To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Abri Bharos
born in Wageningen, the Netherlands

Web Information Systems Research Group
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems

Author: Abri Bharos
Student id: 4665392

Abstract

Powerful predictive AI systems have demonstrated great potential in augmenting human decision-making. Recent empirical work has argued that the vision for optimal human-AI collaboration requires '*appropriate reliance*' of humans on AI systems. However, accurately estimating the trustworthiness of AI advice at the instance level is quite challenging, especially in the absence of performance feedback pertaining to the AI system. In practice, the performance disparity of machine learning models on out-of-distribution data makes the dataset-specific performance feedback unreliable in human-AI collaboration. Inspired by existing literature on critical thinking and explanation-based human debugging, we propose the use of debugging an AI system as an intervention to foster appropriate reliance. In this paper, we explore whether a critical evaluation of AI performance within a debugging setting can better calibrate users' assessment of an AI system and lead to more appropriate reliance. Through a quantitative empirical study ($N = 234$), we found that our proposed debugging intervention does not work as expected in facilitating appropriate reliance. Instead, we observe a decrease in reliance on the AI system after the intervention — potentially resulting from early exposure to the AI system's weakness. We explored the dynamics of user confidence to help explain how inappropriate reliance patterns occur and found that human confidence is not independent of AI advice, which is potentially dangerous when trying to achieve appropriate reliance. Our findings have important implications for designing effective interventions to facilitate appropriate reliance and better human-AI collaboration.

Thesis Committee:

Chair:	Prof. Dr. G.J.P.M. Houben, Faculty EEMCS, TU Delft
University supervisor:	Asst. Prof Dr. U. Gadiraju, Faculty EEMCS, TU Delft
Committee Member:	Asst. Prof Dr. L. Siebert, Faculty EEMCS, TU Delft
Daily supervisor:	MSc Gaole He, Faculty EEMCS, TU Delft

Preface

Dear reader,

I write this with joy, for the document you are about to read marks the end of a journey. During this time, I have had the opportunity to develop myself, meet new people, and make many friends along the way. The only reason I have been able to do this, however, is because of the kind and loving people in my life. I would, therefore, like to dedicate this section to them whom I owe so much. First off, I would like to thank my father and mother for their endless support. No matter their predicaments, they always prioritized our education above all else. Secondly, I would like to thank my friends, who have done so well themselves and are the ones that made this journey enjoyable.

Now, the way this journey will end is by concluding the thesis to which I have dedicated the past nine months. During this time, I have not been working alone. I would therefore like to thank Professor Ujwal, who has taught me much through his insightful answers and provided support throughout my thesis. And most of all, I would like to thank PhD student Gaole He for his continued guidance, feedback, and company (during our long nights of testing).

This has been an incredible journey and I look forward to the next chapter. To whoever is reading this, I hope you enjoy my thesis!
Goodbye,

Abri Bharos
Delft, the Netherlands
November 21, 2022

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and Objectives

Big leaps in computing power combined with excessive demand have caused unprecedented growth in the capabilities of Artificial Intelligence (AI) systems. This has allowed these systems to permeate many different areas, becoming an integral part of many organizations. One particularly important application of AI is supporting human decision-making, where AI systems provide humans with advisory solutions or answers to specific problems or decisions. These systems have been widely adopted in areas like criminal justice, finance [4, 48, 93], healthcare [25, 63], and more [26, 39, 62, 90]. Through leveraging the complementary skills of humans and AI, a harmonious collaboration is created where humans can efficiently access the full range of relevant information to make decisions. If done well, this can result in complementary team performance, which is better performance than would be achieved by either the AI system or humans separately. In practice, however, it is often observed that team performance falls short of AI performance. Despite this superior performance, many practical, ethical, and judicial considerations make full automation of tasks undesirable. For team performance to exceed both human- and AI performance, human decision-makers need to be capable of identifying whenever AI advice is correct (and they should rely on the system) and when it is incorrect (and they should rely on themselves). This concept has been defined as *appropriate reliance*. One way of promoting appropriate reliance has been through conveying the reasoning behind AI advice to users by generating meaningful explanations. According to GDPR, human decision-makers are allowed to receive such meaningful explanations from AI systems they work with [84]. Although this has been successful to some extent [54, 92], this is not always the case [17]. And even when explanations are successful, there is no guarantee that team performance will be superior. This has incited much research into this area and many different interventions aimed at improving appropriate reliance between humans and AI systems have been proposed.

