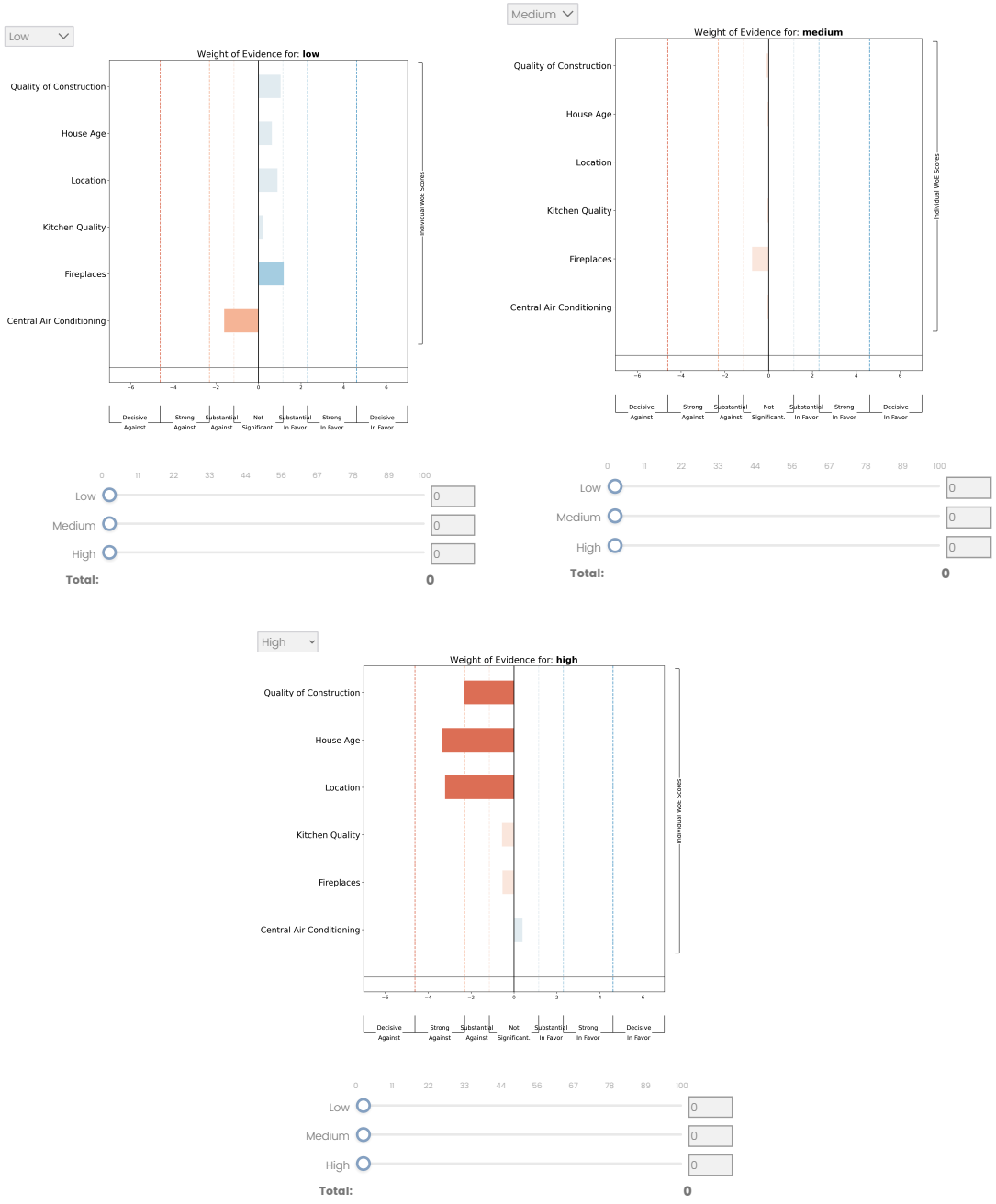



Figure B.6: Screenshots of evidence provided for all three hypotheses in **(C3) Hypothesis-driven**.

# Appendix C
# Visual Evaluative AI

## C.1 Human Experiment's Protocol

I will provide the questions and tasks that I asked the participants in the human experiment. The experiment consists of three phases: (1) a short interview about the participant's background, (2) diagnosis tasks, and (3) a semi-structured interview. The questions are as follows:

### C.1.1 Phase 1's Questions

1. *(Role)* What is your current role?

2. *(Years of experience in the current role)* How many years that you have been in this role?

3. *(Years of experience in skin cancer)* What is your experience in diagnosing skin cancer? (No experience, Beginner, Intermediate, Expert)

4. *(AI expertise)* How would you describe your experience in using AI tools, either in your work or outside of work? (No experience, Beginner, Intermediate, Expert)

5. *(DA tools used)* Have you ever used a decision-support tool to support skin cancer diagnosis? If yes, can you please describe your experience? What tools have you used?

### C.1.2 Phase 2's Questions

*I show participants the two interfaces and ask them to complete the tasks. Instructions and tutorials are provided as in Figures C.1,C.2,C.3.*

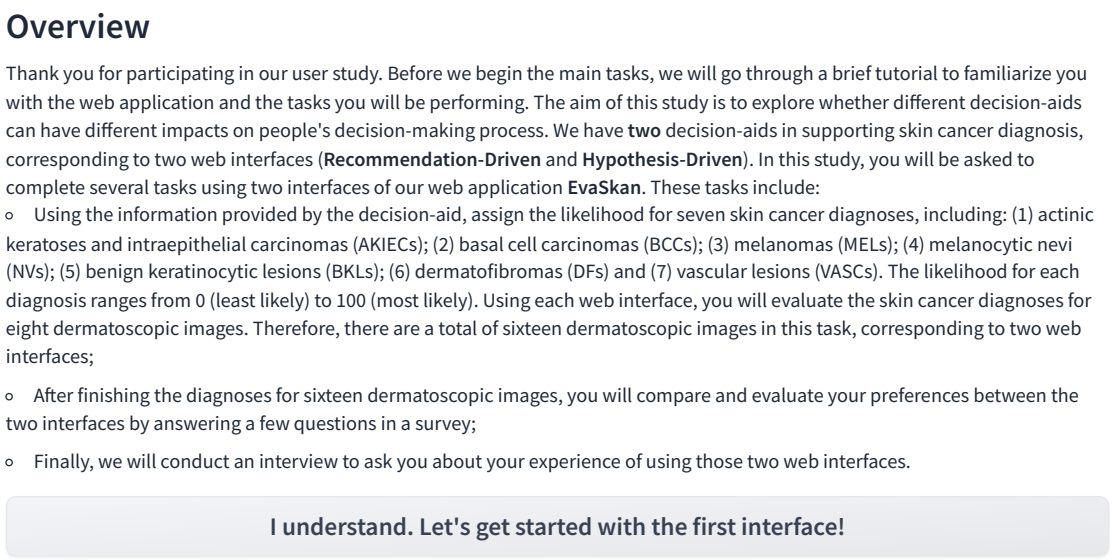



Figure C.1: Overview introduction of the human experiment

*The Qualtrics survey given to participants is shown as in Figure C.4.*

### C.1.3 Phase 3's Questions

*In this phase, I conduct a semi-structured interview by asking them to reflect on how they made the diagnoses in Phase 2 using a think-aloud protocol and open questions about the design of my decision aids.*

1. How accurate and reliable do you think this decision support is, by comparing between the recommendation-driven and hypothesis-driven?

