2023

Probability Expressions in AI Decision Support: Impacts on Human+AI Team Performance

Elias Spinn

Probability Expressions in AI Decision Support: Impacts on Human+AI Team Performance

Elias Spinn

A dissertation submitted in partial fulfilment of the requirements of

Dublin Institute of Technology for the degree of

M.Sc. in Computing - Advanced Software Development

15^{th} June 2023

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing - Advanced Software Development, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Elias Spinn

Date: 15/06/2023

Abstract

AI decision support systems aim to assist people in highly complex and consequential domains to make efficient, effective, and high-quality decisions. AI alone cannot be guaranteed to be correct in these complex decision tasks, and a human is often needed to ensure decision accuracy. The ambition is for these human+AI teams to perform better together than either would individually. To realise this, decision makers must trust their AI partners appropriately, knowing when to rely on their recommendations and when to be skeptical. However, research has shown that decision makers often either mistrust and underutilise these systems, or trust them blindly. Researchers in the fields of HCI and XAI have worked on developing methods that continuously manage an appropriate level of user trust.

Despite the probabilistic nature of ML-based AI, little attention has been given to understand how the research area of uncertainty communication might provide solutions to this challenge. This study draws on that research, and asks how different forms of expressing probability in AI decision support systems might affect human+AI team performance. A series of task-based user tests were conducted to evaluate the use of numerical, verbal, and verbal-numerical probability expressions in communicating AI prediction confidence to decision makers. Results indicated that numerical expressions may be most effective when decision makers use AI decision support. However, findings were inconclusive due to a limited number of participants who used AI decision support during testing.

Acknowledgments

I would like to express my sincere gratitude to Prof. Sarah Jane Delany for the invaluable guidance, support and expertise throughout this dissertation project. Thank you for all those hours spent steering me along this path. I would also like to thank Dr Robert Ross, Dr Luca Longo, and Brendan Tierney for their guidance in finding a dissertation topic, which was arguably the biggest challenge of my whole MSc course. My sincere appreciation to all the course lecturers and staff, which a special mention to Andrea Curley and Dr Emma Murphy for their support during the most challenging times.

I would like to thank Gagan Bansal and Prof. Daniel S. Weld for sharing their work on human+AI teams with me, without which this project would not have been possible. I also thank all my work colleagues in Workday who have supported me throughout this course.

My biggest dept of gratitude goes to my wife Audrey Reilly, who's unwavering encouragement and support has been the bedrock of this project.

Contents

List of Figures

List of Tables

35table.caption.17

List of Acronyms

AI	Artificial intelligence
DSS	Decision support system
HCI	Human Computer Interaction
SUS	System Usability Scale
UI	User interface
XAI	Explainable AI
XUI	Explanation user interface

Chapter 1

Introduction

1.1 Background

Artificial intelligent (AI) systems are increasingly being used by people to make decision. However, not all decisions are equally complex or consequential (Shneiderman, 2021). In domains where complexity and consequences are high, decision support systems and applications are frequently used to assist individuals in the decision-making process (Riveiro, Helldin, Falkman, & Lebram, 2014).

Decision support systems (DSS) aim to support users to make high-quality, efficient, and effective decisions, especially when dealing with large amount of data in diverse and variable contexts (Riveiro et al., 2014; McGuirl & Sarter, 2006). Leveraging artificial intelligence, DS systems automate certain parts of the decision-making process, analysing often huge volumes of data to provide the user with valuable insights and recommendations (Gunning, 2019). The ambition is to enable high performing human+AI teams by combining the intelligence and abilities of both (Buçinca, Malaya, & Gajos, 2021).

AI decision support comes with a certain degree of uncertainty. Uncertainty from the probabilistic nature of machine learning (ML) based AI, (Zhang, Liao, & Bellamy, 2020), the reliability of the underlying data, (Riveiro et al., 2014), and the inherent uncertainty in making any statements of fact or predictions about the future (Teigen, 2022). It becomes important for decision makers to be able to judge whether to

trust or distrust a given recommendation (Zhang et al., 2020). In order to make that judgement, decision makers need to understand the strengths and weaknesses of the systems, and the logic and reasoning behind a recommendation (Gunning, 2019).

Understanding the rational behind a ML model's prediction is often difficult. The limited uptake of many AI systems has been attributed, in part, to the lack of transparency and comprehensibility in ML-based AI (Miller, 2019; Dodge, Liao, Zhang, Bellamy, & Dugan, 2019). This has sparked renewed interest and research in explainable AI (XAI), developing methods to make ML models easier to understand, trust, and control (Gunning, 2019). It is thought that users will feel more confident using an AI system that offers them an explanations for its recommendations (Vilone & Longo, 2021).

The challenge of user distrust is counterbalanced by the issue of over-reliance (Buçinca et al., 2021), where DSS users no longer do their own exploration and analysis of information, and instead rely solely on the automated decision support (McGuirl & Sarter, 2006). Studies in XAI have noted that explanations can potentially increase inappropriate levels of trust in AI recommendations (Bansal et al., 2021). This leads to fragile human+AI teams that underperform when the AI is incorrect (McGuirl & Sarter, 2006), and don't benefit from the collective capability of both.

There is a general recognition for the need of trust calibration, the continuous management of user trust at an appropriate level (Bansal et al., 2021; Buçinca et al., 2021; McGuirl & Sarter, 2006; Dubiel, Daronnat, & Leiva, 2022; Zhang et al., 2020). Trust calibration encourages decision makers to critically evaluate decision support, scrutinizing AI recommendations when necessary. As the user corrects the AI's incorrect results, the team's combined performance surpasses that of either human or AI working individually (Bansal et al., 2021).

Encouraging critical reflection in DSS users has challenges, as people often avoid analytical thinking and may prefer systems that don't reduce over-reliance (Buçinca, Lin, Gajos, & Glassman, 2020). Preferred DS systems may not necessarily improve performance (Buçinca et al., 2021).

1.2 Research problem

A common approach in XAI and AI DSS studies is to display a confidence score along with every AI recommendation (Bansal et al., 2021; Zhang et al., 2020; Jesus et al., 2021; McGuirl & Sarter, 2006). The score expresses the probability of the AI being correct (Zhang et al., 2020), giving users insights into the system's performance on a case-by-case basis (McGuirl & Sarter, 2006).

Confidence scores have shown promise in effective trust calibration (McGuirl & Sarter, 2006; Zhang et al., 2020). However, the best method of communicating these confidence scores seems to be an open question. These studies provide little justification for how confidence scores are shown to users. Communicating uncertainty in these studies seems to, to quote Spiegelhalter, Pearson, and Short (2011), rely on "good intuition rather than well-researched principles". Understanding how different methods of expressing probability impact human+AI decision performance, seems to be an important question, when considering the probabilistic nature of ML-based AI and the objective of enabling human+AI teams.

This study asks 2 research questions.

1. What is the effect of different methods for expressing ML prediction confidence in AI decision support systems on the performance of human+AI teams?

2. To what extent do different methods for expressing ML prediction confidence effectively calibrate decision makers' trust in AI decision support recommendations?

This study draws on work in uncertainty communication, a rich area of research with a long-standing history (Spiegelhalter et al., 2011). Specifically, it focuses on verbal and numerical expressions of probability. The literature offers extensive comparisons and insightful discussions on these 2 forms of communicating uncertainty, but is equivocal about the ideal format to use (Knoblauch, Stauffacher, & Trutnevyte, 2018).

Each has been described by its distinct qualities and characteristics. Numerical form are considered precise, efficient (Spiegelhalter et al., 2011), and are more easily

compared (Jaffe-Katz, Budescu, & Wallsten, 1989). Verbal expressions contain a richer amount of information, and can convey recommendations and warnings more clearly (Teigen, 2022). The qualities of both forms indicates potential advantages in trust calibration. However, there is currently a lack of empirical research that compares them within the context of AI decision support.

1.3 Research Objectives

The objective of this study is to measure decision performance of human+AI teams, in a scenario where the individual performance of human and AI are similar. Teams' performances are compared between conditions where AI confidence scores are expressed either numerically, verbally, or both. As an indicator of trust calibration, these conditions are compared by how much the decision makers improve AI performance when recommendations are incorrect. The final objective is to measure decision makers' subjective preferences, and compare those to team performance.

1.4 Research Methodologies

This is an empirical AI decision support evaluation, using a task-based, user-centered approach. It builds on previous work done by Bansal et al. (2021) on human+AI team performance in the field of XAI.

In a series of unmoderated user tests, study participants reviewed and labeled beer reviews as either positive or negative. AI decision support provided recommendations based on a predicted sentiment, accompanied by a confidence score, expressed in numerical, verbal, or verbal-numerical form. Participants gave a subjective evaluation of the AI decision support after completing the labeling task.

The decision accuracy of participants was measured by comparing their decisions to a ground truth. Performance on all reviews was compared across different conditions, as well as performance on reviews where AI recommendations were incorrect. Conditions were also compared by participants' subjective evaluations.

1.5 Scope and Limitations

Participants completed the unmoderated user tests online, using their own devices, and at a time and location of their choice. Unmoderated testing offers the advantage of obtaining larger samples more quickly. However, a limitation of this approach is that the test environment is not controlled.

This study's participant sample size (∼32 per condition) is moderate compared to other studies in XAI, DSS, and uncertainty communication. Studies in these areas that use unmoderated user tests can have participant sample sizes ranging from 50 to 100 participants per condition (Bansal et al., 2021; Budescu, Weinberg, & Wallsten, 1988). However, there are also studies with smaller participant sample sizes (15, 11, 3), but which employ moderated user tests (Riveiro et al., 2014; McGuirl & Sarter, 2006; Jesus et al., 2021).

The study's choice of a incomplex and inconsequential decision task, which does not require specialized expertise and is accessible for general participation (Bansal et al., 2021), means that the findings may not generalise to complex and consequential domains.

1.6 Document Outline

Chapter 2 presents a review of literature in DS systems, XAI, and uncertainty communication. Chapter 3 describes the study design, methodology, user test design, UI design, application development, and recruitment. Chapter 4 presents the analysis of user test results and discusses observations. Chapter 5 concludes this paper with a discussion on study findings, limitations, and future work.

Chapter 2

Review of Existing Literature

2.0.1 Overview

This chapter begins with a description of decision support systems (DSS) and the use of AI in automating decision processes. It discusses the need for human involvement in decision making, the issue of low user trust in AI decision support, and XAI as an approach to improve trust. It then describes the unintended consequence of blind-trust, highlights the importance of trust calibration, and looks at some of the human factors of mistrust and blind-trust. It moves on to discuss the probabilistic nature of AI DSS, the value of drawing on research on uncertainty communication, and reviews and compares 3 expressions of probability. It concludes with the study's hypotheses.

2.1 AI decision support

Riveiro et al. (2014) provide a real world example of a DSS used by air-traffic controllers to identify potential threats. The DSS presents an air-traffic controller, the decision maker, with information on all the objects that are being tracked at a given time. With this information the air-traffic controller needs to make the decision whether to report any of these objects as a possible threat. Making that decision is highly complex. Objects are evaluated based on multiple types of information. An object's identity data is compared to its origin, flight behaviour, adherence to flight regulations, sensor

readings, etc. And every object type has its own specific attributes. A fighter jet will show different behaviour patterns than a civilian aircraft for example. Based on these complex set of data, the air-traffic controller needs to make this decision for multiple objects in a limited amount of time. And the consequences of an incorrect decision are potentially very high.

The work by Shneiderman (2021) on Human-Centered AI, describes these type of systems that are often found in complex and consequential domains. They are used in industries such as medicine, finance, or defense, where the decision task is "poorly understood and complex with varying contexts of use".

These DS systems are increasingly adopting AI technologies to try and improve decision making processes (Zhang et al., 2020; McGuirl & Sarter, 2006). ML models, trained on historical data, support the human decision maker by providing insights on new data or recommendations on what actions to take. Riveiro et al. (2014) reference work that has been done to automate parts of the air-traffic controllers' task with the use of Bayesian Networks. Another examples of AI decision support in the aviation industry is described by McGuirl and Sarter (2006). A neural-net-based DSS informs pilots about potential ice buildup on the aircraft during a flight, and helps them decide when it is necessary to delay a flight to perform in-flight deicing.

Fully automating these type of decisions is often not desirable. Their complex and highly consequential nature, combined with the probabilistic nature of ML predictions, means correct decisions cannot be guaranteed (Zhang et al., 2020). Full automation remains too risky (Bansal et al., 2021). Shneiderman (2021) describes the Reliable, Safe and Trustworthy system, which has both a high level of automation and a high level of human control. With AI's potential to be both highly beneficial and highly harmful, human involvement is required.

The aim is to combine the strengths of humans and AI in human+AI teams that perform better together than they would individually (Bansal et al., 2021). The decision maker relies on AI decision support to enhance their capabilities, and improves on full automation when AI underperforms (Chromik & Butz, 2021; Zhang et al., 2020). Chromik and Butz (2021) describe this as the "vision of man-computer symbiosis".

2.1.1 A lack of trust in AI decision support

AI decision support systems often suffer from low user trust and low adoption (Dodge et al., 2019; Gunning, 2019; Miller, 2019). A commonly described reason for low user trust is the lack of explanations given with recommendations, which leave decision makers unable to judge whether the recommendations are reliable, trustworthy, and should be acted on (Gunning, 2019; Riveiro et al., 2014). Gunning (2019) depicts a scenario in which a military intelligence analyst, is faced with the decision of whether to report the data insights recommended by an AI decision support for further investigation, without risking raising a false alarm.

A lack of explanations does not seem to be a problem unique to AI systems. Riveiro et al. (2014) describe how decision support systems often give solutions without an explanation or qualification. This, they point out, forces the decision maker to either fully accept the advice or go through the entire decision-making process themselves.

The black-box nature of many AI system only compounds this problem. As Dodge et al. (2019) says, high performance ML algorithms are often "unintelligible even for experts". Chromik, Eiband, Buchner, Krüger, and Butz (2021) describe how people can see a models' inputs and outputs but often have difficulties understanding how they are related. So whilst advances in the field of ML are promising to create systems which "perceive, learn, decide, and act on their own", these systems are incapable of providing explanations for their decisions to human users (Gunning, 2019). For the human decision maker, this lack of explanation is problematic, as they are ultimately responsible for the decision that is made (Zhang et al., 2020). They need to be able to judge whether to accept or reject AI recommendations.

This lack of transparency goes against the principles of Reliable, Safe and Trustworthy systems, which should "support human responsibility" with a high level of automation and a high level of human control (Shneiderman, 2021). It has also been pointed out that it violates many human-computer interactions (HCI) principles, like error correction and predictability (Eiband, Buschek, Kremer, & Hussmann, 2019). Explanations are seen as essential if users are to appropriately use, trust and manage these systems (Gunning, 2019), and if we are to realise the vision of creating highly ef-

fective human+AI teams (Buçinca et al., 2021). As Chromik and Butz (2021) describe it, "explanations are a crucial component for effective cooperation".

This need for more transparency has driven a call for more explainable artificial intelligence (XAI) (Dodge et al., 2019; Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018) and has resulted in a resurgence of research in the field.

2.2 Improving trust with XAI

The recent resurgence in XAI, as Miller (2019) puts it, is driven by the theory that people will more likely trust systems that exhibit transparency, that allow users to interpret their behaviour, and that are able to explain their actions and decisions.

The interest in developing understandable intelligent systems is not new (Liao, Gruen, & Miller, 2020; Miller, 2019). In the 1970s, research focused on making expert systems more understandable. This focus then shifted to neural networks in the 1980s and recommender systems in the 2000s. The current wave of research is focused on making today's increasingly complex and non-linear ML algorithms more understandable (Abdul et al., 2018). Gunning (2019) describes a tension between the performance of machine learning models and their explainability. Models that achieve higher performance, measured by metrics like prediction accuracy, tend to be less explainable. These models often employ newer learning techniques such as random forest, reinforcement learning, or deep learning. On the other hand, models that employ more explainable learning techniques, such as decision trees, often have lower accuracy levels.