Existing work has explored many different user factors that affect human trust in and reliance on AI systems like expertise [24, 70], risk perception [38], machine learning literacy [19], and interaction designs (e.g. performance feedback [7, 78, 95], explanation [89], and user tutorials [55]). These have been successful to varying degrees. A potential inter-

vention that has not been proposed or researched is the usage of debugging exercises as a way of training users, on a case-by-case basis, to think more critically about whether they should trust specific AI advice. Debugging an AI system incentivizes users to understand the system, as to make it easier to find any bugs in the system. It is widely agreed upon in psychology that critical thinking skills contain both generic and domain-specific aspects [53, 29]. These skills can be taught through ill-structured problems that require one to go beyond simply stating previously learned information[53]. As debugging exercises fulfill these criteria, they can potentially improve the domain-specific critical thinking skills of humans in the area of human-AI collaboration, thereby enabling users to make better judgments on the trustworthiness of AI advice. In this work, we, therefore, explore the use of a debugging intervention to facilitate appropriate reliance.

1.2 Research Questions

In many studies in the area of human-AI collaboration, humans are provided with some form of performance feedback (e.g. accuracy). In practice, however, it is not uncommon for AI advice to be devoid of such feedback. Extant works rarely treat cases where performance feedback is absent, which is why research in this area is scarce. This is the setting that will be applied in our work. To realize the goal of appropriate reliance (AR), humans need to be able to evaluate the advice and trustworthiness of AI systems. In practice, when users work together with AI systems, it is common for them to encounter data from unknown distributions and unknown contexts. Inspired by recent works on explanation-based human debugging (EBHD) [6, 58], we propose EBHD as a training intervention to increase appropriate reliance on AI systems. We posit that such a debugging intervention has the potential to help users understand the limitations of AI systems — that neither explanations of the AI advice nor the advice itself are always reliable. Recognizing these limitations can help users better understand when an AI system is trustworthy and thereby increase appropriate reliance on the system. In this study, we aim to empirically evaluate the effectiveness of using a debugging intervention as a means to increase appropriate reliance. To do this, it is paramount that users are able to judge the trustworthiness of machines at both the global and instance level. We, therefore, propose the following research questions:

- RQ1: How can a debugging intervention help users to estimate the performance of an AI system, both at the instance and at the global level?
- RQ2: How does a debugging intervention affect the reliance of users on an AI system?

1.3 Contributions

- A comprehensive study on the effects of a debugging intervention on performance estimation, appropriate reliance, trust, and confidence dynamics.

- Confirmation of previous works showing that humans may be subject to cognitive bias arising from the internal ordering of debugging interventions or interventions similar to it.
- Implications for designing interventions aimed at increasing appropriate reliance of humans on AI systems.

1.4 Thesis Outline

In chapter 2 we introduce the topics upon which our research is based and cover the relevant research related to our work. Chapter 3 states the hypotheses, and their measures and explains the intervention in detail, after which chapter 4 will outline the experiment that has been conducted. The results of this experiment will be shown in chapter 5, where the data analysis will be covered. Chapter 6 presents the findings from the experiment, their implications, and limitations. Finally, there will be some concluding remarks and proposed future work in chapter 7.

Chapter 2

Background and Related Work

This chapter will cover the background knowledge required to understand the research, its purpose, context, and methodology. First, we dive deeper into Human-AI interaction, its dynamics and context. Second, we we talk about Explainable AI (XAI), a tool that is used during the experiment. After that, there will be a section explaining the reasoning behind choosing EBHD as an intervention. Finally, there will be a section explaining the concept of crowd computing, upon which we rely for our experiment.