2. What did not work well when you used this decision support? Is there anything that you are concerned about?

3. What do you think about the quality of the provided evidence? Did you look at the segmentations or the weight of evidence when making the decision?

4. Are there any other evidence that you used to make the decision? Specifically, evidence that you found yourself based on the original image.

5. What are the advantages and disadvantages of the recommendation-driven and the hypothesis-driven interface?

6. What changes would you propose for the DA to help you make better decisions?

## C.2 Web Interfaces

I show example screenshots of the two interfaces used in the human experiment. The recommendation-driven interface is shown in Figure C.5, and the hypothesis-driven interface is shown in Figure C.6.

## Understanding the Web Interface of Recommendation-Driven Decision-Aid

This tutorial will guide you through the key components of the interface.
1. **Image index**: The index of the current image, valued from 0 to 8;

2. **Dermatoscopic image**: The dermatoscopic image that we are evaluating;

3. **Decision aid's recommendation**: The recommendation for the diagnosis of the current image, given by the decision-aid. The recommendation is one of seven possible diagnoses (AKIEC, BCC, MEL, NV, BKL, DF, VASC);

4. **Evidence for**: The evidence that support the decision-aid's recommendation;

5. **Evidence against**: The evidence that refute the decision-aid's recommendation;

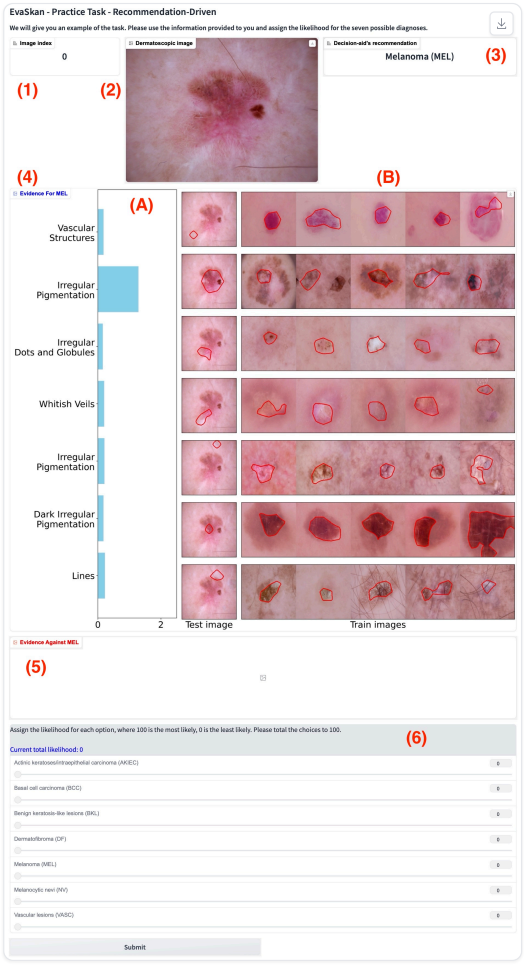
6. **Response - Assigning the likelihood**: Use the seven sliders to answer the task's question. You will need to assign the likelihood for seven possible diagnoses and ensure that the total likelihood is 100.

## How to Read the Evidence

The evidence comprise of two components, including (A) weight of evidence (on the left side) and (B) image segmentations (on the right side), explained as follows.

- **(A) weight of evidence**: We have the weight of evidence (WoE) for each feature being presented as horizontal bar charts. A positive weight of evidence (blue colour) indicates that the feature's value supports the decision-aid's recommendation according to the AI model used in the decision-aid. A negative weight of evidence (red colour) indicates that the feature's value refutes the decision-aid's recommendation according to the AI model used in the decision-aid. The weight of evidence is also measured as how much each feature contributes to the recommendation based on the horizontal axis. Note that this decision-aid can sometimes find wrong evidence or give it the wrong weight.

- **(B) image evidence**: Each feature is represented as evidence on the test image that highlight areas on the skin. We also provide five other example images in the training set with evidence that present the similar feature. Based on these evidence, you can identify the dermatoscopic feature being represented.




I understand

Use via API · Built with Gradio

Figure C.2: Tutorial of the recommendation-driven interface

## Understanding the Web Interface of Hypothesis-Driven Decision-Aid

This tutorial will guide you through the key components of the interface.

1. **Image index**: The index of the current image, valued from 0 to 8;

2. **Dermatoscopic image**: The dermatoscopic image that we are evaluating;

3. **Your hypothesis**: Select one out of seven possible diagnoses for the current image. When you select a hypothesis, the evidence will be shown correspondingly. You do not need to view all hypotheses;

4. **Evidence for**: The evidence that support the selected hypothesis;

5. **Evidence against**: The evidence that refute the selected hypothesis;

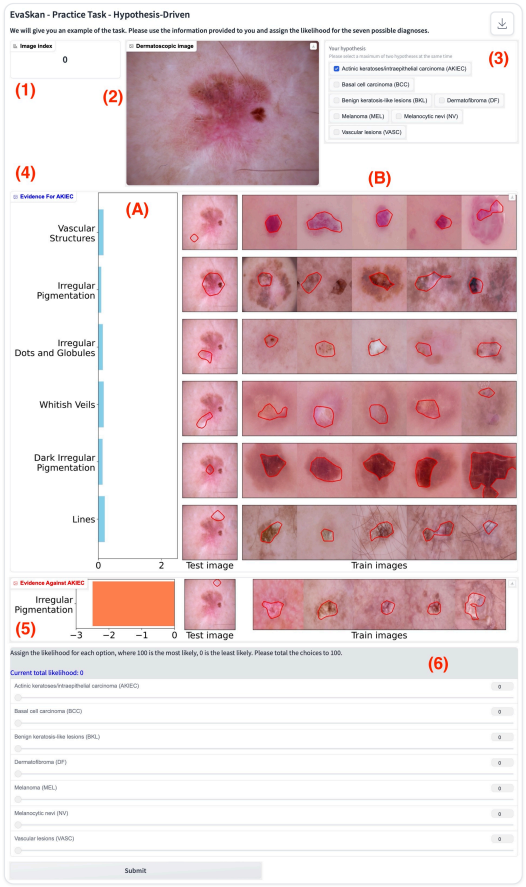
6. **Response: Assigning the likelihood**: Use the seven sliders to answer the task's question. You will need to assign the likelihood for seven possible diagnoses and ensure that the total likelihood is 100.