XAI serves various use cases beyond explaining automated decisions and recommendations to DSS operators. These include autonomous vehicles justifying their actions to their operators (Gunning, 2019), helping developers and researchers to debug and improve ML models (Miller, 2019), assuring stakeholders that AI systems meet requirements, and providing explanations to individuals who have been affected by system behaviour (Vilone & Longo, 2021). Dodge et al. (2019) for example conducted an empirical study on the impact of XAI on people's judgments of model

fairness. They argue that explanations are essential for developers, users, and the general public to ensure fairness in ML systems.

The 7 goals of AI explanations proposed by Tintarev (2007) and referenced by Chromik and Butz (2021) and Balog and Radlinski (2020) offer a good understanding of the intended purposes of XAI features. The list of goals consists of transparency, scrutability, trustworthiness, persuasiveness, effectiveness, efficiency, and satisfaction. Transparency aims to provide users with answers regarding the system's functioning, while scrutability allows users to question and correct the system. Trustworthiness aims to enhance user confidence, while effectiveness and efficiency focus on assisting users in making better and faster decisions, respectively. Lastly, satisfaction aims to improve the overall usability of the system.

2.2.1 Approaches to XAI

Gunning (2019) identified 3 challenges in the development of XAI, creating more explainable models, designing explanation interfaces, and understanding the psychology of explanations. The 1^{st} challenge, creating more explainable models, speaks to the technical aspect of XAI.

Approaches to creating more explainable models have been categorised by Vilone and Longo (2021) into two types: trace-based and reconstructive. Trace-based approaches attempt to give a transparent view of the reasoning process of the predicting model. These approaches involve tracing the logic of how the model arrived at a particular prediction or highlighting the features or inputs that influenced its decision. For examples, the attention mechanism, a technique developed to improve the performance of deep neural networks, can also be used to visualize their internal workings (Parra et al., 2019). Inspired by how our visual systems work, the technique can be used to show specific input features that a model focuses on. By visualizing the areas of an image or document that the model focuses on during classification, users gain insights into the features that have had the most influence on its decision.

Reconstructive approaches, on the other hand, use a secondary model to infer the reasoning process of the predicting model. Instead of tracing a predicting model's

logic, these approaches reconstruct the decision-making process using a separate, more understandable model that mimics the behaviour of the predicting model. This can be done, for example, by training a decision tree-based model to approximate the logic of a neural network-based model. These surrogate models allow people to indirectly interrogate the original black-box models (Jesus et al., 2021).

Fidelity and robustness are two technical measures of evaluating XAI. Robustness measures how consistent explanations are when similar inputs are given, ensuring that similar explanations are provided for similar examples. Fidelity, also referred to as faithfulness, measures how well a reconstructive approach mimics the behaviour of the explained model, ensuring that the explanations accurately represent the reasoning process of the original model (Jesus et al., 2021; Chromik et al., 2021; Liao et al., 2020).

AI explanations are commonly categorised into two types, global explanations and local explanations. (Liao et al., 2020; Chromik et al., 2021; Dodge et al., 2019; Balog & Radlinski, 2020) Global explanations give an overview of "how a system works" (Dodge et al., 2019), its overall behaviour, and logic. These explanations remain consistent for each individual predictions. They allow users to assess the system's overall capabilities, strengths, and weaknesses. Local explanations on the other hand are specific to individual model outputs. They provide insights into the reasoning process and logic for a single prediction, and can change between each one.

2.2.2 XAI as a HCI problem

The last 2 XAI challenges identified by Gunning (2019), designing explanation interfaces and understanding the psychology of explanations, speak to the human factor of the problem. Miller (2019) has described XAI as a human-agent interactions problem, sitting at the intersection of machine learning, social psychology, and human-computer interaction (HCI). There is a recognition for the need of HCI practices in the field of XAI research (Vilone & Longo, 2021; Abdul et al., 2018; Liao et al., 2020).

Dodge et al. (2019) highlight the need for user-friendly explanations in order for people to confidently rely on AI systems. Whilst there are numerous technical XAI

methods, they are often impractical, difficult to use, and therefore ineffective in real-world scenarios (Abdul et al., 2018). To illustrate the point, Liao et al. (2020) question whether popular approaches like listing out influential features in a prediction would satisfy the explanation needs of a doctor. They state that the "effectiveness of an explanation is relative to the recipient", and argue for the importance of human-centered evaluation approaches.

Researchers have developed frameworks and guidelines that draw on HCI methodologies. Holzinger, Carrington, and Müller (2020) proposed the System Causability Scale. This subjective evaluation framework is specifically designed for AI explanations, and was based on the widely accepted System Usability Scale, a common method of measuring application usability in the field of HCI (Brooke, 1995). Chromik and Butz (2021) mapped the 7 explanatory goals proposed by Tintarev (2007) to 7 concepts of interaction proposed by Hornbæk and Oulasvirta (2017). Their aim was to characterize different interaction concepts in XAI and define design principles for interactive explanation interfaces (XUI).

Human-centered evaluative studies are common in the XAI literature. For example, Buçinca et al. (2021) conducted user tests to evaluate the effectiveness of forcing functions in calibrating user trust in a nutritional application. Similarly, Eiband et al. (2019) compared the trust and satisfaction of people using a nutritional application when presented with real explanations versus non-informative *placebic* explanations. Studies by both Bansal et al. (2021) and Zhang et al. (2020) examined the impact of local explanations and confidence scores on human+AI team performance through user tests in decision-making scenarios, whilst Jesus et al. (2021) evaluated 3 popular post-hoc XAI technologies using a human-in-the-loop approach.

2.3 Blind trust in AI decision support with XAI

Providing explanations is also not without its unintended consequences. Human cognitive and social processes influence how users interact with XUI, and how they affect their decision making (Eiband et al., 2019). A common observation in XAI human-

centered studies is a tendency of participants to over-rely on recommendations when explanations are given. This presents a new problem of users blindly trusting AI decision support and not identifying when a model prediction is incorrect (Buçinca et al., 2021; Bansal et al., 2021; Jesus et al., 2021). The human-centred evaluation of common XAI technologies by (Jesus et al., 2021), observed that participant accuracy was highest in conditions without explanations. Whilst decision time increased, their performance based on accuracy and recall fell. The study concludes that there is a trade-off between decision effectiveness and efficiency.

This user behaviour is also not unique to AI systems. Automation bias, which refers to the tendency of DSS users to rely on automated cues rather actively processing information, has been recognized as a user behaviour since the 1990s (McGuirl & Sarter, 2006) . This behaviour as McGuirl and Sarter (2006) point out is a "well adapted response to highly reliable system", but becomes problematic when there is a mismatch between the users' perception of the system's capabilities and its actual performance.

User over-reliance does not fit with the vision of creating human+AI teams that perform better together than either individually. Lack-of-trust and blind-trust are opposing but equally problematic challenges to realising these cooperative teams. After all, the objective should be for humans to maintain a level of control and responsibility (Shneiderman, 2021). As Dubiel et al. (2022) caution, a "misguided reliance" on AI systems may in fact lead to a "loss of agency". The desire is for system users to identify incorrect or improbable AI predictions (Zhang et al., 2020), and improve overall performance.

AI DSS should balance these opposing effects of mistrust and blind trust. They should help users gauge when to trust a recommendation and when to be critical of it (Bansal et al., 2021). For that reason there has been a call for more research in trust-calibration, developing methods that reduce user trust when appropriate (Dubiel et al., 2022; Bansal et al., 2021; Buçinca et al., 2021).

2.4 Human factors of trust and blind-trust

The literature provides a list theories from human psychology and decision science that might play a part in how decision makers perform using these AI DSS.

The dual-process theory is a common explanation for over-reliance. It proposes that humans employ either slower, more effortful analytical thinking or faster heuristic thinking (Joslyn & LeClerc, 2013), and that we more frequently use the latter (Buçinca et al., 2020). While AI explanations are likely designed with the assumption that users will engage analytically with them, it appears that users are more inclined to develop heuristics for determining when to trust the AI (Buçinca et al., 2021). Bansal et al. (2021) observed this in their qualitative analysis of human+AI team performance. Participants described developing a mental model to determine when to trust AI support. They established a threshold for the AI's confidence score, below which they disregarded its recommendations. Although heuristic thinking is efficient, it can also lead to "systematic and predictable errors" (Buçinca et al., 2021).

It is interesting to compare the dual process theory to the formal decision-making strategies described by Riveiro et al. (2014) in their study on target identification. Here we also find 2 similar approaches to decision-making: analytical and naturalistic. Analytical decision making involves weighing options and considering pros and cons, which is the formal strategy taught to air-traffic controllers. However, in practice, the naturalistic strategy is more commonly used due to limited information or time constraints. This strategy relies in part on the decision-maker's past experience of similar situations. Riveiro et al. (2014) describe the tendency of air-traffic controllers to rely too heavily on past experiences, and argue that a decision-maker's level of experience and the complexity of the decision task will impact their decision-making process. It is important to take into account a user's own domain experience when trying to determine how they will use AI DS systems.

XAI researchers have suggested that people can have similar levels of trust when provided with explanations, regardless of whether the explanations offer genuine informational justifications. Langer, Blank, and Chanowitz (1978) conducted a study

based on an interesting finding in social psychology, which suggested that individuals are more likely to agree to a request when given reasons, even if those reasons do not provide any meaningful information.

Langer et al. (1978) studied whether different ways of asking would change people's willingness to allow someone to go ahead of them at a photocopying machine. People approaching a Xerox machine at a New York university were asked in 1 of 3 ways: "Excuse me, I have 5 pages. May I use the Xerox machine?", "Excuse me, I have 5 pages. May I use the Xerox machine, because I have to make copies?", or "Excuse me, I have 5 pages. May I use the Xerox machine, because I'm in a rush?". Only the 3^{rd} request provided any meaningful justification. The study found that in the last 2 conditions, people were more and equally likely to comply with the request, regardless of whether the explanation provided was *placebic* or conveyed real information.

Based on these findings, Eiband et al. (2019) conducted an experiment to test how different types of explanations affect user trust in the context of a nutritional application. Similar to the previous study, they compared 3 conditions: no explanations, *placebic* explanations, and real explanations. They concluded that "placebic explanations can elicit similar levels of perceived trust as real explanations". This has obvious implications on the issue of XAI and blind-trust.

Similarly, Bansal et al. (2021) reference the Truth-Default Theory to explain their qualitative findings of how participants used AI recommendations. According to this theory, individuals have a natural inclination to assume that a speaker is telling the truth unless there is sufficient evidence suggesting otherwise. Bansal et al. (2021) argue that participants' use of model confidence as a threshold can be explained by this theory. When confidence scores are high, participants tend to trust the recommendations as they assume truthfulness. Low scores on the other hand, encouraged participants to abandon their truth-default behaviour.

The influence of personal values and beliefs on trust is another obvious factor to consider. The study by Dodge et al. (2019) on the effect of XAI on perceptions of model fairness, found that participants' preexisting views of ML significantly influenced their judgments. A similar observation was made by Knoblauch et al. (2018) in their study

on communicating risks related to natural resource extraction. Despite presenting identical risk information, respondents' perception of risk was significantly different between deep geothermal energy and shale gas. Although some of these studies lie outside the field of XAI, they shed light on the decision-making processes and the role of trust in various contexts.

The challenges in designing effective AI DS systems are highlighted by this non-exhaustive list of human factors. There is in fact mounting evidence that, although human+AI teams can outperform unassisted humans, better performance can often be achieved by the AI alone (Bansal et al., 2021; Buçinca et al., 2021, 2020). This measure of performance does not account for all the reasons why it is necessary to involve people in a decision process. As discussed earlier, accuracy is not the only consideration in the Human-Centered AI framework (Shneiderman, 2021). However, these observations do challenge the feasibility of creating highly effective human+AI teams that outperform both individually.

Chromik et al. (2021) also challenge the degree to which XAI will be able help people understand how these systems work. They studied the effect of the Illusion of Explanatory Depth on XAI, a theory that describes people's tendency to over-estimate their understanding of complex systems. They argue that humans are unlikely to every be able to "correctly predict the behaviour of complex non-linear ML models".

Some work has been done in developing interactions to help address these human-factors. For instance, Buçinca et al. (2021) studied the use of forcing functions in XAI interfaces as a way to mitigate heuristic thinking in users. They tested 3 approaches: delaying AI recommendations, making AI recommendations optional, and only providing a recommendation after participants had made their own decision. Similar trust calibration methods were recommended by Dubiel et al. (2022): allowing users to enable or disable recommendations, and for recommendations to foster user reflection and encourage them to consider alternatives.

Buçinca et al. (2021) saw a significant reduction in over-reliance with the use of forcing functions. However, the study also observed a negative correlation between participants' task performance and their subjective rating. The conditions in which

participants performed best were rated lower in terms of trust and preference. They concluded that there is a "trade-off between subjective trust and preference in a system and performance with the system". This view aligns with the arguments made by Buçinca et al. (2020) that user preference does not necessarily predict decision-making performance.

More work is needed to develop explanation interfaces that appropriately manage user trust, encourage critical thinking, and ensure user satisfaction (Bansal et al., 2021; Buçinca et al., 2021).

2.5 State of the art in XAI research

Gunning (2019) gives 3 user-centered approaches to measuring the effectiveness of XAI, user satisfaction, users' mental model of an AI, and task performance. However, the challenges posed by human factors should serve as a reminder not to overly rely on subjective measures of satisfaction.

Gunning (2019) does recognise the importance of evaluating XAI systems by how well they help system users, including task performance. Jesus et al. (2021) also argue for the need to objectively measure how users perform using these systems. Buçinca et al. (2020) even claim that evaluations with proxy tasks or subjective measurements are misleading, and that by not evaluating XAI systems by measuring performance on actual decision-making tasks, the field may be slowing the progress toward realising human+AI teams. Their study showed how proxy tasks, ones that force study participants to engage with AI assistance and explanations, don't necessarily predict results of real tasks. They show the importance of using decision-task scenarios in which participants can choose whether and how much to use AI decision support.

User-centered XAI studies commonly use task-based approaches where study participants make decisions based on some given information and AI recommendations. Various types of decision scenarios have been used, and are often designed to be suitable for participants with general backgrounds. Scenarios include text classification tasks such as labeling the sentiment of beer reviews (Bansal et al., 2021), making

meal decisions using a nutritional app (Buçinca et al., 2020; Eiband et al., 2019), and financial scenarios, like evaluating loan applications (Chromik et al., 2021; Zhang et al., 2020). Real-world scenarios are less common in XAI studies. It is recognised that studying real-world decision tasks is costly in terms of both time and money (Jesus et al., 2021). Jesus et al. (2021) is an example of a study of real-world tasks using real end-users. They evaluated popular XAI technologies in the domain of fraud detection, with real fraud analysts.

Task-based studies in XAI employ a range of approaches, including moderated and unmoderated user tests, sometimes taking a mixed approach. Unmoderated user tests involve participants completing tests independently, using a web application, in an uncontrolled environment. These tests are often conducted on crowd-worker platforms like Amazon Mechanical Turk, allowing for larger participant sample sizes. For instance, a study by Bansal et al. (2021) included 3 different tasks and 4 conditions for each, with approximately 100 participants per condition. Similarly, a study by Buçinca et al. (2020) conducted 2 experiments with 3 conditions each, involving approximately 300 participants. Not all studies have such large participant samples. For example, Zhang et al. (2020) tested 8 conditions with only 9 participants in each.