2.1 Human-AI interaction

Technological advances in the last decade have allowed AI to become an extremely powerful and useful tool for solving problems. This has allowed it to become a widespread practice for automation in many domains. Some AI systems have even proven to be more performant than humans at solving certain problems. However, many practical, ethical, and judicial considerations make full automation undesirable. AI, therefore, often has a supporting role by providing predictions to the humans making the final decision. This generic process of providing predictions will be referred to as human-AI decision-making. One such way of utilizing AI systems is through Decision Support Systems (DSS). They are a subset of AI systems that aim to leverage the strengths of humans and AI systems. Throughout this thesis, when referring to AI systems, we will mainly target these systems. Compared to humans, AI systems are often very limited in the type of problems they can solve but possess superior quantitative reasoning skills. This is the ability to solve specific problems by applying mathematical concepts to analyze large amounts of data. Humans, on the other hand, outperform AI when it comes to cross-functional reasoning skills. This encompasses skills that involve reasoning based on a wide variety of areas. In complex environments, humans are often presented with an amount of data large enough to make analyzing it in its entirety impractical or impossible. They are therefore often obliged to make decisions based on their own perceptions, which are not only incomplete but are also subject to (systemic) bias. By having an AI system aggregate this data, humans can be presented with advice that relies on a much more complete view of the data. Ideally, this results in complementary team performance; performance resulting from human-AI interaction that is superior to the

performance of either one separately. However, this does not always happen, especially with well-performing AI systems.

2.1.1 Trust and AI

Many studies show that humans interact with technology in a similar fashion as they would with a human collaborator [79]. Extant research suggests that emotional and attitudinal factors that affect human-to-human interaction may likewise influence human-machine interaction. In these interactions, trust seems to be critically important. In particular, machine reliance appears to be guided by trust [57], where trusted machines and untrusted machines are likely to be accepted and rejected, respectively. We define trust as suggested by Mayer, Davis, and Schoorman [67] to be "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party". As the focus of this research is human-AI collaboration, humans will assume the role of the trustor, and AI systems the role of trustee. Many studies have shown that AI systems used as decision support systems are initially highly trusted, which decreases when erroneous behavior is observed [21, 23, 66]. Restoring this trust tends to be a lengthy process. A general tendency that is observed is a drop in trust when users are presented with AI weakness. XAI backs this up by reporting a drop in trust when low-quality models revealed weaknesses to the user [85]. Another study in this field observed that, even though machine predictions were correct, trust decreased whenever discrepancies between the user's and machine's reasoning occurred (i.e. the prediction is correct, but the human does not agree with the given explanation) [83]. The trust drop resulting from this can be mitigated by communicating the AI's rationale and abilities to the user, which lowers the impractically high initial trust and increases the recovery rate of trust after erroneous behavior is encountered [44].

2.1.2 Confidence and AI

Through interaction and feedback, humans build mental models. These are internal representations of real-world objects or concepts. When interacting with AI systems, human reliance is guided by their mental models. The way these models are built is dependent on many factors, an important one being performance feedback. Previous work shows that presenting users with machine accuracy [54, 95], information on its confidence [97], and correctness feedback [7, 8] helps build these models, thereby affecting their reliance on an AI system.

However, performance feedback is not always available, forcing users to rely on other forms of feedback. Recent work suggests confidence could be a significant influencing factor in these situations. When users agree on decisions in which they are highly confident, they tend to rely more on a model. This means that, as long as the users and AI system are sufficiently adept at performing the task they are presented with, there will be a high level of appropriate reliance. When this is not the case, however, users might underestimate machine performance, potentially leading to under-reliance. Additionally, if this occurs whenever users' and AI systems' predictions are not independent, it could lead to over-

reliance on models that are prone to the same fallacies as the user and under-reliance on complementary models.

In our research, no performance feedback will be provided to users outside of the debugging intervention. We will therefore track their confidence in their decisions, and analyze how it changes after being subjected to the intervention.