## How to Read the Evidence

The evidence comprise of two components, including (A) weight of evidence (on the left side) and (B) image segmentations (on the right side), explained as follows.

- **(A) weight of evidence**: We have the weight of evidence (WoE) for each feature being presented as horizontal bar charts. A positive weight of evidence (blue colour) indicates that the feature's value supports the selected hypothesis according to the AI model used in the decision-aid. A negative weight of evidence (red colour) indicates that the feature's value refutes the selected hypothesis according to the AI model used in the decision-aid. The weight of evidence is also measured as how much each feature contributes to the hypothesis based on the horizontal axis. Note that this decision-aid can sometimes find wrong evidence or give it the wrong weight.

- **(B) image evidence**: Each feature is represented as evidence on the test image that highlight areas on the skin. We also provide five other example images in the training set with evidence that present the similar feature. Based on these evidence, you can identify the dermatoscopic feature being represented.




I understand



Figure C.3: Tutorial of the hypothesis-driven interface

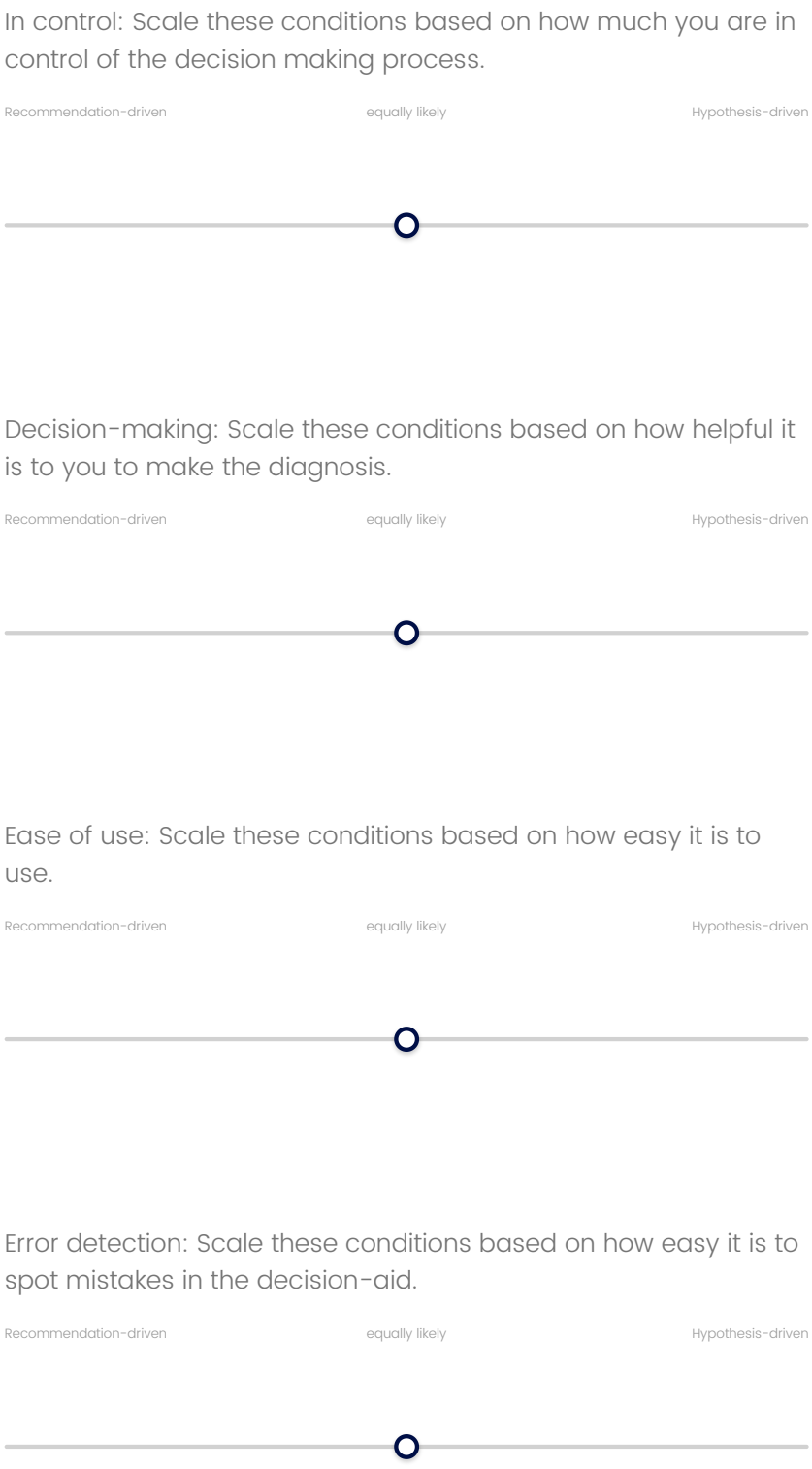



Figure C.4: Qualtrics survey for the human experiment - Bipolar scale questions

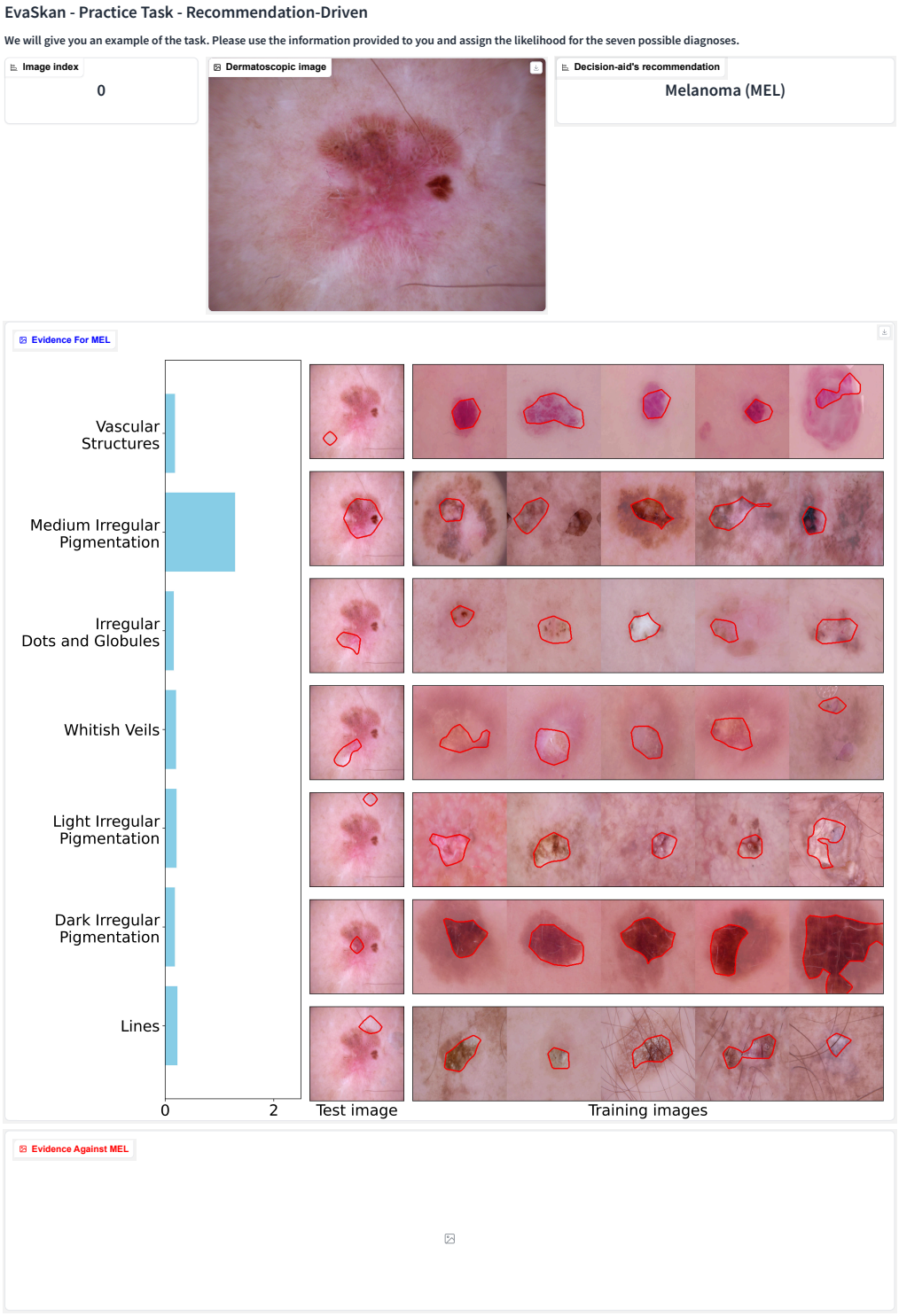



Figure C.5: The recommendation-driven interface

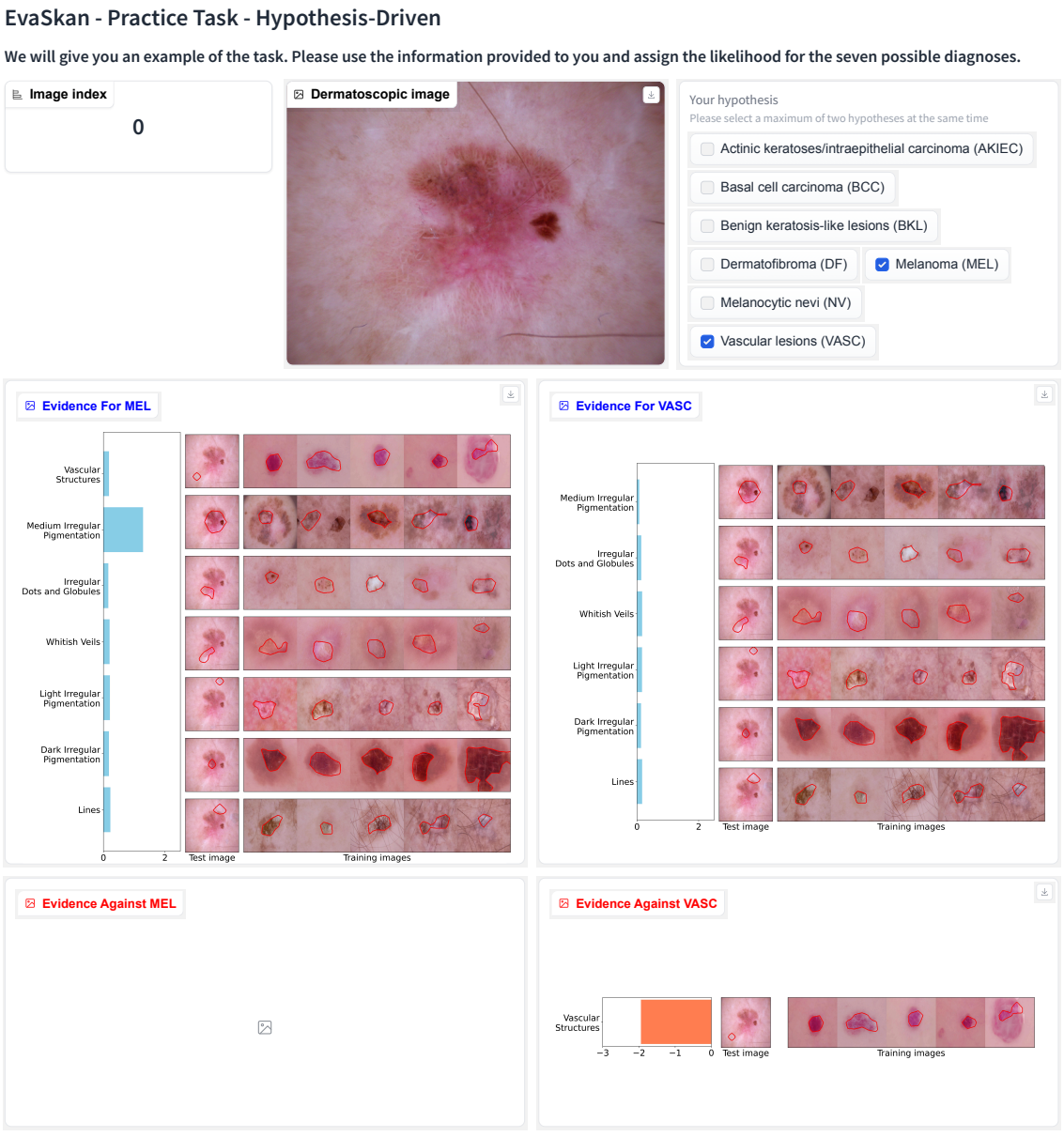



Figure C.6: The hypothesis-driven interface

# Bibliography


[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[2] Derek A Albert and Daniel Smilek. Comparing attentional disengagement between Prolific and MTurk samples. *Scientific Reports*, 13(1), 2023.