Other studies in XAI have used moderated methods, where researchers are present during the testing process. Some have used a mix of both. Moderated tests are often more time-consuming, and can result in smaller participant numbers sizes. For example, Jesus et al. (2021) included only 3 participants in their study. This was also due to their use of a real-world fraud detection task, involving actual fraud analysts. As a mixed method example, Chromik et al. (2021) conducted moderated tests with 40 participants and unmoderated tests with 107 crowd workers.

In task-based studies, the sample sizes of decision tasks are also important to consider, as well as the number of participants. For instance, although Jesus et al. (2021) included only 3 study participants, they collected a total of 300 decisions across those participants. Zhang et al. (2020) had participants complete 40 decision tasks, while Bansal et al. (2021) included 50 tasks for each participant to complete.

To evaluate the performance of combined human+AI teams, it is necessary to

include incorrect AI decision support examples in the decision tasks and motivate participants to improve upon them. In these studies, model performance is often reduced or matched to that of an unassisted decision maker (Bansal et al., 2021; Zhang et al., 2020; Buçinca et al., 2020). To incentivize participants to perform well, a bonus and penalty system is commonly used in crowd-sourced participant studies. Participants are rewarded for making correct decisions, penalized for incorrect ones, or offered bonuses for achieving a certain level of accuracy (Zhang et al., 2020; Bansal et al., 2021; Buçinca et al., 2021).

In these studies, decision task performance, trust, and over-reliance are primarily measured quantitatively. Decision performance is measured by comparing participants' final decisions to a ground truth, to calculate accuracy, recall, and false positive rates (Jesus et al., 2021; Bansal et al., 2021; Buçinca et al., 2021). Decision time is another common metric of performance (Stoll, Urban, Ballin, & Kammer, 2022; Jesus et al., 2021). As a measure of user trust or over-reliance, the agreement rate between AI recommendations and participants' final decisions have been used (Buçinca et al., 2021; Zhang et al., 2020; Jesus et al., 2021).

Most of these studies include a post-task subjective evaluations, in which participants provide qualitative feedback or quantitative ratings of the application, the AI support, or AI explanations (Stoll et al., 2022; Bansal et al., 2021; Dodge et al., 2019; Jesus et al., 2021). Quantitative evaluations are used to rate preferences (Stoll et al., 2022; Buçinca et al., 2021), while qualitative approaches can provide additional insights into participants' perceptions and decision-making tactics (Bansal et al., 2021).

2.6 Communicating uncertainty in AI DSS

ML-based AI is commonly described as probabilistic in nature. Holzinger et al. (2020) differentiate between scientific models and ML models. Scientific models typically aim to describe causation, whereas ML models primarily rely on concepts of correlation, similarity, or distance.

Confidence scores are a commonly used example of the probabilistic nature of ML

(Bansal et al., 2021; McGuirl & Sarter, 2006; Zhang et al., 2020). Calculated by a model's performance during training, confidence scores represent the probability of a model prediction being correct (Bansal et al., 2021). A higher confidence score indicates a higher likelihood of a correct prediction, a lower score suggests a higher probability of an error.

Performance information helps people to develop an understanding of a model's error boundaries. Performance information can be given at a global level to show overall system performance, or at a local level, specific to individual predictions. Confidence scores are an example of local performance information that helps users gauge performance and reliability on a case-by-case basis.

The classification of confidence scores as an XAI method is questionable as they do not meet the definition of XAI provided by Miller (2019), as "an explanatory agent revealing underlying causes to its or another agent's decision making." Confidence scores do not describe model reasoning processes or logic. Liao et al. (2020) only includes confidence scores within the broader scope of XAI. They consider them alongside model input and output data as descriptive information that contributes to the goal of enhancing model transparency.

Confidence scores do align with some of the 7 goals of XAI discussed earlier, namely enhancing user confidence (trustworthiness) and improving decision making effectiveness and efficiency. As an industry example, Google's guidance for designing with AI proposes confidence scores as a readily-available alternative to describing how an AI came to a decision. And they have been show to be effective in AI decision support.

Studies by Zhang et al. (2020) and Bansal et al. (2021) explored the use of confidence scores and local feature-based explanations as trust calibration mechanisms in AI-assisted decision making. Their findings showed the effectiveness of confidence scores in managing user trust, while local feature-based explanations were shown not to provide additional benefits. These studies suggest that confidence scores are a promising tool for balancing user trust and critical thinking in AI DS systems.

In their study on a flight DSS, McGuirl and Sarter (2006) found that continuously showing confidence scores effectively calibrated trust and reduced automation bias in

pilots. By comparing the presentation of DSS recommendations alone to the presentation of recommendations with a visualization of model confidence over time, they observed a significant improvement in trust calibration without any noticeable decline in task performance.

Liao et al. (2020) gave a contrasting perspective on confidence scores based on interviews with UX designers at IBM. They found that performance type explanations, including confidence scores, were consistently ranked lowest amongst designers. Designers described them as lacking actionable information and being unnatural in their communication style. Despite these reservations, confidence scores have shown to be effective in various studies. In fact, Bansal et al. (2021) suggest that, in light of their findings that local explanations did not improve on simply showing confidence scores, new XAI methods might be developed that work in tandem with confidence scores.

Given the probabilistic nature of ML, drawing on research in uncertainty communication would seem valuable. The knowledge from this area should help in effectively conveying probabilistic information, including confidence scores, with the aim of improving human+AI teams performance and trust calibration.

Abdul et al. (2018) mapped the XAI research landscape to assess the interconnectedness of various research communities and to identify trends and opportunities of HCI research in XAI. They showed that the topic of uncertainty appears to be relatively isolated within two prominent research communities, namely Intelligent and Ambient Systems, and Psychology of Explanations and Causality. This indicates a research gap in exploring uncertainty within the context of XAI.

When it comes to displaying confidence scores to users in XAI studies, different methods have been used. For instance, McGuirl and Sarter (2006) presented confidence trends over time using a bar graph, while Bansal et al. (2021) conveyed confidence as a percentage value within a sentence. Zhang et al. (2020) also incorporated confidence within a sentence but expressed it as a natural frequency. There is little justification given for the chosen methods to convey confidence to the user, and none on how these methods align with the goals of XAI.

This raises the questions: What is the effect of different methods for expressing ML

prediction confidence in AI decision support systems on the performance of human+AI teams? To what extent do different methods for expressing ML prediction confidence effectively calibrate decision makers' trust in AI decision support recommendations?

Research question 1

What is the effect of different methods for expressing ML prediction confidence in AI decision support systems on the performance of human+AI teams?

Research question 2

To what extent do different methods for expressing ML prediction confidence effectively calibrate decision makers' trust in AI decision support recommendations?

2.7 Expressions of probability

To reference Teigen (2022), "past facts and future outcomes are rarely known exactly", there is an uncertainty to them. These uncertainties need to be expressed in "tentative or approximate of ways" that communicate the inexact nature of these statements. Whilst uncertainty has been described as unjustifiable in an economic context, it is measured by the statistical means of probability (Knoblauch et al., 2018). Aleatory probability takes an objective stance, and quantifies the probability of an event based on empirically observed occurrences. Epistemic or Bayesian probability, provides a subjective measure, and is a function of the assessors subjective knowledge and beliefs (Knoblauch et al., 2018).

Other qualifications of uncertainty have been used that expand on this. The Intergovernmental Panel on Climate Change (IPCC) for example has a framework for quantifying levels of confidence in what they communicate. Confidence in this framework is a function of both probability and the quantity, quality, and variance of evidence (Bradley, Helgeson, & Hill, 2017). Spiegelhalter et al. (2011) provide an easy working definition of probability as "betting odds constructed from knowledge and information".

Communicating probability and uncertainty has a long and rich history. It is difficult to do effectively, in particular when communicating to a lay audience (Spiegelhalter et al., 2011), and it is difficult to know how well people incorporate uncertainty information in their decision making process (Joslyn & LeClerc, 2013; Knoblauch et al., 2018). And the objective of communicating is not always the same. The communication may aim to simply inform people, to change their behaviour, to give detailed information, or communicate its essence (Spiegelhalter et al., 2011). The way it is expressed is critical to how effective it can be. It needs to be tailored to suit both the decision task and the user's capabilities (Joslyn & LeClerc, 2013). Expressing probabilities in percentages (10%) for example, is considered too abstract for many people, whilst changing to natural frequencies (1 in 10) is more commonly understood. And, echoing the tension between user preferences and user performance described by Buçinca et al. (2021), Spiegelhalter et al. (2011) reminds us that the most effective forms of communication may not be the one readers like.

Probability and uncertainty can be expressed in 3 ways, verbally, numerically, or visually (Jaffe-Katz et al., 1989; Spiegelhalter et al., 2011). The IPCC categorise confidence levels as verbal expressions; very low, low, medium, high, and very high (Bradley et al., 2017). Bansal et al. (2021) display their ML model's prediction confidence as a numerical percentage point. Kay, Kola, Hullman, and Munson (2016) use density plots and dot plots to visualise the probabilities of bus arrival times, and McGuirl and Sarter (2006) display model confidence over time as a line-graph. Each of these type of expressions are topics of research in their own right, and are often comparatively evaluated on their effectiveness.

2.7.1 Numerical and verbal expressions of probability

Verbal and numerical forms of communications have a long been discussed and compared. An argument for either format can be found in the literature. Arguments can be found in the literature for using verbal (categorical) forms, due their natural quality and richness of information. Opposing arguments for using numerical formats can also be found, due to their precise nature. Finding definitive guidance on how best to

communicate uncertainty can be difficult, and it has been described as being "more of an art than a science", that relies more on a designer's intuition than scientific principles (Spiegelhalter et al., 2011; Jaffe-Katz et al., 1989).

The subject is complex. Like all forms of communication, it is subject to the speaker's intent, reader's perception, and the preferences of both. As Teigen (2022) points out, "words and perhaps number... are never neutral", they often convey additional, implied information. The communicator may, for example, choose to present a particular perspective on the information, referred to as framing. When being show natural frequencies or fractions, a reader will often perceive the same probability differently, depending on the value of the numerator. 1/10 is often read as less probable than 10/1000 (Spiegelhalter et al., 2011). The communication paradox, described by Erev and Cohen (1990) states that the communication style preferred by the communicator and recipient are not necessarily the same. This is echoed in a study that Jaffe-Katz et al. (1989) reference, that found that people often prefer to express uncertainty verbally, whilst preferring to receive it numerically.

Numerical expressions

Describing probability with numbers is succinct and accurate (Spiegelhalter et al., 2011). Numbers are precise forms of communication and are easy to compare. The work by Jaffe-Katz et al. (1989) studied the cognitive processes involved in comparisons of numerical and nonnumerical expressions of uncertainty. They observed that participants, when asked to choose the higher or lower of 2 values, were consistently faster when comparing numerical expressions than comparing nonnumerical forms. This could not be attributed to the cognitive difference of reading digits or words, because numerical names were compared to verbal categories of probability, i.e. "five%" versus "improbable". This has been explained as the result of the reader assigning a numerical value to a category themselves, when one isn't provided (Knoblauch et al., 2018). The reader spends effort sampling values within an implied value range, in an attempt to resolve the vagueness of the category (Budescu et al., 1988).

The literature provides overwhelming evidence of the great variability and over-

lap of the values people assign to probability words. And between-subject variability far exceeds within-subject variability (Budescu et al., 1988; Jaffe-Katz et al., 1989). Numerical expressions of probability result in less misunderstanding than its verbal counterpart (Joslyn & LeClerc, 2013). In the study by Budescu et al. (1988) participants were asked to bid or rate lotteries of different probabilities. Participants were observed to attach more extreme values to lotteries with verbal probabilities than numerical ones.

The vagueness of probability terms have been measured by Wallsten, Budescu, Rapoport, Zwick, and Forsyth (1986). Terms such as doubtful, chance, possible, and good chance were represented by a membership function on a 0 to 1 probability interval. These membership functions, measures of a term's membership of a particular category, show the range of numerical values that people assign to verbal probability expressions, as well as the overlap between them (Wallsten et al., 1986).

Studies have compared numerical and verbal probability expressions on decision making. The study by Budescu et al. (1988) showed that, based on a profit-loss evaluation of participants' lottery bids, numerical probabilities were significantly superior to verbal ones. Joslyn and LeClerc (2013) evaluated and compared the weather-related decisions students made under the 2 expressions of probability. Participants decided whether or not to salt roads, based on a given likelihood of overnight freezing, and a budget-penalty analysis. Their decisions were evaluated against an economically rational model that states, based on the cost of salting and the potential penalty of not salting if it freezes, to salt at or above a 17% probability of freezing. The results showed participants shown numerical probabilities performed significantly closer to the rational model.

To summarise, numbers are considered overall more reliable, precise, and consistent (Jaffe-Katz et al., 1989). And as Spiegelhalter et al. (2011) reminds us, they also avoid any literacy or language barriers.

Verbal expressions

There is general agreement that verbal expressions of probability are still the preferred form of people communicating uncertainty, including experts (Wallsten et al., 1986; Budescu et al., 1988). In the study by Erev and Cohen (1990), that described the communication paradox referred to above, sportswriters and broadcasters were asked to give a probability of upcoming basketball games. Most were observed to express their opinions verbally. Experts often believe that people "should not be burdened with precise numerical estimates" (Joslyn & LeClerc, 2013).

Wallsten et al. (1986) reference Zimmer (1984) for a possible explanation. Zimmer (1984) points out that concepts of probability were not formalised until the 17^{th} century. However, expressions of uncertainty existed in language long before that. It is rules of conversations by which people handle these expressions, not numbers. This perspective is interesting when considering the argument by Miller (2019), that people will expect AI explanations to adhere to frameworks used to describe human explanations.

The probability of basketball games made by the sportswriters and broadcasters, in the study by Erev and Cohen (1990), were given to students to decide the attractiveness of the gamble. Whilst the students were shown to prefer receiving numerical probabilities, this study concluded that there was no difference in efficiency between numerical and verbal forms.

Whilst verbal expressions are vague, they can convey a much richer amount of information. Teigen (2022) gives a strong argument for the benefits of verbal probability expressions. They can convey additional information that numbers do not. They can imply the source and limitations of the information, the speaker's credibility, attitude and intentions, and speak to the valence and severity of an outcome.

The source of knowledge can be implied in a verbal expression, indicating whether the probability statement takes a subjective/epistemic or objective/aleatory stance. Epistemic verbs such as *believe* and *doubt* describe a subjective evaluation, whilst auxiliary verbs like *will* and *could* describe a more objective one. As Teigen (2022) illustrates, "it (not I) can happen, I (not it) believe it will".

Words can indicate how desirable a probability outcome is, which Teigen (2022) refers to as the expression's valence. Terms like *risk* are a negative evaluation of an outcome, whilst *chance* or *hope* suggests it to be desirable. Other terms, such as *certain* and *likely*, can be used for both.

Verbal expressions have also been described as having *directionality*, the ability to point either towards the occurrence or non-occurrence of a probable event. As Teigen (2022) describes, probability statements have a double meaning, that of an event occurring or not. Here again words can convey more information than numbers, as they can point towards either of these outcomes. When an event is described as *possible*, *likely*, or *almost certain*, the statement points "upwards" to the event's occurrence. A statement that describes an event as *uncertain*, *doubtful*, or *not completely certain* points "downwards" to its non-occurrence.

The direction of a statement can influence a reader's decisions. Teigen (2022) references a study that asked participants whether they would recommend the use of a new and controversial migraine treatment to a patient. In one condition the treatment was described as having "some possibility" of being helpful, in the other as "quite uncertain" to be. With the assumption that both expressions fell within a probability range of 30-35%, the results were significant. Nearly all participants said they would recommend the treatment in the 1^{st} condition. That dropped to only 1 in 3 participants in the 2^{nd}. A 3^{rd} group was shown a numerical probability, and the responses were more evenly split with 58% recommending the treatment and 42% not.