2.1.3 Appropriate Reliance

A critical factor in achieving complementary team performance is appropriate reliance. Humans appropriately rely on AI systems when they accept correct advice and reject incorrect advice from an AI system. The widespread usage of AI has incited much research exploring how to increase appropriate reliance. It has been observed that reliance can be affected by human-, automation-, and context-related factors [69, 81, 15]. In practice, however, the human's failure to appropriately rely on the system is often the cause of achieving sub-optimal performance [74, 28]. This can be described in terms of misuse and disuse, which respectively refer to relying on automation when it performs poorly and rejecting automated predictions when it is correct [57]. Figure 2.1 shows how (in)appropriate reliance occurs. To account for this sub-optimal performance, existing works propose various interventions like user-tutorials [20, 55], cognitive force functions [12], and improving the AI literacy of the use case [18]. Other works propose ways of improving the transparency of AI systems using effective explanations [54, 89], performance feedback [64], and global model properties [13]. The common denominator between these different approaches is to, in addition to AI, provide users with additional information about the AI system or change their attitude towards and knowledge of the system.

2.2 Explainable AI

Explainable AI (XAI) is a relatively new field that aims to help people understand the decisions and predictions made by AI systems. While this research does not focus on XAI specifically, it still forms an integral part of the experiments. This section will elaborate on the concept of XAI by explaining why, where, and how it is used. It will also give a brief overview of the different methods and techniques that currently exist, as well as characterize them based on aspects that are relevant in Human-AI interaction research.

2.2.1 Reasons for XAI

In general, there are four reasons for using XAI, as posed by Adadi and Berrada [1], which are to justify, control, improve, and discover.

Justify

Performant AI solutions often utilize machine learning (ML), which uses functions that are too complex to understand, effectively making them black-box models. To mitigate these effects, much work has gone into providing explanations for these models. The main focus

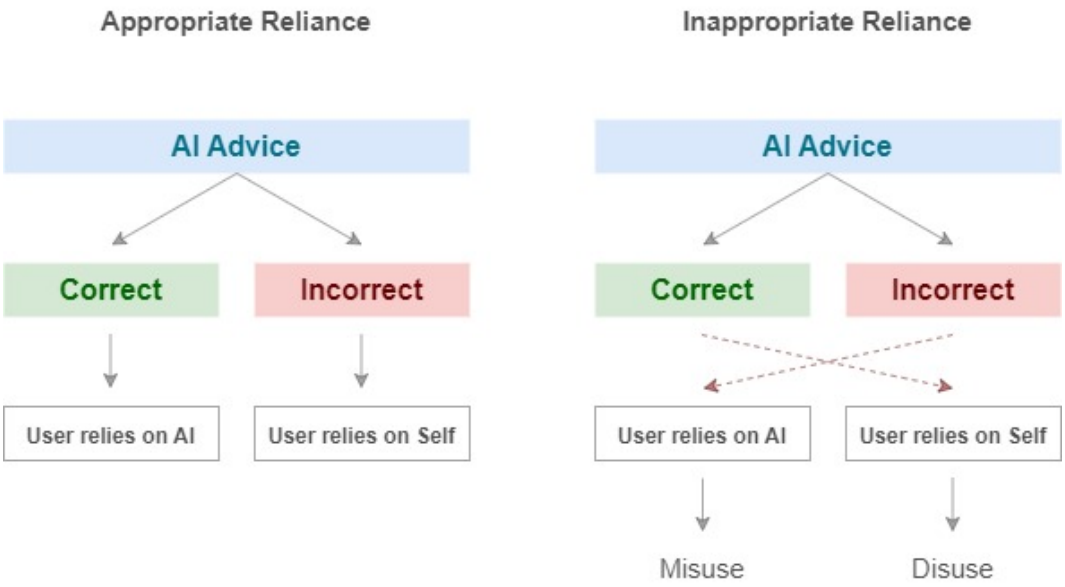

Figure 2.1: To appropriately rely on an AI system, users have to adopt AI advice when it is correct, and reject it when it is incorrect. Failing to do this will lead to misuse or disuse of the AI system.

of these explanations is to help users understand the system and the decisions it makes. Especially with potential bias and discrimination that can occur in ML models [14, 46], it becomes important that it can be shown that predictions were not made erroneously.