[3] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. Does Explainable Artificial Intelligence Improve Human Decision-Making? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6618–6626, 2021.

[4] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. User Study on Interpretability - Tutorial. https://github.com/dmelis/interpretwoe/blob/master/notebooks/WoE_UserStudy_Tutorial.ipynb, 2021. Accessed: 2023-05-30.

[5] Leila Amgoud and Henri Prade. Using arguments for making and explaining decisions. *Artificial Intelligence*, 173(3-4):413–436, 2009.

[6] Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. Towards Improving Trust in Context-Aware Systems by Displaying System Confidence. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, pages 9–14, 2005.

[7] Javier Antoran, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a CLUE: A Method for Explaining Uncertainty Estimates. In *9th International Conference on Learning Representations ICLR Virtual Event, Austria*, 2021.

[8] Daniel W Apley and Jingyu Zhu. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 2020.

[9] G. Argenziano, I. Zalaudek, and H.P. Soyer. Which is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology? *British Journal of Dermatology*, 151(2):512–513, 2004.

[10] Giuseppe Argenziano, Gabriella Fabbrocini, Paolo Carli, Vincenzo De Giorgi, Elena Sammarco, and Mario Delfino. Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions: Comparison of the ABCD Rule of Dermatoscopy and a New 7-Point Checklist Based on Pattern Analysis. *Archives of Dermatology*, 134(12):1563–1570, 1998.

[11] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally Weighted Learning. *Artificial Intelligence Review*, 11(1):11–73, 1997.

[12] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. Explanation in AI and law: Past, present and future. *Artificial Intelligence*, 289, 2020.

[13] Abdullah Awaysheh, Jeffrey Wilcke, François Elvinger, Loren Rees, Weiguo Fan, and Kurt L Zimmerman. Review of Medical Decision Support and Machine-Learning Methods. *Veterinary pathology*, 56(4):512–525, 2019.

[14] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter Lasecki, Dan Weld, and Eric Horvitz. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, pages 2–11, 2019.

[15] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:2429–2437, 2019.

[16] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[17] Catarina Barata, Veronica Rotemberg, Noel C. F. Codella, Philipp Tschandl, Christoph Rinner, Bengu Nisa Akay, Zoe Apalla, Giuseppe Argenziano, Allan Halpern, Aimilios Lallas, Caterina Longo, Josep Malvehy, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. A reinforcement learning model for AI-based decision support in skin cancer. *Nature Medicine*, 29 (8):1941–1946, 2023.

[18] Saba Bashir, Usman Qamar, and Farhan Hassan Khan. IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. *Journal of biomedical informatics*, 59:185–200, 2016.

[19] Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, Reciprocity, and Social History. *Games and economic behavior*, 10(1):122–142, 1995.

[20] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 78–91, 2022.

[21] 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 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.

[22] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 2006.

[23] Eliza Bobek and Barbara Tversky. Creating visual explanations improves learning. *Cognitive research: principles and implications*, 1:1–14, 2016.

[24] Andrei Bondarenko, Phan Minh Dung, Robert A Kowalski, and Francesca Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial intelligence*, 93(1-2):63–101, 1997.

[25] Ralph Peter Braun, Harold S Rabinovitz, Margaret Oliviero, Alfred W Kopf, and Jean-Hilaire Saurat. Dermoscopy of pigmented skin lesions. *Journal of the American Academy of Dermatology*, 52(1):109–121, 2005.

[26] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[27] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

[28] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[29] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. Explainable AI in Fintech Risk Management. *Frontiers in Artificial Intelligence*, 3, 2020.

[30] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *International Conference on Healthcare Informatics*, pages 160–169, 2015.

[31] Ruth M. J. Byrne. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, pages 6276–6282, 2019.

[32] Ruth MJ Byrne and Alessandra Tasso. Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory & cognition*, 27:726–740, 1999.

[33] Federico Cabitza, Chiara Natali, Lorenzo Famiglini, Andrea Campagner, Valerio Caccavella, and Enrico Gallazzi. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial Intelligence in Medicine*, 150, 2024.

[34] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW), 2019.

[35] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.

[36] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 378–389. Springer, 2005.

[37] Tirtha Chanda, Katja Hauser, Sarah Hobelsberger, Tabea-Clara Bucher, Carina Nogueira Garcia, Christoph Wies, Harald Kittler, Philipp Tschandl, Cristian Navarrete-Dechent, Sebastian Podlipnik, et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications*, 15(1), 2024.

[38] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining Image Classifiers by Counterfactual Generation. In *7th International Conference on Learning Representations, ICLR*, 2019.

[39] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. Machine Explanations and Human Understanding. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023.

[40] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2), 2023.

[41] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept Whitening for Interpretable Image Recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[42] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[43] Lingwei Cheng and Alexandra Chouldechova. Overcoming Algorithm Aversion: A Comparison between Process and Outcome Control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–27, 2023.

[44] Chun-Wei Chiang and Ming Yin. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces*, pages 148–161, 2022.

[45] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 103–119, 2024.

[46] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, Francesca Toni, et al. Argumentative xai: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4392–4399, 2021.

[47] Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y Zou. SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Advances in Neural Information Processing Systems*, 35:18157–18167, 2022.

[48] Dean De Cock. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

[49] Eoin Delaney, Derek Greene, and Mark T Keane. Uncertainty Estimation and Out-of-Distribution Detection for Counterfactual Explanations: Pitfalls and Solutions. *arXiv preprint arXiv:2107.09734*, 2021.

[50] Eoin Delaney, Derek Greene, and Mark T. Keane. Instance-Based Counterfactual Explanations for Time Series Classification. In *Case-Based Reasoning Research and Development*, pages 32–47, 2021.

[51] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 590–601, 2018.

[52] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 2015.

[53] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 275–285, 2019.

[54] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos One*, 18(3), 2023.

[55] Dheeru Dua, Casey Graff, et al. UCI Machine Learning Repository, 2017. URL http://archive.ics.uci.edu/ml.

[56] Phan Minh Dung. On the acceptability of arguments and its fundamental role in

nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[57] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. *Advances in Neural Information Processing Systems*, 35:21400–21413, 2022.

[58] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[59] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. The Impact of Algorithmic Risk Assessments on Human Predictions and Its Analysis via Crowdsourcing Studies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 2021.

[60] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 1362–1374, 2022.

[61] Maximilian Förster, Philipp Hühn, Mathias Klier, and Kilian Kluge. Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations. In *54th Hawaii International Conference on System Sciences, HICSS*, pages 1–10, 2021.

[62] Jerome H Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of statistics*, pages 1189–1232, 2001.

[63] Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim. On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, 109(4):853–898, 2020.