Teigen (2022) argues that when the objective is to convey recommendation or warnings, verbal expressions are less ambiguous than numbers. Implied recommendations are more clear when expressed verbally than numerically. If the aim of the uncertainty communication, as Spiegelhalter et al. (2011) describes it, is to change people's behaviour, then verbal expressions seem more effective.

To summarise, verbal expressions are thought to appeal to people's intuitions and emotions (Spiegelhalter et al., 2011). They have the ability to convey a richer amount of information, and provide clearer guidance and recommendations. They can also help when communicating to an audience with low numeracy (Spiegelhalter et al.,

2011), and have been recommended in particular when dealing with small probability values (Knoblauch et al., 2018).

Verbal-numerical

Recommendations have been made to use both forms as verbal-numerical probabilities expressions (Teigen, 2022; Knoblauch et al., 2018). The study by Knoblauch et al. (2018), referenced earlier, compared the perceived risks of natural resource extraction when communicated either verbally or verbal-numerically. Participants rated the combined format highest, as being easier to understand, most exact, and most liked.

The two forms of probability expressions posses individual qualities, and offer their own benefits, and draw-backs. These qualities could be complementary, when combining the precision of numbers and the richness of words. Combined they may "lead to better understanding than either format taken in isolation" (Teigen, 2022).

2.8 State of the art in uncertainty communication

The survey conducted by Hullman, Qiao, Correll, Kale, and Kay (2019) provides a valuable insight into how uncertainty visualisations have been evaluated. The goal of evaluating uncertainty visualizations, as they describe it, is to test the effectiveness of different techniques in conveying the variability of a point estimate to readers and helping them to make informed decisions. They describe 3 types of evaluations; theoretical evaluations based on design principles, low-level visual evaluations, and task-oriented user studies. Similar to the field of XAI, arguments have been made for the need to evaluate uncertainty visualisations based on their effectiveness in supporting people in their tasks. They should be assessed on how well they help in realistic user tasks (Hullman et al., 2019).

Hullman et al. (2019) classified evaluative studies in terms of their aims, expected effects, evaluation goals, elicitation methods, and analysis methods. Their analysis shows both commonalities with what is found in XAI research, and novel approaches that might inform future work in XAI.

Considering task-oriented evaluations, the most common aim in these studies is to measure user performance in two key areas; reading and understanding information, and decision making (Hullman et al., 2019). Researchers most commonly focus on participants' accuracy in reading and understanding probability values, and determining the difference between the perceived and actual values. These can involve absolute or relative measures. Absolute measures ask participants to estimate probability values, while relative measures might involve ranking probabilities. Results are then compared against data-based values or rankings (Hullman, 2016). The study by Kay et al. (2016) is a good example of this, where dotplots and density plots were evaluated on participants' accuracy in judging the probability of bus arrivals. As another example, in the study by Galesic, Garcia-Retamero, and Gigerenzer (2009), the accuracy of people's understanding of medical risks were compared when presented as numerical values or using icon arrays.

Similar to task-based XAI evaluations, other studies in uncertainty visualizations also focus on the effect on decision-making and decision quality. These asses how visualisation methods affect participants' decision-making processes or the quality of decisions using a rational decision standard. Another study focus, which could be of interest in the field of XAI, is the effect on participants' confidence in making probability judgments or decisions (Hullman et al., 2019).

Performance has been measured in various ways (Hullman et al., 2019). Subjective measures include self-reported satisfaction and confidence. Both of these subjective measures can be seen in the studies by Riveiro et al. (2014) and Kay et al. (2016). Decisions have been measured by asking participants to make hypothetical choices based on given information, or to place a value on a probability. Decisions can also be measured by how close participants' decisions align with optimal decision under utility theory. The study by Joslyn and LeClerc (2013) is a good examples of these decision measures. Based on weather predictions and a given budget, participants repeatedly made the hypothetical decisions of whether or not to salt roads. Participants considered the predicted nighttime temperatures, the cost of choosing to salt roads, and the potential penalty of not salting and temperatures dropping below freezing.

Their performance was evaluated against an economically rational decision model.

Participant samples in uncertainty visualization studies are similar to those seen in XAI research. Task-oriented studies that involve real-world scenarios tend to have smaller sample sizes. For instance, in the study by McGuirl and Sarter (2006), 2 conditions were tested with 15 certified flight instructors in each condition. The study on target identification by Riveiro et al. (2014) simulated a real decision support system and included professional air-traffic controllers, with 11 participants in each condition. In contrast, online studies that are suitable for public participation tend to have higher participant numbers. Knoblauch et al. (2018) tested how different forms of risk communication influenced public perception of risk by conducting surveys with 49 respondents in each of their 12 conditions. In the study on bus arrival estimates by Kay et al. (2016), an online survey was conducted with 172 participants across 2 conditions.

As was seen in XAI studies, the size of decision samples is important in task-oriented studies. For example, McGuirl and Sarter (2006) involved 30 pilots, each completing 28 flights. In the study conducted by Riveiro et al. (2014), the target identification task comprised 119 objects that participants had to evaluate.

2.8.1 Study and Hypothesis

The discussion and comparison of verbal and numerical probability expressions suggests that they may impact decision performance differently. Numerical expressions, preferred by readers, may lead users to prefer confidence scores shown as numbers. Alternatively, users may prefer verbal expressions as they feel more natural.

How they effect decision performance also seems to be an open question. Task-oriented studies have come to different conclusions, and the decision tasks are often more mathematical in nature than those seen in XAI e.g. rating lotteries of different probabilities vs labeling the sentiment of beer reviews.

If verbal expressions are more influential on people's decisions compared to numbers, they may be more effective in encouraging analytical thinking when prediction confidence is low. Alternatively, if combining both forms leads to better outcomes

compared to using either alone, expressing model confidence verbal-numerically may result in the highest levels of user trust and decision performance. This study hypothesizes that the 3 different ways of expressing confidence scores will result in significant differences in human+AI performance and trust calibration.

Research hypothesis 1

The method of displaying AI recommendation confidence in AI decision support systems, whether numerically, verbally, or verbal-numerically, will have a significant impact on human+AI team performance, measured by final human decisions accuracy.

Research hypothesis 2

The method of displaying AI recommendation confidence in AI decision support systems, whether numerically, verbally, or verbal-numerically, will result in significantly different human+AI team performance, when dealing with examples where AI recommendations are incorrect, measured by final human decisions accuracy.

Chapter 3

Experiment design and methodology

3.1 Experiment overview

3.1.1 Experiment foundations

This experiment was based on the work done by Bansal et al. (2021). Their research asked how effective XAI is in enabling human + AI teams that perform better than either human or AI individually. This experiment followed their approach with some deviation, and used one of their data-sets.

Bansal et al. (2021) tested team performance on 2 types of tasks, 1 of which was text classification. The task involved labeling either beer or book reviews as positive or negative. Decision support was provided by a natural language processing (NLP) model that predicted the sentiment of every review. As participants labeled these reviews, the AI decision support gave its recommendations as to the sentiment, qualified by a confidence score.

This study used the beer data-set from that study, which they shared with the researcher. Of the 3 data-sets used in the original study, the beer data-set was chosen because it was shown that participants labeling these reviews benefited more from AI decision support.

During a pilot study, Bansal et al. (2021) measured the average task performance of unassisted people. They observed people to be about 87% accurate at labeling beer reviews without AI decision support. A RoBERTa-based text classification model from AllenNLP was then trained to match that accuracy. 50 beer reviews were classified at 84% accuracy, giving a set of 42 correctly classified and 8 misclassified reviews, with equal number of false positives and false negatives. Each prediction included a confidence score which was generated by RoBERTa with further post-hoc calibration done using an isotonic regression.

This study deviated from their method and study design to some degree. These differences will be noted throughout this chapter. The recruiting and reward method is the primary difference, and is the reason for many of the other deviations. Bansal et al. (2021) recruited crowd-workers through Amazon Mechanical Turk. Crowd-workers were compensated for their time, and were given an incentive to perform well using a bonus and penalty system. This study relied on voluntary participation, and an optional raffle was used as an incentive and reward for taking part. Certain methods were judged inappropriate in this context.

3.1.2 Experiment overview

This was an empirical, quantitative, human-grounded, task-based AI decision support evaluation. Data was collected during a series of unmoderated user tests. Study participants completed an independent test under 1 of 3 conditions, and in an uncontrolled environment. Quantitative data was collected during the labeling task, and participants' made a quantitative subjective evaluation of the AI decision support.

Volunteers were able to participate through a web-application using their own device and web-browser. Taking part comprised of 3 steps. 1^{st} step, after agreeing to participate, they completed a short demographic survey. 2^{nd} step, they completed the task of labeling either 25 or 50 beer reviews as positive or negative. A dummy AI assistant provided decision support. AI recommendations included a confidence score, which was expressed in 1 of 3 ways. 3^{rd} step, participants evaluated the AI decision support by rating 10 statements on a Likert scale.

The experiment design can be divided into the following topics, which will be described in this chapter.

1. User test design
2. User interface design
3. Application development
4. Recruitment

3.2 User test design

3.2.1 Test conditions

The experiment comprised of 3 conditions. It aimed to measure task performance when AI confidence was expressed either numerically (*Num.*), verbally (*Ver.)*, or verbal-numerically (*VNum.*). See figure 3.1.

The beer data-set included a confidence score with each review. In the 1^{st} condition this was displayed as a percentage value to one decimal point. e.g. 90.5%. This was in-line with how confidence was expressed by Bansal et al. (2021).

The verbal expressions were taken from NATO's approximate probability scale, referenced by Teigen (2022). The scale maps verbal statements to numeric probabilities. Various scales have been defined in different domains, including defense, medicine and climate research. The NATO scale was chosen for being the smallest set of 5 statements. In order of probability it is comprised of the terms; *highly likely*, *likely*, *even chance*, *unlikely*, and *highly unlikely*. The corresponding numerical values can be seen in table 3.1. In the 2^{nd} condition only these verbal terms were displayed, as they mapped to the data-set's confidence values.

In the 3^{rd} condition, the NATO scale and percentage value were shown together e.g. Highly likely (90.5%). The format followed the verbal-numerical expression examples given by Teigen (2022), where the numerical is shown in brackets after the verbal.

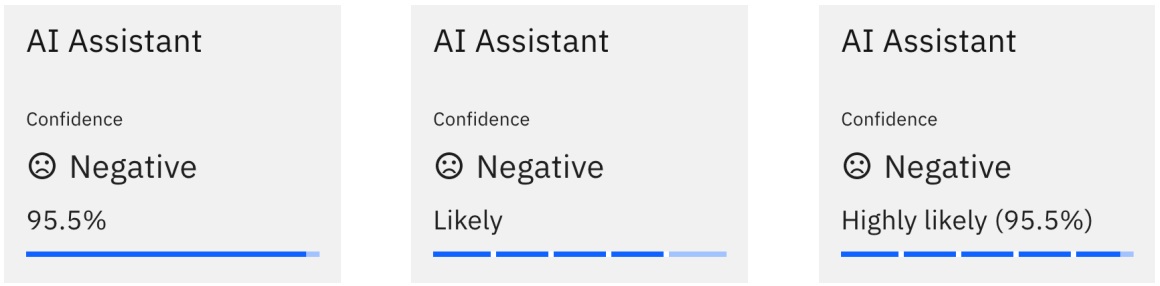
Numerical assessment	Verbal statement
>90%	Highly likely
60-90%	Likely
40-60%	Even chance
10-40%	Unlikely
<10%	Highly unlikely

Table 3.1: NATO(2016) approximate probability scale (Teigen, 2022)

In all 3 conditions, AI confidence for both positive and negative sentiments were shown. The confidence of P in the predicted class, and the confidence of 1 - P in the alternative. Showing both was necessary in the 2^{nd} conditions, when confidence was expressed only verbally and the AI assistant gave an *even chance* probability of a beer review being positive or negative. To control this variable, both confidence scores were shown in all cases.

Showing the confidence in both sentiments has also been suggested to reduce over-reliance on AI decision support. Whilst explanations were shown to increase reliance by Bansal et al. (2021), this tendency was less pronounced when explanations were given for both negative and positive. They presumed this may be because the user is encouraged to consider the alternative sentiments as well as the predicted. Showing confidence in both sentiments seemed appropriate when one of the research questions was related to trust calibration.

In the conditions where verbal expressions were included in the confidence scores, the predicted sentiment would be shown as being either *highly likely*, *likely*, or *even chance*. The alternative was shown as *highly unlikely*, *unlikely*, or also *even chance*. The predicted class was visually highlighted in all cases, except for an *even chance* confidence in the 2^{nd} condition. When both sentiments are shown to have an *even chance* likelihood, making an ultimate prediction contradicts the probability statement, and might cause confusion. This is different in the 3^{rd} condition, when an *even-chance* probability is further qualified by a percentage value. In this case, a



(a) Numerical (b) Verbal (c) Verbal-numerical

Figure 3.1: Expressions of confidence

definitive prediction can be made.

The local saliency-based explanations studied by Bansal et al. (2021) were not included in this experiment. This was decided for 2 reasons. 1^{st}, Bansal et al. (2021) concluded that they did not significantly improve performance over just showing AI confidence. 2^{nd}, local XAI methods were not part of this study's aim, and omitting them reduced the number of experiment variables.

The personification of the AI decision support was toned-down compared to the illustrated Marvin character used by Bansal et al. (2021). It was simply referred to as the AI Assistant. Guidance on how to mitigate over-trust in robots, referenced by Buçinca et al. (2021), suggests to avoid anthropomorphic features.

The buttons that participants used to classify the beer reviews were changed from being labeled *mostly positive* and *mostly negative* to *positive* and *negative*. The term *mostly* was thought to add ambiguity when combined with the verbal expressions of probability. Teigen (2022) describes how people use *probable*, *average*, and *most likely* to mean the same thing. *Mostly positive* might be interpreted as meaning *most likely positive*. To illustrate the point, if the AI assistant gave a recommendation that a beer review had an *even chance* of being *mostly positive* or *mostly negative*, it may not be immediately clear to a participant which was a statement of confidence. These were semantic questions which were outside the bounds of this study. Taking into consideration that the effectiveness of verbal expressions are subject to the readers' literacy (Spiegelhalter et al., 2011), simplifying the labels seemed appropriate.

3.2.2 Demographic survey

The test included a short demographic survey. It comprised of 5 closed-ended questions related to participants' age, gender, education, reading literacy, and whether they had a background in computer science or engineering.

Participants were asked to indicate their level of comfort reading English due to the nature of the task and the study's independent variables. Text classification requires reading comprehension, and verbal expressions of probability can be a barrier for people with literacy limitations (Spiegelhalter et al., 2011). Understanding how comfortable participants were in reading English was necessary for that reasons.

Participants indicated whether they had an educational and/or professional background in Computer Science or Engineering. This was asked because recruitment was going to rely, in part, on professional contacts in the software industry, and academic contacts within Technical University Dublin. This may have resulted in a high representation of people with those backgrounds. Responses to that questions would show the extent of that. Also, as discussed above, peoples' attitudes toward technology can influence their trust in AI systems (Knoblauch et al., 2018; Dodge et al., 2019). Participants' background in these technologies may have been a contributing factor.

Age, gender, and education is a common way to describe a study sample, and were included for that reason (Dodge et al., 2019; Zhang et al., 2020; Chromik et al., 2021). Age groups and education levels were based on those used by the Irish Central Statistics Office. Gender options followed guidance from the UK National Health Service.

3.2.3 Subjective evaluation

Participants were asked to give their subjective evaluation of the AI decision support after completing the labeling task. This is a common approach in user-centered evaluations (Jesus et al., 2021; Stoll et al., 2022). A post-task survey was also used by Bansal et al. (2021), albeit using different statements. A subjective evaluation was also used by Buçinca et al. (2021), allowing them to correlate participants' performance

and preferences.