Control

When working with AI systems, it is always desired to make as few erroneous predictions as possible. As explanations provide information on given decisions, they can help rapidly recognize flaws and vulnerabilities. This can be paramount, especially in high-stakes domains where decisions have much impact.

Improve

Rather than just recognizing errors and justifying predictions, explanations can help improve existing systems. They can provide information that, based on the prediction, can highlight sub-optimal procedures or inconsistencies in the AI's belief system. Traditionally, this is one of the main contributions of EBHD [58].

2.2.2 Methods and Characteristics

There are several ways to distinguish XAI approaches by their methods and characteristics. This section will show them from the perspective of their complexity, level of engagement, and the extent to which they are dependent on a specific model.

Complexity

There exists a direct correlation between the interpretability and the complexity of a ML model. Generally speaking, the more complex a model is, the less interpretable it is. Because of this, it would be useful to have performant models that are intrinsically interpretable. Several works have explored this subset of models [59, 14]. Among the best-performing Machine Learning models, however, this is very uncommon. Examples of these models are transformers like XLNet and BERT [88, 94, 22]. The two state-of-the-art methods of generating explanations for these models are Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) [80, 65]. Both are so-called post-hoc explanation methods, which generate explanations after the model has been trained and rely on input perturbations to generate their explanations. These are the type of explanations that have been used in our experiments.

Level of Interpretability

A different way of differentiating between explanations is through their level of Interpretability. Extant work distinguishes between global and local interpretability.

Global interpretability focuses on the model on a holistic level, trying to explain the belief system of a model and how this correlates with all possible outputs. This information can be useful when making decisions at a higher level of abstraction (e.g. in the context of climate change). In such cases, the reasoning of the machine can provide useful insights into the dynamics of a specific area or field. Drawbacks of these models are that they are generally hard to create. This gets amplified as models become more complex.

Local interpretability focuses on explanations at the instance level. They are often unique for every input and therefore different on a case-by-case basis. While there are many methods for providing local explanations, the most popular ones are the aforementioned SHAP and LIME.

The explanations provided in our experiments are mainly local, with a small part (arguably) being global.

Model-specificity

Model-specificity is a binary property of XAI methods. It describes whether a method is model-specific or model-agnostic; whether a method can only be used on one specific model or whether it can be used independently of the model that is used. The method used in this research is model-agnostic.

2.2.3 Techniques

There are several types of techniques for presenting explanations to users. Adadi and Berrada [1] has broadly categorized them into visualization, knowledge extraction, influence methods, and Example-based explanations.

Visualization

Visual pattern recognition is something that comes naturally to humans. Much work has therefore gone into exploring ways of presenting model explanations to humans through visualizations. Three popular techniques in this area are Surrogate models, Partial Dependence Plots, and Individual Conditional Expectations. Surrogate models try to derive simplified versions of black-box models and visualize them in a way that is understandable to humans. Partial Dependence Plots aim to provide a graphical representation of complex models, where the nodes and vertices present relations between one or more input and output variable(s). Individual Conditional Expectations are a refinement of Partial Dependence Plots.

Knowledge Extraction

Well-performing ML models store their algorithms/belief systems in their network of nodes. Knowledge Extraction tries to translate (some of) that data into a human-comprehensible form. The two main techniques in existing work are Rule Extraction and Model Distillation. Rule Extraction tries to find rules that represent the decision-making process of a model, thereby being a good descriptor of how the system works. Model Distillation tries to transfer to a more shallow, easily interpretable model while retaining the key properties of the original model.

Influence Methods

Influence methods rely on perturbations of the input or internal components of a model. Explanations are derived based on the difference in output resulting from a specific (set of) perturbation(s). There are three main methods for doing this: Sensitivity Analysis, Layer-wise Relevance Propagation, and Feature Importance. Sensitivity Analysis looks at whether a model output stays stable when data is perturbed. It is often used to test a model's stability or find unimportant input attributes. Layer-wise Relevance Propagation identifies the core properties of a model by going over it, starting from the output layer. Feature Importance decides the importance of input features by permutating them and observing the differences this causes in the output.