[64] Krzysztof Z. Gajos and Lena Mamykina. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*, pages 794–806, 2022.

[65] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

[66] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, 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), 2021.

[67] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9273–9282, 2019.

[68] Soumya Ghosh, Q Vera Liao, Karthikeyan Natesan Ramamurthy, Jiri Navratil, Prasanna Sattigeri, Kush R Varshney, and Yunfeng Zhang. Uncertainty Quantification 360: A Holistic Toolkit for Quantifying and Communicating the Uncertainty of AI. *arXiv preprint arXiv:2106.01410*, 2021.

[69] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[70] IJ Good. Weight of evidence: A brief survey. In *Bayesian statistics*, volume 2, pages 249–270. Elsevier, 1985.

[71] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual Visual Explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2376–2384, 2019.

[72] Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. Interpretable Credit Application Predictions With Counterfactual Explanations. *arXiv preprint arXiv:1811.05245*, 2018.

[73] Mara Graziani, Vincent Andrearczyk, Stéphane Marchand-Maillet, and Henning Müller. Concept attribution: Explaining CNN decisions to physicians. *Computers in Biology and Medicine*, 123, 2020.

[74] Dale Griffin and Amos Tversky. The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24(3):411–435, 1992.

[75] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems*, 34(6):14–23, 2019.

[76] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

[77] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL https://www.gurobi.com.

[78] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[79] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[80] Sarita Herse, Jonathan Vitale, Benjamin Johnston, and Mary-Anne Williams. Using Trust to Determine User Decision Making & Task Outcome During a Human-Agent Collaborative Task. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 73–82, 2021.

[81] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[82] Robert R Hoffman, Timothy Miller, and William J Clancey. Psychology and AI at a

crossroads: How might complex systems explain themselves? *The American journal of psychology*, 135(4):365–378, 2022.

[83] 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.

[84] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence*, 301, 2021.

[85] Aya Hussein, Sondoss Elsawah, and Hussein A Abbass. Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions. *Human Factors*, 62(8):1237–1248, 2020.

[86] 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 antidepressant selection. *Translational Psychiatry*, 11(1), 2021.

[87] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635, 2021.

[88] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, 2021.

[89] M Janda and HP Soyer. Can clinical decision making be enhanced by artificial intelligence? *British Journal of Dermatology*, 180(2), 2019.

[90] Monika Janda, Caitlin Horsham, Uyen Koh, Nicole Gillespie, Dimitrios Vagenas, Lois J Loescher, Clara Curiel-Lewandrowski, Rainer Hofmann-Wellenhof, and

H Peter Soyer. Evaluating healthcare practitioners' views on store-and-forward teledermoscopy services for the diagnosis of skin cancer. *Digital Health*, 5, 2019.

[91] 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.

[92] Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya Gupta. To Trust or Not to Trust A Classifier. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5546–5557, 2018.

[93] Jinglu Jiang, Surinder Kahai, and Ming Yang. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies*, 165, 2022.

[94] Weiwei Jiang, Zhanna Sarsenbayeva, Niels van Berkel, Chaofan Wang, Difeng Yu, Jing Wei, Jorge Goncalves, and Vassilis Kostakos. User Trust in Assisted Decision-Making Using Miniaturized Near-Infrared Spectroscopy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[95] Susan Joslyn and Sonia Savelli. Visualizing Uncertainty for Non-Expert End Users: The Challenge of the Deterministic Construal Error. *Frontiers in Computer Science*, 2, 2021.

[96] Susan L Joslyn and Jared E LeClerc. Uncertainty Forecasts Improve Weather-Related Decisions and Attenuate the Effects of Forecast Error. *Journal of Experimental Psychology: Applied*, 18(1), 2012.

[97] Daniel Kahneman. Thinking, Fast and Slow. *Farrar, Straus and Giroux*, 2011.

[98] Patricia K Kahr, Gerrit Rooks, Martijn C Willemsen, and Chris CP Snijders. Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions. *ACM Transactions on Interactive Intelligent Systems*, 14(4):1–30, 2024.

[99] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees. In *International Conference on Artificial Intelligence and Statistics*, pages 1846–1870, 2022.

[100] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

[101] Mark T. Keane and Barry Smyth. Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). In *Case-Based Reasoning Research and Development: 28th International Conference, ICCBR*, pages 163–178, 2020.

[102] 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. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4466–4474, 2021.

[103] 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, 2021.

[104] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2668–2677, 2018.

[105] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[106] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. Explaining the Gap: Visualizing One's Predictions Improves Recall and Comprehension of Data. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1375–1386, 2017.

[107] Harald Kittler, Ashfaq A Marghoob, Giuseppe Argenziano, Cristina Carrera, Clara Curiel-Lewandrowski, Rainer Hofmann-Wellenhof, Josep Malvehy, Scott Menzies, Susana Puig, Harold Rabinovitz, et al. Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the International Society of Dermoscopy. *Journal of the American Academy of Dermatology*, 74 (6):1093–1106, 2016.

[108] Joshua Klayman. Varieties of Confirmation Bias. *Psychology of learning and motivation*, 32:385–418, 1995.

[109] Gary Klein. A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition*, 4(3):164–168, 2015.

[110] Gary Klein, Brian Moon, and Robert R Hoffman. Making sense of sensemaking 1: Alternative perspectives. *IEEE intelligent systems*, 21(4):70–73, 2006.

[111] Gary Klein, Brian Moon, and Robert R Hoffman. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems*, 21(5):88–92, 2006.

[112] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. A Data–Frame Theory of Sensemaking. In *Expertise Out of Context*, pages 118–160. Psychology Press, 2007.

[113] Gary A Klein. *Sources of Power: How People Make Decisions*. MIT press, 2017.

[114] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. Trust and reliance on AI - An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160:108352, 2024.

[115] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *International conference on machine learning*, pages 5338–5348, 2020.

[116] Janet Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, 2014.

[117] Daniel C Krawczyk. Chapter 11 - Decision Making and Abductive Reasoning. In *Reasoning*, pages 255–282. Academic Press, 2018.

[118] Volodymyr Kuleshov and Percy S Liang. Calibrated Structured Prediction. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[119] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M. Burnett, Stephen Perona, Amy J. Ko, and Ian Oberst. Why-Oriented End-User Debugging of Naive Bayes Text Classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1):1–31, 2011.

[120] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 126–137, 2015.

[121] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *International conference on machine learning*, pages 5491–5500. PMLR, 2020.

[122] Peter D Kvam and Timothy J Pleskac. Strength and weight: The determinants of choice and confidence. *Cognition*, 152:170—180, 2016.

[123] Vivian Lai and Chenhao Tan. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38, 2019.

[124] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 1369–1385, 2023.