After a considering a number of questionnaires from the literature, the System Usability Scale (SUS) was chosen. It is a widely accepted, used, and cited measure of system usability. It was developed and described by Brooke (1995) as a "quick and dirty", low cost and effective global usability assessment, suitable to use for a variety of products and services.

Usability is measured by 10 statements which respondents agree or disagree with on a 5 point Likert scale. Half of the statement scales are inverted, where a high score of 5 is a negative evaluation. To prevent response bias, the sequence of the 5 positively and 5 negatively scaled statements are alternated throughout the questionnaire. This forces the evaluator to read each statement and think about their response (Brooke, 1995).

The sum of the 10 scored statements is the System Usability score. It is not meant as an absolute measure but a comparative score, a way to rank and compare systems with same intended purpose e.g. competitors and predecessors. The SUS would allow the 3 experimental condition in this study to be compared by a subjective measure, using a tried and tested method.

The statements were updated to refer to *the AI assistant* instead of *this system*. Statements 1 and 10 were altered, after hallway-testing indicated that, within the context of the test, they were a source of confusion for participants. The new statements were phrased in such a way as to try and stay true to their original meaning.

The 1^{st} statement was also rephrased to assess how much participants were using AI decision support. Individual statement scores are not considered meaningful in the SUS method (Brooke, 1995). The objective however was to indicate if participants did or did not use the AI decision support, and not to evaluate it. Having participants agree or disagree with the statement "I frequently used the AI assistant during the task", was thought to achieve that. Bansal et al. (2021) employed a qualitative, open-ended question to make that assessment. The 1st SUS statement was thought to provide a simple, closed-ended, quantitative way of gaining similar insights.

Statement
I frequently used the AI assistant during the task.
I found the AI assistant unnecessarily complex.
I thought the AI assistant was easy to use.
I think that I would need the support of a technical person to be able to use the AI assistant.
I found the various functions in this AI assistant were well integrated.
I thought there was too much inconsistency in the AI assistant.
I would imagine that most people would learn to use the AI assistant very quickly.
I found the AI assistant very cumbersome to use.
I felt very confident using the AI assistant.
I needed a lot of time before I could get going with the AI assistant.

Table 3.2: Updated SUS statements

3.2.4 Test procedure

The experiment consisted of 3 steps.

1. Consenting to take part and completing the demographic survey.
2. On-boarding and completing the task of labeling 25 or 50 beer reviews.
3. Completing the SUS survey.

Consent and demographic survey

The website's landing page gave an overview of the study, an estimated time commitment (30 minutes; a conservative estimate based on the time taken by participants during hallway-testing), and information about the optional raffle. It gave details on data that would be collected, and warning statements regarding the topic of the task, which will be discussed in more detail below. Visitors gave their implied consent by clicking a button labeled *take part*, and entering a password that was included in their invitation to participate.

Participants then completed the demographic survey, at which point they were given the option to enter the raffle by providing their email.

On-boarding and beer review labeling

On-boarding. Participants were introduced to the task and the AI assistant during an on-boarding phase. They were given instructions on labeling the beer reviews as either positive or negative. They were informed that the AI assistant would provide a sentiment analysis of each review, and give a confidence score with each recommendation.

Here again the study method deviated slightly from Bansal et al. (2021). It did not include a practice round. Zhang et al. (2020) justified including practice tasks in their study because crowd-workers were unlikely to be familiar with the domain. That was not deemed to be a concern in this study. During hallway-testing, participants were clear about the task, and there was no indication that they would have benefited from a practice round. Practice rounds may be an ethical requirement when participants are rewarded or penalised based on performance, which was the case in the studies by Bansal et al. (2021) and Zhang et al. (2020). That was not the case here.

Labeling. During the labeling task, participants were shown each beer review one at a time. The AI assistant recommended the predicted sentiment. The AI assistant's confidence in each sentiment was displayed, a confidence of P in the predicted class, and a confidence of 1 - P in the alternative. The AI assistant gave its recommendation by highlighting its prediction. Participants labeled each beer review as positive or negative by clicking the appropriate button. See figure 3.2.

After 25 reviews, participants were given the option to finish the task and move onto the SUS survey. This was not done by Bansal et al. (2021), but seemed appropriate when participants were volunteering their time.

The order of the 50 beer reviews was randomised before being split into two groups of 25. The process of splitting ensured that each set included 21 correctly classified, and 4 misclassified beer reviews, maintaining a model accuracy of 84%.

Participants were not shown whether they had labeled reviews correctly or incor-

rectly. This is another difference between this study and that by Bansal et al. (2021). Again, because participants were not rewarded or penalised for their performance, the purpose of showing performance was not obvious. It was thought to potentially influence behaviour and detract from the illusion of dealing with a real AI decision support system.

Metrics. Participants' decisions and the time they took to decide was recorded along with the predicted sentiment, ground truth, and confidence score. Participants' time to decide was measured by recording the time a review was initially displayed and the time the participant clicked the relative button.

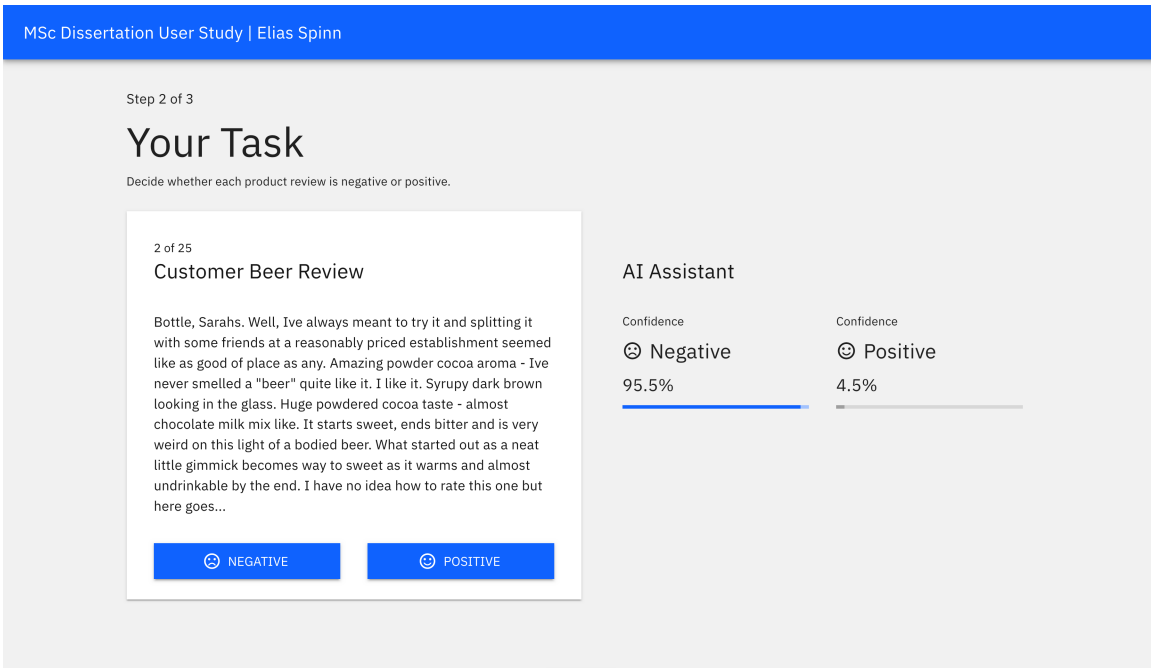


Figure 3.2: Beer review labeling task showing a numerical confidence score

3.2.5 Ethical considerations

A number of design decision were made in regards to the ethics of the study. Experimental ethics were reviewed by a professional UX researcher before testing.

To allow potential participants to make an informed decision on whether to take part or not, details about the study were provided on the website landing page, as described above. Information was given about the data that was to be collected if

they participated. As described above, consent was implied when participants clicked a button labeled *take part* and entering a participant password. No data was collected before that point. All demographic questions included an option not to disclose.

As already discussed, this study relied on voluntary participation. Participants were given the option to enter a raffle for a €50 Amazon Gift Voucher, both as an incentive to participate, and to provide some form of compensation. The decision to give participants the option to finish the task after completing 25 beer reviews was also done out of consideration that people were volunteering their free time.

The landing page informed participants that the experiment was not a test of their abilities. They were informed, and again reminded during on-boarding, that they were free to stop at any time.

Considering the task's topic of alcohol, two warning statements were added to the website landing page. 1^{st} that by taking part, participants were 18 years of age or older. 2^{nd} that people who suffer from alcohol dependency issues may wish not to take part. Links to Irish and international alcohol support organisations were provided.

3.3 UI design

One of the study's objectives was to ensure that the application was usable and guidance was clear. It was important that the user interface and content did not impede participants from completing any aspect of the test. Low usability was a stated limitation in the study by Stoll et al. (2022). They discussed the importance of ensuring general usability that does not interfere with experimental measurements.

The process of designing the application included 2 iterations of design and evaluation. The 1^{st} iteration included visual design and information architecture of the application before development started. After initial implementation, the application was evaluated with 2 expert reviews. Design updates were made and implemented in the 2^{nd} iteration. The application was then evaluated more formally by conducting 3 usability tests. Final changes were minor, and were reviewed informally.

Application design and development used the Material UI component library, an

open-source implementation of Google's Material Design. Using a widely adopted component library ensured the application UI followed good visual and interaction standards.

3.4 Application development

A full-stack application, running on Amazon Web Services (AWS) cloud infrastructure, was developed to run the user tests. The client application was a React (Javascript library) based web-application, served by an Ngnix web-server running in a Docker container on AWS's Elastic Container Service (ECS). The web-service application was a Hasura GraphQL API, also running in a Docker container on ECS. Data storage was provided by a PostgreSQL data base running on AWS's Relational Database Service (RDS). The website's domain, secure HTTP, and routing was handled by an AWS application load balancer.

The application architecture was developed to meet the following study requirements and security considerations.

1. Randomly assigning participants to 1 of the 3 conditions.
2. Preventing people from participating more than once.
3. Securing the test data.

3.4.1 Independent sampling

To ensure independent sampling, participants were randomly assigned to 1 of the 3 conditions. Each experimental condition was served by a separate web-server, running in a separate container service. The application load balancer distributed incoming traffic equally across the 3 web-servers. In this way, a visitor to the website was assigned to 1 of the 3 conditions.

The application load balancer used a session token to persist which web-server was serving which web-browser. Known as a sticky session, this ensured that, as long as the same browser was used, participants would be served the same condition if they

returned at a later date. Due to the uncontrolled nature of unmoderated user testing, it was presumed that some participants would start the experiment out of curiosity and complete the task at a later, more convenient time. This was observed during testing.

3.4.2 Preventing multiple attempts

To maintain sample independence the application was required to prevent participants from completing the test more than once. Participants' progress was recorded as they completed each step i.e. the demographic survey, the task, and the SUS survey. State management on the client application re-routed participants away from steps they had already completed. One limitation was that participants' labeling task progress state was not managed. The task would reset and restart if a participant closed and re-opened, or refreshed the browser at any point during the task. However, a unique id showed the number of beer reviews a participant had labeled, indicating multiple attempts. Their data could be discarded before data analysis. The same email address associated with more than 1 participant id would also indicate multiple attempts. It was also presumed that labeling 50 beer reviews would not be an exercise someone would want to complete more than once. However, these mechanisms can not guarantee multiple attempts were prevented in full.

3.4.3 Data security

The security of participants' personal information was a primary concern. The website was SSL certified to secure network communication over HTTPS. Participation was password protected. A password was shared in the invitations to participate.

API access was managed using access tokens, and access policies. Read and write permissions were limited to the bearer of a JSON Web Token (JWT) which was shared with the client application. Read access was further limited to non-sensitive data only. Access to all AWS services, including the database, was heavily restricted using the cloud provider's security policies.

3.5 Recruitment

The user test was online and open to participation for 18 days. During that time, participants were recruited with open calls sent out to academic, professional, and personal contacts.

Invitations were sent incrementally. Smaller groups were initially invited to allow the application to be monitored for any defects. Invitations were sent to larger groups after the application was shown to work as intended.

Chapter 4

Results, evaluation and discussion

4.1 Sample description

105 tests were recorded where participants labeled ≥25 reviews. In line with Bansal et al. (2021), data from participants whose median labeling time was <2 seconds, or who labeled all reviews the same, were discarded. 2 were removed. The data was reviewed for indications of multiple attempts. 2 more were removed after filtering for email addresses associated with more than 1 participant id and 1 was removed for labeling 75 beer reviews.

During initial analysis, 2 outliers were seen in both the accuracy (0.5, 0.54) and inter-rater reliability (κ -0.08, 0.08). With a binary classification problem, an accuracy of 0.5 can be achieved with random labeling, and low to negative inter-rater reliability values can be interpreted as showing no agreement (McHugh, 2012). This suggested that the 2 participants may have labeled reviews at random. For that reason their data were discarded.

Data from 98 tests were included in the final analysis. See table 4.1 for sample details and table 4.2 for demographic details.

Sample	Num.	Ver.	VNum.
Participants	32	32	34
Reviews	1407	1425	1498
SUS respondents	31	31	32

Table 4.1: Sample sizes

Question	Option	Num.	Ver.	VNum.
Age	18-24	1	5	2
	25-44	24	23	26
	45-64	7	4	5
	65-100	0	0	1
Gender	Female	19	14	15
	Male	12	18	18
	Other/not stated	1	0	1
Education	Level 1	2	4	1
	Level 3	11	16	15
	Level 4	19	12	18
Reading	Somewhat comfortable	1	0	0
	Very comfortable	31	32	34
Comp. Sci./Eng.	None	14	16	15
	Educational	8	5	2
	Professional	3	4	4
	Both	7	7	13

Table 4.2: Demographic description of participants

4.1.1 Significance tests

Before results were tested for significance, sample distributions were checked for normality. A density plot was used as a visual check, and the Shapiro-Wilk method was used as the normality test.

A pairwise Mann-Whitney U test was used on non-normally distributed samples. A pairwise Student's T Test was used on normally distributed samples

Sample distributions tended to be non-normal. Significance tests described in this chapter were made with a Mann-Whitney U test, unless stated otherwise.

Pairwise comparisons were corrected for family-wise error rate with the Bonferroni method (Bansal et al., 2021; Jesus et al., 2021). Alpha (0.05) was divided by the number of comparisons (3).

$$\frac{\alpha 0.05}{3} = \alpha 0.0167$$

4.2 Accuracy

Based on the beer reviews' ground truth, each participant's labeling decision was categorised as a true positive, true negative, false positive, or false negative label. Each participant's accuracy was calculated as:

$$\frac{tp + tn}{tp + tn + fp + fn}$$

Participant accuracy was compared between condition on all reviews, misclassified reviews where by the AI assistant was incorrect in its recommendation, and by categories of confidence i.e. *highly likely*, *likely*, and *even chance*. See table 4.4, figure 4.1, and figure 4.2 for results.

Beer review	Num.	Ver.	VNum.
All reviews	1407	1425	1498
Misclassified	226	240	228
Highly likely	474	482	507
Likely	846	855	900
Even chance	87	88	91

Table 4.3: Number of reviews labeled

Data-set	Accuracy	Num.	Ver.	VNum.
All reviews	Mean	0.874	0.840	0.872
	Std	0.050	0.064	0.060
Misclassified	Mean	0.864	0.765	0.841
	Std	0.130	0.200	0.138
Highly likely	Mean	0.885	0.851	0.878
	Std	0.082	0.100	0.109
Likely	Mean	0.865	0.838	0.862
	Std	0.064	0.075	0.076
Even chance	Mean	0.901	0.770	0.931
	Std	0.163	0.273	0.159

Table 4.4: Participant accuracy

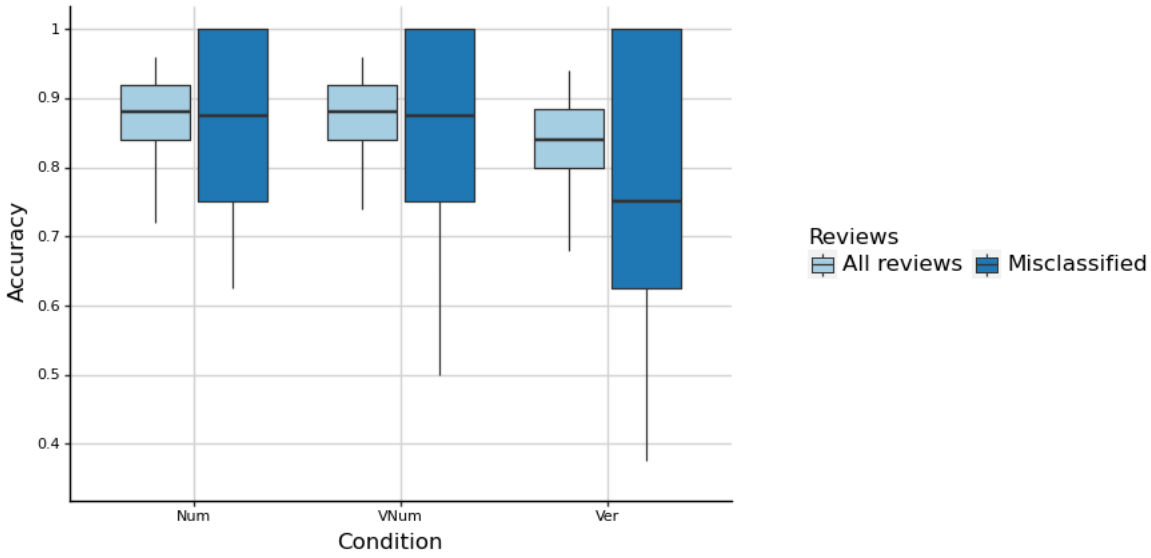


Figure 4.1: Participant decision accuracy

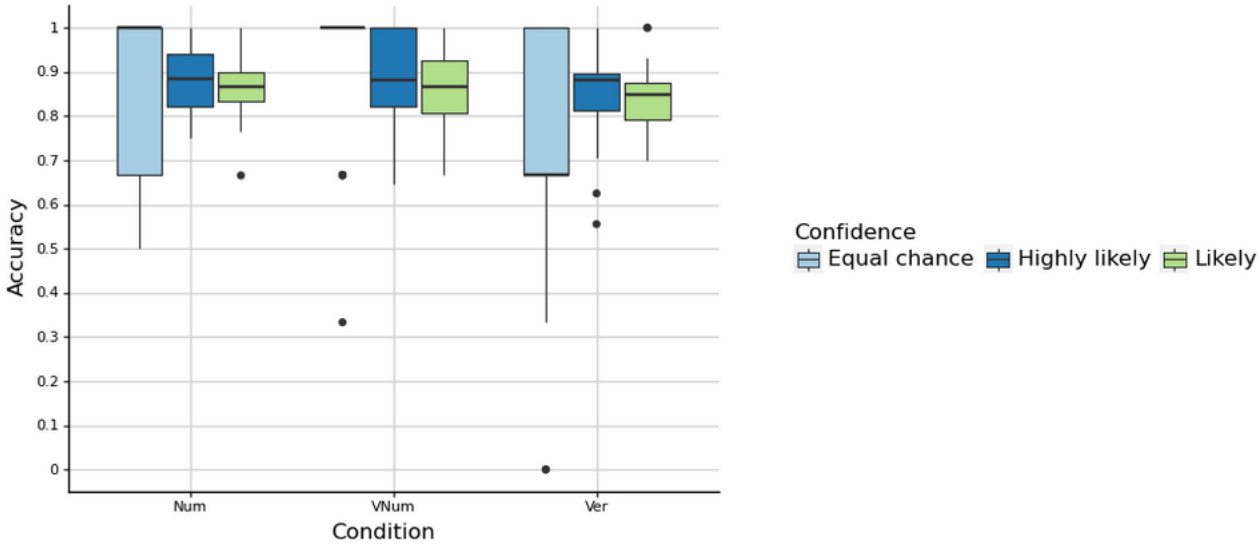


Figure 4.2: Participant decision accuracy by confidence

4.2.1 Accuracy results

All reviews

No significant differences were observed on overall accuracy. *Num.* and *VNum.* conditions both showed higher accuracy compared to *Ver.* The biggest differ-

ence was seen between the *Num.* and *Ver.* conditions (p=0.024). A similar difference was see between *VNum.* and *Ver.* (p=0.042). Accuracy scores in *Num.* and *VNum.* were shown to be similar (p=0.969).

Misclassified

No significant difference was observed on misclassified reviews. *Num.* and *VNum.* again showed higher accuracy compared to *Ver.*. The biggest difference was also seen between *Num.* and *Ver.* conditions (p=0.049), and then between *VNum.* and *Ver.* (p=0.125). In this sub-set, a slightly more pronounced difference was seen between *Num.* and *VNum.* (p=0.59).

By confidence

Only small differences were seen in the 2 sub-sets of reviews, which had confidence score values within the probability scale ranges of *highly likely* and *likely*. Again *Num.* accuracy was always slightly higher, closely followed by *VNum.*, and lastly *Ver.*. The highest differences were again between *Num.* and *Ver.* (*highly likely* p=0.180, *likely* p=0.0850), and then *VNum.* and *Ver.* (*highly likely* p=0.120, *likely* p=0.307). In both *highly likely* and *likely*, *Num.* and *VNum.* showed very similar results (p=0.968, p=1.0).

A significant difference was observed in the sub-set of reviews that had a confidence score within the probability range of *even chance*. The *VNum.* condition showed significantly higher accuracy compared to *Ver.* (p=0.003). The next largest difference, but not significant, was between *Num.* and *Ver.* (p=0.038). Only in this sub-set did *VNum.* achieve higher accuracy compared to *Num.* (p=0.337). See figure 4.3. It should be noted that the number of labeled reviews in this probability range is an order of magnitude smaller than any of the other sub-sets.

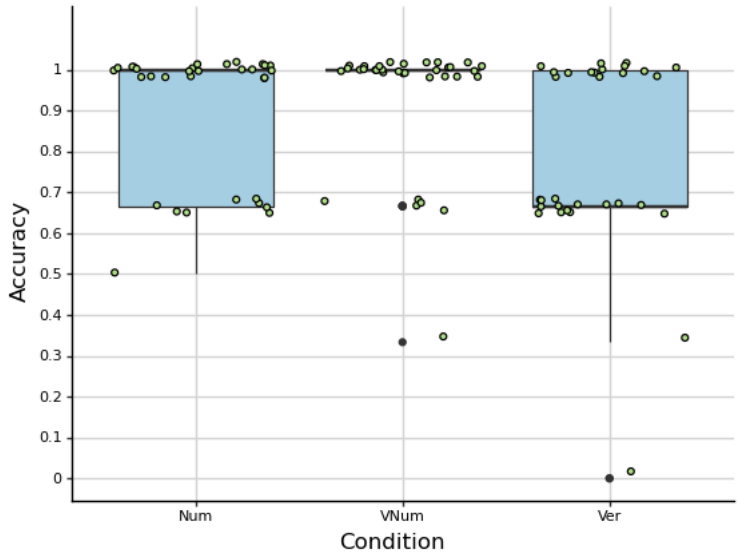


Figure 4.3: Participant decision accuracy on even chance confidence

4.2.2 Discussion on accuracy

Because overall accuracy results, and the results of misclassified reviews, showed no significant differences, this study does not reject either the 1^{st} or 2^{nd} null hypotheses.

Results suggest that numerical expressions of probability might be more effective than verbal ones. Both conditions that included the numerical form did show higher accuracy on all sub-sets, albeit not significant. And the *VNum.* condition did show significantly higher accuracy than the *Ver.* condition, on reviews within the *even chance* confidence range.

One interesting observation is that human+AI team performance did not improve on AI performance in the *Ver.* condition. Average team accuracy was 84%, equal to the prediction accuracy in the beer review data-set. Team performance only showed improvement in the 2 conditions that included a percentage confidence score.

The order of performance stayed constant in all results; *Num.* showed highest accuracy, followed by *VNum.*, and *Ver.* last. Except for in the *even chance* sub-set, where the *VNum.* condition had the highest accuracy.

Teigen (2022) argues that verbal expressions are more effective in conveying recommendations or warnings compared to numerical ones. Showing that a recommendation

has an *even chance* of being correct or incorrect alongside a numerical confidence score, may have been more effective in influencing participants to be critical of that recommendation, than simply showing a percentage value.

This might point to a way forward in developing interactions that better influence decision makers to be critical of recommendations when confidence is low, and could be a worthwhile topic for future studies.

4.3 Agreement levels

Following the method by Bansal et al. (2021), levels of agreement between participants and the AI assistant's recommendations were analysed using Cohen's kappa (κ). Every participant's κ was calculated on all reviews, misclassified reviews, correctly classified, and by categories of confidence i.e. *highly likely*, *likely*, and *even chance*. See table 4.5 for results. Averages were mapped to agreement levels, based on guidance by McHugh (2012), on how to interpret κ. See 4.6 for agreement levels.

4.3.1 Discussion on agreement levels

The results show an overall weak level of agreement on all reviews. Agreement seems to be correlated to the probability scale. Agreement is moderate on reviews within the *highly likely* probability range, and drops down to weak at *likely*, and none at *even chance*. On misclassified reviews, κ was so low in all conditions, as to be classified by McHugh (2012) as a strong disagreement or random data. With relatively high accuracy (0.86, 0.77, 0.84) on misclassified reviews, low agreement makes sense. However, even on correctly classified reviews, agreement levels are only moderate.

Agreement observed in this study is lower compared to those observed by Bansal et al. (2021). They saw an overall moderate agreement level (κ 0.71). Their analysis differed slightly but still makes for an interesting comparison. Based on qualitative observations, they divided reviews by a confidence threshold of 83%, below which their participants were more likely to distrust AI recommendations. Agreement levels on reviews above that threshold were almost perfect, and were moderate below it. This

Data-set	κ	Num.	Ver.	VNum.
All reviews	Mean	0.515	0.509	0.524
	Std	0.094	0.128	0.120
Misclassified	Mean	-0.729	-0.531	-0.683
	Std	0.260	0.400	0.277
Correct Class.	Mean	0.753	0.707	0.755
	Std	0.106	0.133	0.133
Highly likely	Mean	0.766	0.693	0.749
	Std	0.165	0.201	0.224
Likely	Mean	0.458	0.471	0.475
	Std	0.146	0.147	0.127
Even chance	Mean	0	0	0
	Std	0	0	0

Table 4.5: Inter-rater reliability

Data-set	Num.	Ver.	VNum.
All reviews	Weak	Weak	Weak
Misclassified	Strong disagreement	Strong disagreement	Strong disagreement
Correct Class.	Moderate	Moderate	Moderate
Highly likely	Moderate	Moderate	Moderate
Likely	Weak	Weak	Weak
Even chance	None	None	None

Table 4.6: Level of agreement

	Num.	Ver.	VNum.
SUS respondents	31	31	32

Table 4.7: SUS sample sizes

study observed much lower agreement levels. Agreement was moderate on reviews with a confidence $> 90\%$, and was at best weak on $< 90\%$.

Bansal et al. (2021) found that 23% of their participants mostly ignored AI recommendations. The observed lower levels of agreement in this study suggested a higher ratio of participants may have disregarded AI decision support. SUS results provided more evidence for this behaviour.

4.4 Subjective evaluation

SUS results were used to compare the 3 conditions by their SU score, and to further investigate to what extent participants used AI decision support. 4 participants did not completed the evaluation. See table 4.7 for details.

Before analysis, SUS scores were inverted for statements 1,3,5,7,9. A respondent's SU score was calculated as the sum of all scores multiplied by 2.5 (scaled to a range of 0-100). (Brooke, 1995)

4.4.1 SU score results

SU score results show similar subjective evaluations in all conditions. See table 4.8. *Num.* was rated highest, followed by *Ver.*, and lastly *VNum.* A pairwise Student's T test showed small differences between them (*Num./Ver.* p=0.606, *Num./VNum.* p=0.492, *Ver./VNum.* p=0.845).

4.4.2 Indicated use of AI decision support

Response counts of the 1^{st} SUS statement, *I frequently used the AI assistant during the task*, gives further evidence that the majority of participants did not use AI decision

SU score	Num.	Ver.	VNum.
Mean	68.54	67.01	66.32
Std	10.22	12.90	14.76

Table 4.8: System Usability score

Response	Num.	Ver.	VNum.
Strongly disagree	9	10	7
Disagree	13	10	16
Neither	4	4	6
Agree	4	7	3
Strongly agree	1	0	0

Table 4.9: Response to: *I frequently used the AI assistant during the task.*

support. As shown in table 4.9, 65/94 (69%) of participants either disagreed or strongly disagreed with the statement. Only 15/94 respondents (16%) agreed in any way.

4.4.3 Discussion on subjective evaluation

Overall subjective evaluations were similar in all 3 conditions. Responses to the 1^{st} statement gave a clear indication that few participants used AI decision support. The ratio of participants who disregarded the AI (at least 69%) was higher than that observed by Bansal et al. (2021) (23%).

These results also align with anecdotal evidence from informal conversations with volunteers. Participants reported that they mostly ignored AI recommendation. 3 people even went so far as to cover it up with a sheet of paper.

Bansal et al. (2021) suggested that participants tend to ignore decision support in domains where the AI does not provide the decision maker with any additional expertise. This aligns with the argument by Riveiro et al. (2014), that the difficulty of the decision task, and the decision maker's own domain experience, influences how much they will use decision support. When considering the nature of the decision task

in this study, these findings support that argument.

As is common in other XAI task-based studies (Buçinca et al., 2021; Zhang et al., 2020), this and the Bansal et al. (2021) study chose decision tasks that don't require specialized domain knowledge and are suitable for general participation and crowd sourcing. These experimental methods of testing human+AI team performance deliberately create scenarios where "humans have comparable domain knowledge" to the AI (Zhang et al., 2020). But if people tend to ignore AI decision support, unless it gives them additional expertise that they don't perceive to have themselves, then these type of studies are susceptible to low adoption. Results in these studies may not predict the results in conditions where the AI does provide the decision maker with additional domain knowledge and expertise.

Future work might look to develop decision-task scenarios suitable for general participation, where human and AI performance is comparable, but the difference between their domain expertise is somehow greater.

4.5 Analysis by usage of AI decision support

For further analysis, participants were divided by how much they indicated using AI decision support. Participants who either disagreed or strongly disagreed with the 1^{st} SUS statement were grouped as *disregarding AI recommendations*, and those that neither agreed or disagreed, agreed, or strongly agreed were grouped as *considering AI recommendations*.

The overall accuracy and SU scores of these 2 groups was calculated. Only participants who completed the SUS survey were included. See table 4.10 and figure 4.4 for details.

Group	Sample	Num.	Ver.	VNum.
Considered AI	Participants	9	11	9
	Reviews	425	475	400
Disregarded AI	Participants	22	20	23
	Reviews	952	900	1048

Table 4.10: Participants grouped by agreement to statement "I frequently used the AI assistant during the task." Strongly disagree and disagree are group as disregarding AI. Neither, agree, and strongly agree grouped as considered AI.

4.5.1 Accuracy results

Accuracy of participant who considered AI recommendations

Num. and *VNum.* conditions in this group, had the highest average overall accuracy seen in this study (0.891, 0.884). *Ver.* on the other hand had the lowest (0.807).