[125] Kathryn Ann Lambe, Gary O'Reilly, Brendan D Kelly, and Sarah Curristan. Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review. *BMJ Quality & Safety*, 25(10):808–820, 2016. ISSN 2044-5415.

[126] Thao Le. Explaining the Uncertainty in AI-Assisted Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):16119–16120, 2023.

[127] Thao Le, Tim Miller, Ronal Singh, and Liz Sonenberg. Explaining Model Confidence Using Counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):11856–11864, 2023.

[128] Thao Le, Tim Miller, Liz Sonenberg, and Ronal Singh. Towards the New XAI: A Hypothesis-Driven Approach to Decision Support Using Evidence. In *27th European Conference on Artificial Intelligence*, pages 850–857, 2024.

[129] Thao Le, Tim Miller, Peter Soyer, Ronal Singh, and Liz Sonenberg. EvaSkan: An Evaluative Skin Cancer Tool for Decision Support. 2024. URL osf.io/d9csz.

[130] Thao Le, Tim Miller, Ruihan Zhang, Liz Sonenberg, and Ronal Singh. Visual Evaluative AI: A Hypothesis-Driven Tool with Concept-Based Explanations and Weight of Evidence. *arXiv preprint arXiv:2407.04710*, 2024.

[131] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[132] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19, 1995.

[133] Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job

Candidate Screening. *Explainable and interpretable models in computer vision and machine learning*, pages 197–253, 2018.

[134] Brian Y Lim and Anind K Dey. Assessing Demand for Intelligibility in Context-Aware Applications. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 195–204, 2009.

[135] Brian Y Lim and Anind K Dey. Design of an Intelligible Mobile Context-Aware Application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, pages 157–166, 2011.

[136] 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.

[137] Shihong Ling, Yutong Zhang, and Na Du. More Is Not Always Better: Impacts of AI-Generated Confidence and Explanations in Human–Automation Interaction. *Human Factors*, 2024.

[138] Peter Lipton. Contrastive Explanation. *Royal Institute of Philosophy Supplements*, 27: 247–266, 1990.

[139] Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257, 2007.

[140] Ana Lucic, Hinda Haned, and Maarten de Rijke. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 90–98, 2020.

[141] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Computer Methods and Programs in Biomedicine*, 215, 2022.

[142] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[143] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2024.

[144] Shuai Ma, Chenyi Zhang, Xinru Wang, Xiaojuan Ma, and Ming Yin. Beyond Recommender: An Exploratory Study of the Effects of Different AI Roles in AI-Assisted Decision Making. *arXiv preprint arXiv:2403.01791*, 2024.

[145] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

[146] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable Reinforcement Learning through a Causal Lens. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2493–2500, 2020.

[147] Roger C Mayer, James H Davis, and F David Schoorman. An Integrative Model of Organizational Trust. *Academy of management review*, 20(3):709–734, 1995.

[148] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9(1):35–47, 2021.

[149] Georg Meyer, Gediminas Adomavicius, Paul E Johnson, Mohamed Elidrisi, William A Rush, JoAnn M Sperl-Hillen, and Patrick J O'Connor. A Machine Learning Approach to Improving Dynamic Decision Making. *Information Systems Research*, 25(2):239–263, 2014.

[150] Tim Miller. Explanation in Artificial Intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[151] Tim Miller. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021.

[152] Tim Miller. Are we measuring trust correctly in explainability, interpretability, and transparency research? *CHI Workshop on Trust and Reliance in AI-Human Teams*, 2022.

[153] Tim Miller. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 333–342, 2023.

[154] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv preprint arXiv:1712.00547*, 2017.

[155] Sanjay Modgil and Henry Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.

[156] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.

[157] Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

[158] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–39, 2024.

[159] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[160] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. The ABCD

rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994.

[161] Limor Nadav-Greenberg and Susan L Joslyn. Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, 3(3):209–227, 2009.

[162] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option. In *Machine Learning in Systems Biology*, pages 65–81. PMLR, 2009.

[163] Raymond S Nickerson. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of general psychology*, 2(2):175–220, 1998.

[164] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International journal of qualitative methods*, 16(1), 2017.

[165] Conor Nugent, Dónal Doyle, and Pádraig Cunningham. Gaining Insight through Case-Based Explanation. *Journal of Intelligent Information Systems*, 32:267–295, 2009.

[166] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3), 2018.

[167] MM Omodei, AJ Wearing, J McLennan, GC Elliott, and JM Clancy. More is better? Problems of self-regulation in naturalistic decision making settings, 2005.

[168] Stuart Oskamp. Overconfidence in case-study judgments. *Journal of consulting psychology*, 29(3), 1965.

[169] Lace Padilla, Matthew Kay, and Jessica Hullman. *Uncertainty Visualization*, pages 1–18. John Wiley & Sons, Ltd, 2021.

[170] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[171] Cristiano Patrício, João C. Neves, and Luís F. Teixeira. Explainable Deep Learning Methods in Medical Image Classification: A Survey. *ACM Computing Surveys*, 2023.

[172] Pavansubhash. IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset, 2017. Accessed: 2023-03-01.

[173] Judea Pearl. *Causality*. Cambridge university press, 2009.

[174] Charles Sanders Peirce. *Collected papers of Charles Sanders Peirce*, volume 5. Harvard University Press, 1974.

[175] Oskar Pfungst. *Clever Hans (the horse of Mr. Von Osten) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston, 1911.

[176] Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based Explainable Artificial Intelligence: A Survey. *arXiv preprint arXiv:2312.12936*, 2023.

[177] Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russ Greiner, D. S. Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. Visual Explanation of Evidence in Additive Classifiers. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence - Volume 2*, pages 1822–1829, 2006.

[178] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and Measuring Model Interpretability. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021.

[179] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. Understanding Uncertainty: How Lay Decision-makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th international conference on intelligent user interfaces*, pages 379–396, 2023.

[180] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding Fast and Slow: The Role of Cognitive Biases in AI-

Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), 2022.

[181] Roger Ratcliff and Philip L Smith. A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review*, 111(2):333–367, 2004.

[182] Amy Rechkemmer and Ming Yin. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2022.

[183] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, 12(1), 2022.

[184] Valerie F Reyna and Charles J Brainerd. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and individual differences*, 18 (1):89–107, 2008.

[185] 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, 2016.

[186] Maria Riveiro and Serge Thill. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence*, 298, 2021.

[187] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. Understanding the Role of Explanation Modality in AI-Assisted Decision-Making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 223–233, 2022.

[188] Enrico Rukzio, John Hamard, Chie Noda, and Alexander De Luca. Visualization of Uncertainty in Context Aware Mobile Applications. In *Proceedings of the 8th Confer-*

*ence on Human-Computer Interaction with Mobile Devices and Services*, pages 247–250, 2006.