A significant difference was seen between *Num.* and *Ver.* A pairwise Student's T Test showed that the *Num.* condition had significantly higher accuracy than *Ver.* (p=0.0131). *VNum.* showed higher, but not significantly higher accuracy than *Ver.* . *Num.* and *VNum.* showed small differences (p=0.756). See figure 4.5

Accuracy of participant who disregarded AI recommendations

This group showed similar accuracy in all conditions as seen in table 4.11 and figure 4.11. The order of accuracy matched general findings. *Num.* again had the highest accuracy, followed by *VNum.*, and *Ver.* last. Their similarity was seen in a pairwise Student's T test. (*Num./Ver.* p=0.600, *Num./VNum.* p=0.402, *Ver./VNum.* p=0.138).

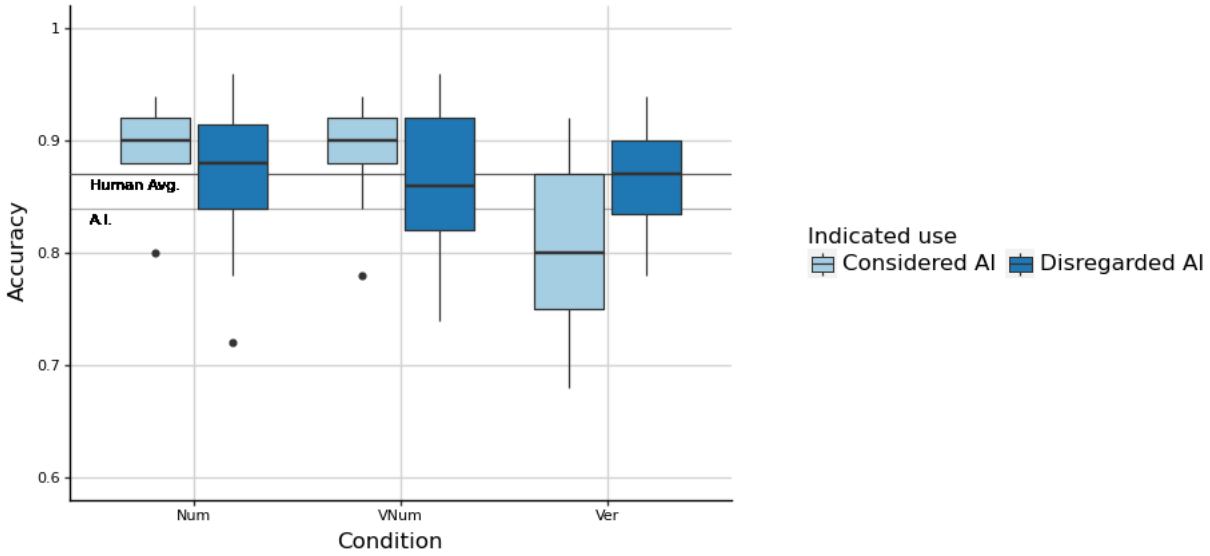


Figure 4.4: Participant decision accuracy by indicated use of AI decision support

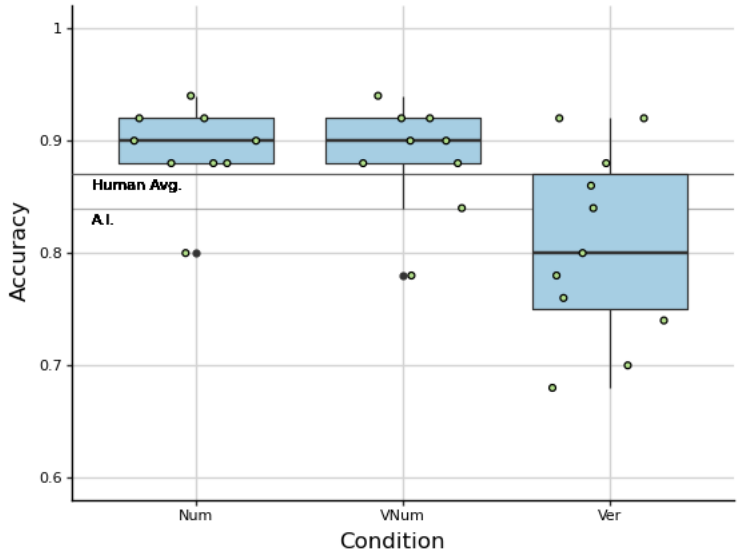


Figure 4.5: Decision accuracy of participants who considered AI recommendations

4.5.2 SU score results

SU scores of participant who considered AI recommendations

As seen in table 4.12, this group gave the highest subjective evaluation scores seen in this study. The group also showed the biggest differences between conditions, although

Group	Accuracy	Num.	Ver.	VNum.
Considered AI	Mean	0.891	0.807	0.884
	Std	0.040	0.083	0.048
Disregarded AI	Mean	0.868	0.859	0.861
	Std	0.055	0.046	0.063

Table 4.11: Participant accuracy on all reviews

Group	SU Score	Num.	Ver.	VNum.
Considered AI	Mean	75.00	70.22	75.55
	Std	7.18	10.98	10.51
Disregarded AI	Mean	65.90	65.25	62.71
	Std	10.22	13.78	14.78

Table 4.12: Participant System Usability Scale evaluation

none were significant. (*Num./Ver.* p=0.302, *Num./VNum.* p=0.100, *Ver./VNum.* p=0.320).

SU scores of participant who disregarded the AI

As would be expected, this group scored the AI assistant lowest. This group also showed the smallest differences between conditions. (*Num./Ver.* p=0.714, *Num./VNum.* p=0.707, *Ver./VNum.* p=0.732).

4.5.3 Discussion on indicated use of AI decision support

The bigger differences and one significant difference seen in the accuracy of the group that *considered AI recommendations*, and the negligible differences seen in the group that *disregarded AI recommendations*, give further insights to the initial results. It explains how initial results showed some differences between conditions, whilst agreement levels were shown to be so low.

Numerical expression of confidence did show significantly higher accuracy compared to verbal expressions in the group that *considered AI recommendations*. Sample sizes however were much smaller due to overall low adoption of AI decision support. These results are only based on 29 tests for all 3 conditions.

An interesting observation is that there is no apparent trade off between performance and preference in the group that *considered AI recommendations*. As seen in table 4.11 and 4.12, numerical and verbal-numerical expressions of confidence showed both higher accuracy and higher SU scores. This might point to a way forward in developing interactions that encourage critical thinking without the cost of user acceptance. (Buçinca et al., 2021)

One final observation is that the average accuracy of participants who *considered AI recommendations* in the *Ver.* condition (80%), was lower than AI accuracy alone (84%). The average accuracy of participants who *disregarded AI recommendations* (87%, 86%, 86%) was comparable to the average unassisted human accuracy (87%) observed by Bansal et al. (2021). Only groups who *considered AI recommendations* and were shown numerical expressions of confidence achieved average human+AI team accuracy (89%, 88%) that was higher than either working solo. But only just.

Chapter 5

Conclusion

5.0.1 Discussion

This study was an empirical AI decision support evaluation, using a task-based, user-centered approach. It measured decision performance of human+AI teams through a series of unmoderated user tests, where participants reviewed and labeled the sentiment of beer reviews. Participants' final decision accuracy was compared across conditions where AI confidence scores were expressed either numerically, verbally, or verbal-numerically.

To evaluate human+AI team performance, decision accuracy was analysed on all beer reviews. To evaluate the effect on trust calibration, decision accuracy was analysed on examples where AI recommendations were incorrect. Participants' subjective evaluations were analysed and compared to team performance.

The hypotheses expected significant differences, both on overall decision accuracy and decision accuracy on examples where AI decision support was incorrect. Results showed differences, but not significant. Numerical expressions showed the highest decision accuracy both overall, and on examples of incorrect AI recommendations.

Analysis of results indicated that a low number of participants used the AI decision support during testing. This was evident in both the low inter-rater agreement levels, and subjective evaluation results.

This may be explained by the study design and the nature of the decision task. The

study design did not include incentives for participants to perform well and improve on AI performance. Incentives have been used in similar studies, where a bonus is given for correct decisions and a penalty for incorrect ones. Low use of AI decision support might also be explained by the nature of the decision task. Is has been argued that the complexity of a decision task, and the experience of the decision maker will impact how they will use AI DS systems (Bansal et al., 2021; Riveiro et al., 2014). This argument would provide a reason for why study participants were observed to mostly disregard AI recommendations, because AI performance was comparable to that of an unassisted human, and the decision task did not require specialised expertise.

Significant differences were found when analysing the performance of participants who indicated considering AI recommendations. In that group, expressing confidence numerically showed significantly higher decision accuracy than verbal expressions. Verbal-numerical expressions showed higher decision accuracy than verbal expressions, but not significantly. One other significant difference was seen when analysing the performance of all participants on examples where AI confidence was within the probability scale range of *even chance*. The verbal-numerical expressions showed significantly higher decision accuracy than verbal expressions.

These results suggest that expressing confidence scores numerically may be a more effective in supporting human+AI team performance when AI decision support is used by system operators. All significant results showed higher performance in conditions where a numerical expression was included. Analysis of groups based on their indicated use of AI decision support, showed that only participants who considered AI recommendations displaying a percentage confidence score, achieved higher average human+AI team performance compared to individual decision making. Expressing confidence verbally did not improve on AI performance, and actually decreased performance below that of the AI and the average unassisted human decision maker.

Numerical confidence scores may also be preferred by people who use AI decision support. Subjective evaluations in this group rated numerical and verbal-numerical expressions higher than the verbal expressions, but not significantly. This conclusion would be in agreement with the literature in uncertainty communication, that readers

of probability statements prefer numbers to words (Erev & Cohen, 1990; Jaffe-Katz et al., 1989).

These findings highlight the importance of carefully considering how probability is expressed in AI DSS, and the importance of drawing on research in probability and uncertainty communication.

However, neither of the study's hypotheses were accepted, because no significant differences were found in overall decision accuracy or on examples where AI decision support was incorrect. In line with previous work, this study did use a low significance level (0.016) to correct for family wise-error rates. It has been argued that while the Bonferroni correction reduces the risk of a type I error, it can increase the likelihood of a type II error (Armstrong, 2014).

However, analysis showed that participants who considered AI recommendations had the most influence on overall observed differences, and this was a small sub-group of participants. The majority of participants, who indicated dismissing AI recommendations, showed similar decision accuracy.

Given the study's moderate participant sample size compared to similar studies, and that significant observations were on small sub-group of this sample, the null hypotheses were not rejected. Further evaluative research would need to be done with larger participant numbers.

5.0.2 Limitations

As discussed, this study did not give participant incentives to perform well, which likely contributed to low use of AI decision support. The use of unmoderated online user tests also means that the test environment was not controlled. Participant engagement levels may have varied significantly due to these limitations. And, as already mentioned, participant sample sizes were moderate compared to other studies in the field of XAI research.

Whilst AI DSS is of particular interest in complex and consequential domains, the decision-task in this study was neither complex or consequential. Study results may not generalise to those domains.

This study assumed the data provided by Bansal et al. (2021) was accurately labeled, and predictions and confidence scores were reliable.

It must be noted that an oversight was made in the design of the verbal-numerical condition. All 3 conditions included a progress bar which visualised the confidence score. In the verbal and verbal-numerical conditions, tick marks were included to illustrate the probability scale categories. These ticks divided the bar into 5 equal categories for the verbal condition. These should have been re-scaled to match the NATO approximate probability scale in the verbal-numerical condition.

5.0.3 Future work & recommendations

Future work might replicate this study but within a condition where participants are more likely to consider AI recommendations. This might be achieved through the use of incentives, or by developing new decision-task scenarios where participants are less likely to ignore AI decision support. This leads to another thought on future work.

Conducting evaluative studies in real-world scenarios is costly in terms of time and money (Jesus et al., 2021). But studies that employ decision tasks that don't require specialized domain knowledge and are suitable for general participation have been shown to result in low use of AI decision support. Results in these studies may not predict results in conditions where AI does provide the decision maker with additional domain knowledge and expertise. Future work may look at developing decision-task scenarios that are suitable for general participation, but where AI performance is complementary rather than equivalent to that of an unassisted human. Methods found in uncertainty communication and decision science should inform that work.

Finally, future work might study the influential quality of verbal expressions in verbal-numerical forms, with the question of how they could encourage analytical thinking in DSS users. This work could focus on comparing numerical and verbal-numerical expressions, and measure signals of influencing decision makers. This may be a way forward in solving the challenge of developing interactions that encourage analytical thinking, without reducing user acceptance (Buçinca et al., 2021; Bansal et al., 2021).

Bibliography

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 chi conference on human factors in computing systems* (p. 1–18). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3173574.3174156 doi: 10.1145/3173574.3174156

Armstrong, R. A. (2014). When to use the bonferroni correction. *Ophthalmic and Physiological Optics*, *34*(5), 502-508. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/opo.12131 doi: https://doi.org/10.1111/opo.12131

Balog, K., & Radlinski, F. (2020). Measuring recommendation explanation quality: The conflicting goals of explanations. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval* (p. 329–338). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3397271.3401032 doi: 10.1145/3397271.3401032

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., ... Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 chi conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3411764.3445717 doi: 10.1145/3411764.3445717

Bradley, R., Helgeson, C., & Hill, B. (2017). Climate change assessments: Confidence, probability, and decision. *Philosophy of Science*, *84*(3), pp. 500–522. Retrieved 2023-05-03, from https://www.jstor.org/stable/26551841

Brooke, J. (1995, 11). Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, *189*.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces* (p. 454–464). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1145/3377325.3377498 doi: 10.1145/3377325.3377498

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021, apr). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1). Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1145/3449287 doi: 10.1145/3449287

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(2), 281 - 294. Retrieved from https://login.ezproxy-ta.tudublin.ie/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,url,shib&db=pdh&AN=1988-25447-001&site=ehost-live&scope=site

Chromik, M., & Butz, A. (2021). Human-xai interaction: A review and design principles for explanation user interfaces. In *Human-computer interaction – interact 2021: 18th ifip tc 13 international conference, bari, italy, august 30 – september 3, 2021, proceedings, part ii* (p. 619–640). Berlin, Heidelberg: Springer-Verlag. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1007/978-3-030-85616-8_36 doi: 10.1007/978-3-030-85616-8_36

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think i get your point, ai! the illusion of explanatory depth in explainable ai. In *26th international conference on intelligent user interfaces* (p. 307–317). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3397481.3450644 doi: 10.1145/3397481.3450644

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces* (p. 275–285). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3301275.3302310 doi: 10.1145/3301275.3302310

Dubiel, M., Daronnat, S., & Leiva, L. A. (2022). Conversational agents trust calibration: A user-centred perspective to design. In *Proceedings of the 4th conference on conversational user interfaces.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1145/3543829.3544518 doi: 10.1145/3543829.3544518

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 chi conference on human factors in computing systems* (p. 1–6). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1145/3290607.3312787 doi: 10.1145/3290607.3312787

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, *45*(1), 1-18. Retrieved from https://www.sciencedirect.com/science/article/pii/074959789090002Q doi: https://doi.org/10.1016/0749-5978(90)90002-Q

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, *28*(2), 210 - 216. Retrieved from https://login.ezproxy-ta.tudublin.ie/login?url=

https://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,url,shib&db=pdh&AN=2009-03297-009&site=ehost-live&scope=site

Gunning, D. (2019). Darpa's explainable artificial intelligence (xai) program. In *Proceedings of the 24th international conference on intelligent user interfaces* (p. ii). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3301275.3308446 doi: 10.1145/3301275.3308446

Holzinger, A., Carrington, A., & Müller, H. (2020, jun). Measuring the quality of explanations: The system causability scale (scs). *Künstliche Intelligenz*, *34*(2), 193-198. Retrieved from https://doi.org/10.1007/s13218-020-00636- doi: 10.1007/s13218-020-00636-z

Hornbæk, K., & Oulasvirta, A. (2017). What is interaction? In *Proceedings of the 2017 chi conference on human factors in computing systems* (p. 5040–5052). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3025453.3025765 doi: 10.1145/3025453.3025765

Hullman, J. (2016). Why evaluating uncertainty visualization is error prone. In *Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization* (p. 143–151). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1145/2993901.2993919 doi: 10.1145/2993901.2993919

Hullman, J., Qiao, X., Correll, M., Kale, A., & Kay, M. (2019). In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 903-913. doi: 10.1109/TVCG.2018.2864889

Jaffe-Katz, A., Budescu, D. V., & Wallsten, T. S. (1989). Timed magnitude comparisons of numerical and nonnumerical expressions of uncertainty. *Memory Cognition*, *17*, 249-264. Retrieved from https://link.springer.com/article/10.3758/BF03198463 doi: https://doi.org/10.3758/BF03198463

Jesus, S., Belém, C., Balayan, V., Bento, J. a., Saleiro, P., Bizarro, P., & Gama, J. a. (2021). How can i choose an explainer? an application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (p. 805–815). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1145/3442188.3445941 doi: 10.1145/3442188.3445941

Joslyn, S., & LeClerc, J. (2013). Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, *22*(4), 308-315. Retrieved from https://doi.org/10.1177/0963721413481473 doi: 10.1177/0963721413481473

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 chi conference on human factors in computing systems* (p. 5092–5103). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1145/2858036.2858558 doi: 10.1145/2858036.2858558

Knoblauch, T. A. K., Stauffacher, M., & Trutnevyte, E. (2018). Communicating low-probability high-consequence risk, uncertainty and expert confidence: Induced seismicity of deep geothermal energy and shale gas. *Risk Analysis*, *38*(4), 694-709. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.12872 doi: https://doi.org/10.1111/risa.12872

Langer, E., Blank, A., & Chanowitz, B. (1978, 06). The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology*, *36*, 635-642. doi: 10.1037/0022-3514.36.6.635

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 chi conference on human factors in computing systems* (p. 1–15). New York, NY, USA: Association

for Computing Machinery. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1145/3313831.3376590 doi: 10.1145/3313831.3376590

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*(4), 656-665. Retrieved from https://doi.org/10.1518/001872006779166334 (PMID: 17240714) doi: 10.1518/001872006779166334

McHugh, M. (2012, 10). Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, *22*, 276-82. doi: 10.11613/BM.2012.031

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38. Retrieved from https://www.sciencedirect.com/science/article/pii/S0004370218305988 doi: https://doi.org/10.1016/j.artint.2018.07.007

Parra, D., Valdivieso, H., Carvallo, A., Rada, G., Verbert, K., & Schreck, T. (2019, August). Analyzing the design space for visualizing neural attention in text classification. In *Proc. ieee vis workshop on vis x ai: 2nd workshop on visualization for ai explainability (visxai).*

Riveiro, M., Helldin, T., Falkman, G., & Lebram, M. (2014). Effects of visualizing uncertainty on decision-making in a target identification scenario. *Computers Graphics*, *41*, 84-98. Retrieved from https://www.sciencedirect.com/science/article/pii/S0097849314000302 doi: https://doi.org/10.1016/j.cag.2014.02.006

Shneiderman, B. (2021). Tutorial: Human-centered ai: Reliable, safe and trustworthy. In *26th international conference on intelligent user interfaces - companion* (p. 7–8). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3397482.3453994 doi: 10.1145/3397482.3453994

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393-1400. Retrieved from https://www.science.org/doi/abs/10.1126/science.1191181 doi: 10.1126/science.1191181

Stoll, E., Urban, A., Ballin, P., & Kammer, D. (2022). Can explainable ai foster trust in a customer dialogue system? In *Proceedings of the 2022 international conference on advanced visual interfaces.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1145/3531073.3534481 doi: 10.1145/3531073.3534481

Teigen, K. H. (2022). Dimensions of uncertainty communication: What is conveyed by verbal terms and numeric ranges. *Current Psychology*, 1936-4733. Retrieved from https://link.springer.com/article/10.1007/s12144-022-03985-0 doi: https://doi.org/10.1007/s12144-022-03985-0

Tintarev, N. (2007). Explanations of recommendations. In *Proceedings of the 2007 acm conference on recommender systems* (p. 203–206). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1297231.1297275 doi: 10.1145/1297231.1297275

Vilone, G., & Longo, L. (2021, dec). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion*, *76*(C), 89–106. Retrieved from https://doi-org.ezproxy-ta.tudublin.ie/10.1016/j.inffus.2021.05.009 doi: 10.1016/j.inffus.2021.05.009

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115*(4), 348 - 365. Retrieved from https://login.ezproxy-ta.tudublin.ie/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,url,shib&db=pdh&AN=1987-09386-001&site=ehost-live&scope=site

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In

Proceedings of the 2020 conference on fairness, accountability, and transparency (p. 295–305). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3351095.3372852 doi: 10.1145/3351095.3372852

Zimmer, A. C. (1984). A model for the interpretation of verbal predictions. *International Journal of Man-Machine Studies*, *20*(1), 121-134. Retrieved from https://www.sciencedirect.com/science/article/pii/S0020737384800097 doi: https://doi.org/10.1016/S0020-7373(84)80009-7

Appendix A

Additional content

A.1 Diagrams, designs and screenshots

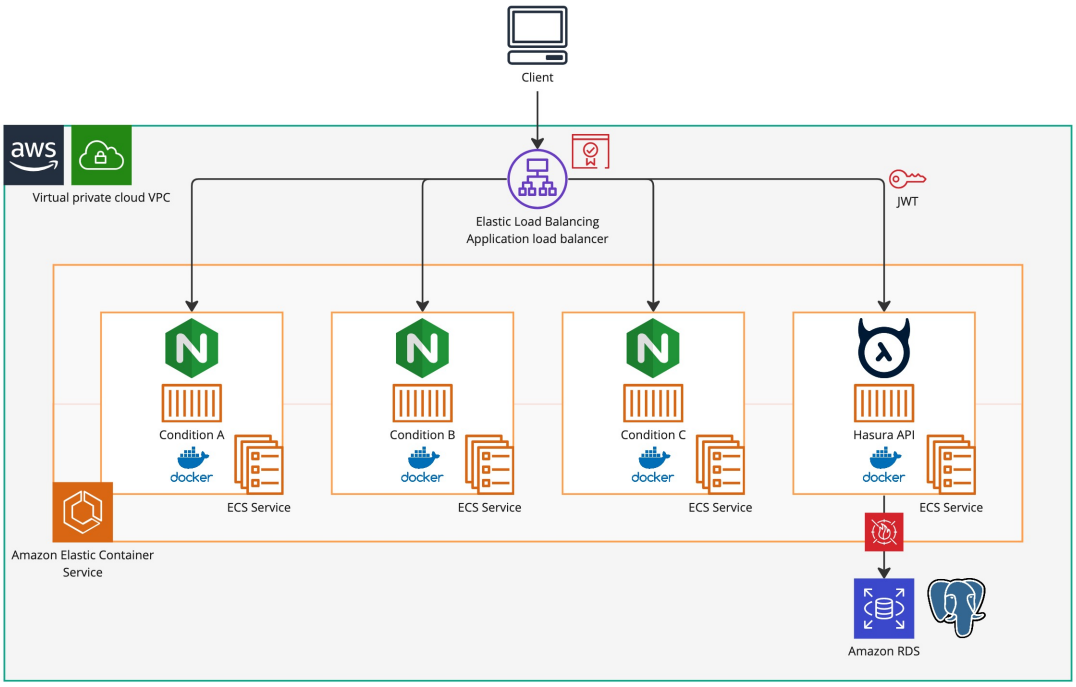


Figure A.1: Application architecture

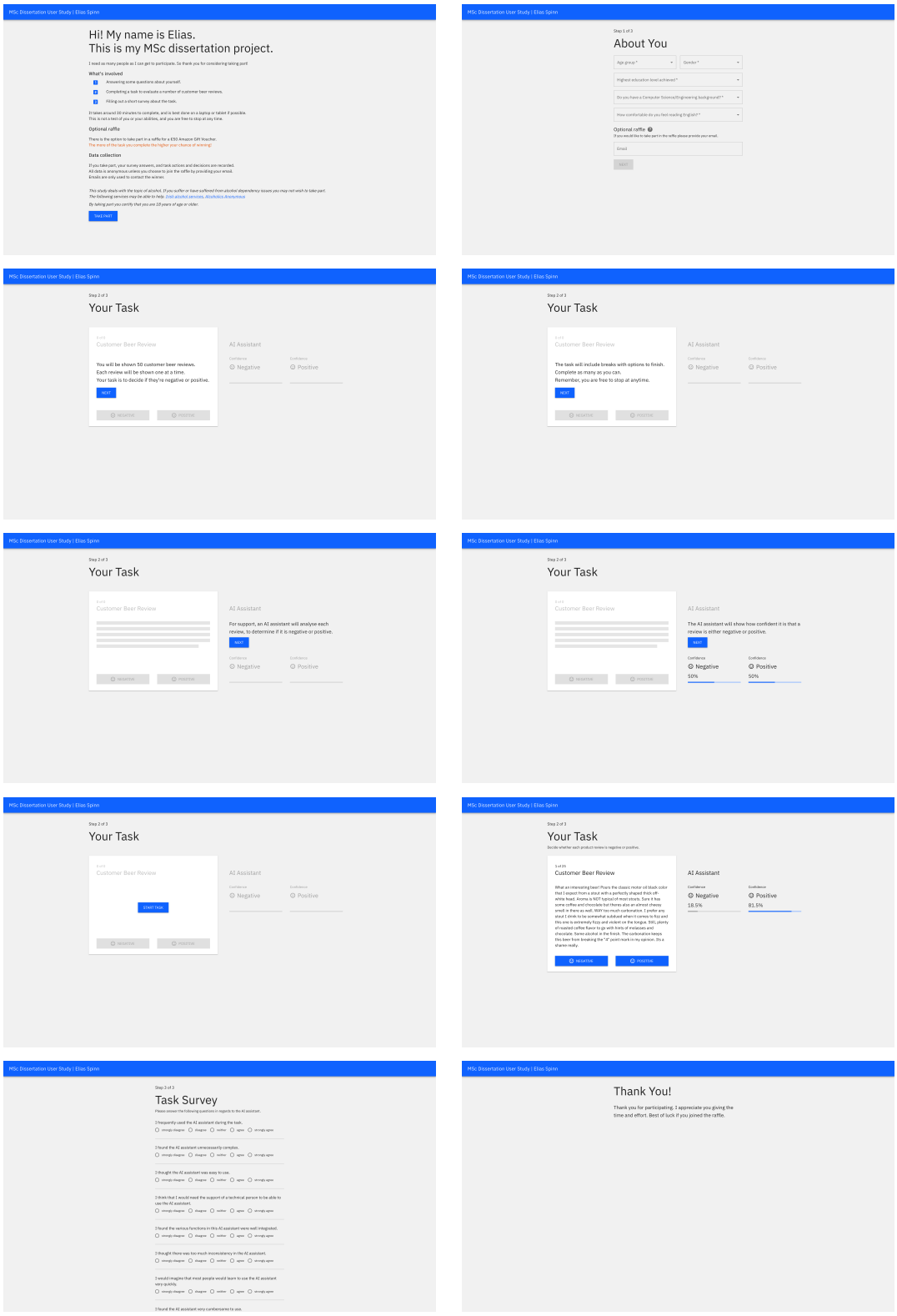
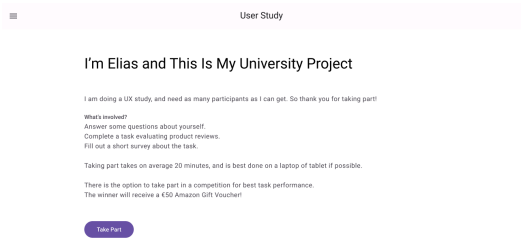
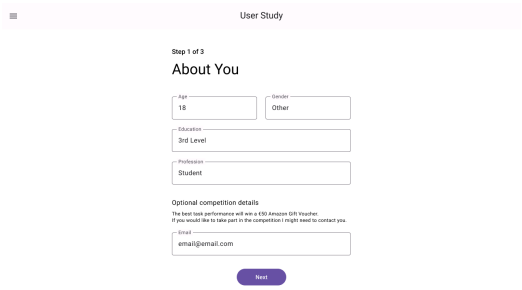
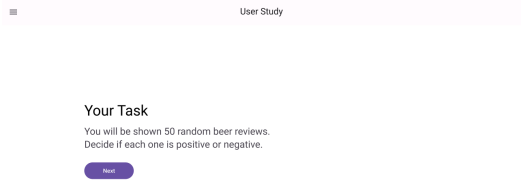
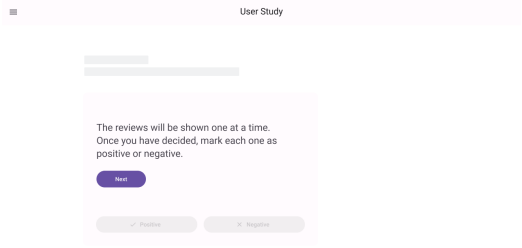


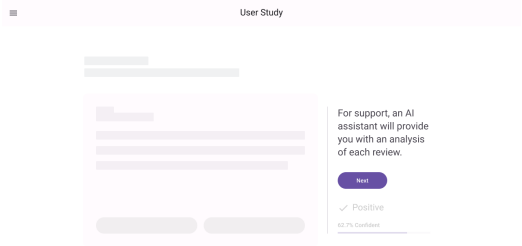
Table A.1: End-to-end user test

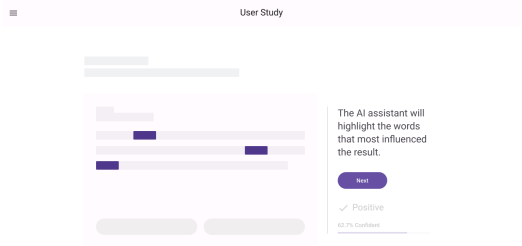


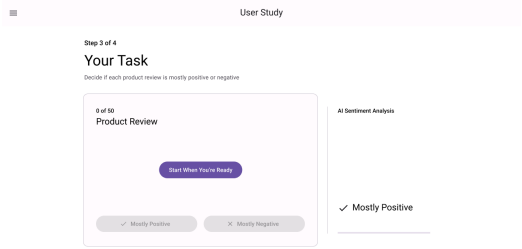


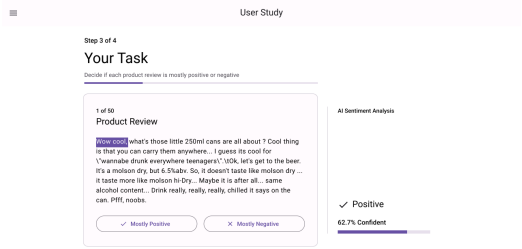












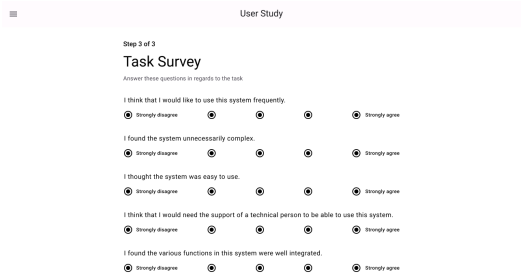


Table A.2: UI designs

A.2 Resources

Dissertation project repository

https://github.com/elspinn/msc-dissertation

NHS Gender Identity Guidance

https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/mental-health-services-data-set/submit-data/data-quality-of-protected-characteristics-and-other-vulnerable-groups/gender-identity

System Usability Scale (SUS)

https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

Usability testing

https://www.nngroup.com/articles/usability-testing-101/

Expert reviews

https://www.nngroup.com/articles/ux-expert-reviews/

Material UI

https://mui.com/

Hasura GQL

https://hasura.io/

Google's People+AI Guidebook

https://pair.withgoogle.com/chapter/explainability-trust/