[189] Chris Russell. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

[190] Swati Sachan, Jian-Bo Yang, Dong-Ling Xu, David Eraso Benavides, and Yang Li. An Explainable AI Decision-Support-System to Automate Loan Underwriting. *Expert Systems with Applications*, 144, 2020.

[191] Sonia Savelli and Susan Joslyn. The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. *Applied Cognitive Psychology*, 27 (4):527–541, 2013.

[192] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. Trust and reliance in XAI–distinguishing between attitudinal and behavioral measures. *arXiv preprint arXiv:2203.12318*, 2022.

[193] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel Van Den Bosch. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154, 2021.

[194] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[195] Burr Settles. *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences, 2009.

[196] Dominik Seuß. Bridging the gap between explainable ai and uncertainty quantification to enhance trustability. *arXiv preprint arXiv:2105.11828*, 2021.

[197] Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recom-

mendations. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 1330–1340, 2022.

[198] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, 6th edition, 2016.

[199] Ronal Singh, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. Directive Explanations for Actionable Explainability in Machine Learning Applications. *ACM Transactions on Interactive Intelligent Systems*, 13(4):1–26, 2023.

[200] Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pages 32633–32653, 2023.

[201] Alvin T. Tan. Cracking the Ames Housing Dataset with Linear Regression. https://github.com/at-tan/Cracking_Ames_Housing_OLS, 2021. Accessed: 2023-05-30.

[202] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, 1(4), 2020.

[203] Stephen E Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

[204] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), 2018.

[205] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

[206] Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, 1974.

[207] Amos Tversky and Daniel Kahneman. The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481):453–458, 1981.

[208] Anne Marthe Van der Bles, Sander Van Der Linden, Alexandra LJ Freeman, James Mitchell, Ana B Galvao, Lisa Zaval, and David J Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society open science*, 6(5), 2019.

[209] Anne Marthe Van Der Bles, Sander van der Linden, Alexandra LJ Freeman, and David J Spiegelhalter. The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, 117(14):7672–7683, 2020.

[210] Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, and Mark Neerincx. Contrastive Explanations with Local Foil Trees. *arXiv preprint arXiv:1806.07470*, 2018.

[211] Jasper van der Waa, Tjeerd Schoonderwoerd, Jurriaan van Diggelen, and Mark Neerincx. Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, 144, 2020.

[212] 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, 2021.

[213] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

[214] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. Argumentation and explainable artificial intelligence: A survey. *The Knowledge Engineering Review*, 36:e5, 2021.

[215] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. The Effects of Explanations on Automation Bias. *Artificial Intelligence*, 322, 2023.

[216] Andrey V. Vlasov, Oksana O. Zinchenko, Zhenjie Zhao, Mansur Bakaev, and Arsenjy Karavaev. The Design of a Trust-based Game as a Conversational Component of Interactive Environment for a Human-agent Negotiation. In *Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI*, pages 19–23, 2021.

[217] Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. Uncalibrated Models Can Improve Human-AI Collaboration. *Advances in Neural Information Processing Systems*, 35:4004–4016, 2022.

[218] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31, 2017.

[219] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. 2011.

[220] Charles Wan, Rodrigo Belo, and Leid Zejnilovic. Explainability's Gain is Optimality's Loss? How Explanations Bias Decision-Making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 778–787, 2022.

[221] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2019.

[222] Danding Wang, Wencan Zhang, and Brian Y Lim. Show or suppress? Managing input uncertainty in machine learning model explanations. *Artificial Intelligence*, 294, 2021.

[223] Jennifer Wang and Angela Moulden. AI Trust Score: A User-Centered Approach to Building, Designing, and Measuring the Success of Intelligent Workplace Features.

In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[224] 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.

[225] Xinru Wang and Ming Yin. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Transactions on Interactive Intelligent Systems*, 12 (4):1–36, 2022.

[226] Greta Warren, Barry Smyth, and Mark T. Keane. "Better" Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI). In *Case-Based Reasoning Research and Development: 30th International Conference, ICCBR*, pages 63–78, 2022.

[227] Greta Warren, Ruth MJ Byrne, and Mark T Keane. Categorical and Continuous Features in Counterfactual Explanations of AI Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 171–187, 2023.

[228] Greta Warren, Eoin Delaney, Christophe Guéret, and Mark T Keane. Explaining Multiple Instances Counterfactually: User Tests of Group-Counterfactuals for XAI. In *International Conference on Case-Based Reasoning*, pages 206–222, 2024.

[229] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. Can Large Language Model Agents Simulate Human Trust Behaviors? *arXiv preprint arXiv:2402.04559*, 2024.

[230] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017.

[231] Siyuan Yan, Zhen Yu, Xuelin Zhang, Dwarikanath Mahapatra, Shekhar S Chandra, Monika Janda, Peter Soyer, and Zongyuan Ge. Towards Trustable Skin Cancer

Diagnosis via Rewriting Model's Decision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11568–11577, 2023.

[232] J. Frank Yates and Georges A. Potworowski. Evidence-Based Decision Management. In *The Oxford Handbook of Evidence-Based Management*, pages 198–222. Oxford University Press, 2012.

[233] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On Completeness-aware Concept-Based Explanations in Deep Neural Networks. *Advances in neural information processing systems*, 33:20554–20565, 2020.

[234] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc Concept Bottleneck Models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[235] Fatima Al Zegair, Nathasha Naranpanawa, Brigid Betz-Stablein, Monika Janda, H Peter Soyer, and Shekhar S Chandra. Application of Machine Learning in Melanoma Detection and the Identification of 'Ugly Duckling' and Suspicious Naevi: A Review. *arXiv preprint arXiv:2309.00265*, 2023.

[236] Zhiwei Zeng, Xiuyi Fan, Chunyan Miao, Cyril Leung, Jing Jih Chin, and Yew Soon Ong. Context-based and explainable decision making with argumentation. 2018.

[237] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2022.

[238] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:11682–11690, 2021.

[239] 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 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

[240] Zelun Tony Zhang and Heinrich Hußmann. How to Manage Output Uncertainty: Targeting the Actual End User Problem in Interactions with AI. In *Joint Proceedings of the ACM IUI Workshops co-located with 26th ACM Conference on Intelligent User Interfaces*, volume 2903, 2021.

[241] Qiaoting Zhong, Xiuyi Fan, Xudong Luo, and Francesca Toni. An explainable multi-attribute decision model based on argumentation. *Expert Systems with Applications*, 117:42–61, 2019.

[242] 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.

[243] Jianlong Zhou, Syed Z. Arshad, Simon Luo, and Fang Chen. Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making. In *Human-Computer Interaction – INTERACT*, pages 23–39, 2017.