Example-based

Example-based techniques explain machine behavior based on instances from the training data. Two examples are Prototypes and Criticisms, and Counterfactuals. Prototypes and Criticisms tries to explain models by finding instances representing the dataset (prototypes) and instances representing exceptions (criticisms). Counterfactual Explanations denote the minimum amount of conditions that would have to change before a different decision is reached.

2.2.4 Effectiveness

As XAI techniques have started to become more widespread, researchers in many areas have tried to incorporate them into their workflows. Most notable amongst these fields are healthcare, criminal justice, and military [14, 45, 47, 16, 86, 9, 40, 43], where XAI could be a promising technology. As mentioned, they aim to help people understand the decisions and predictions made by AI systems. Many agree, and believe that XAI can provide additional insights to assist decision-making [56, 89, 54, 92].

The extent to which this is the case, however, is very context-dependent. In section 2, it was mentioned that complexity can limit the techniques that can be used. Another factor that might influence this is the data type that is used by machines. Examples of data types that can be used are text, images, video, audio, and tabular data. Using the same AXI technique on these types can result in varying effectiveness. An example of this is Alqaraawi et al. [3] where attention explanations were shown to be limited in their utility on images because attention data might be noisy and confusing compared to other types like text data, where attention is presented using highlighted words.

There are several works examining the effects of XAI with respect to trust and appropriate reliance. A number of studies support the idea that explanations lead to appropriate reliance by improving human understanding of the AI model [54, 92]. Additionally, many studies support the idea that explanations arouse trust in humans. This sometimes leads to over-reliance and subsequent misuse and is therefore not always desired. Moreover, there are often covariates involved in engendering trust, like the popularity of the software generating the applications, or the visual appeal of the explanations. On the other hand, some research reports trust and reliance do not necessarily increase trust and reliance. Cheng et al. [17], for example, noticed explanations increased their subjects' machine comprehension but failed to increase trust in the system in high-stakes applications.

2.3 Critical Thinking, Belief Systems, and EBHD

The debugging intervention relies on existing work on critical thinking, belief systems, and debugging. This section will give an overview of relevant information related to those fields and explain the reasoning for proposing EBHD as an intervention.

2.3.1 Critical Thinking

Human belief systems and decision-making processes can be heavily influenced by the extent to which critical thinking is performed. This complex subject has no universally accepted definition. Different academic disciplines offer their own approaches to defining the subject. To best reflect our concerns, we rely on the work of Robert J. (1986) and define critical thinking to be the mental processes, strategies, and representations people use to solve problems, make decisions, and learn new concepts [11]. It is generally agreed upon that critical thinking has both general- and domain-specific aspects [53, 29] [31, 30, 76]. Some oppose this proposition, maintaining that critical thinking skills and abilities are not domain-specific [41, 61, 35], and others that critical thinking skills are purely domain-

specific [91, 5]. Despite this, the importance of domain-specific aspects has been established in much of the existing literature. By altering these aspects in humans, it is therefore expected to affect their critical thinking. Prior work has evaluated many domain-specific aspects of Human-AI collaboration, like knowledge [51, 10, 40] and trust [36]. Most researchers agree on the important role of domain-specific knowledge. As McPeck (1990) notes, to think critically, there needs to be something critical to think about [68]. Also, well-established is the role of trust in interaction with technology, with low trust leading to disuse and possible abuse, and high trust potentially causing over-trust which could lead to undesirable outcomes [57].

2.3.2 Debugging

According to the ANSI/IEEE standard glossary of software engineering terminology debugging is to detect, locate, and correct faults in a computer program [71]. While overall consensus has been achieved on the meaning of 'bugs' in software engineering, they have been ascribed to various meanings in ML research. These definitions range from implementation errors to particularly damaging or inexplicable test errors. Following the works of [58] we adopt the definition of Adebayo et al. [2] which defines bugs in ML to be "contamination in the learning and/or prediction pipeline that makes the model produce incorrect predictions or learn error-causing associations". Bugs that might occur include spurious correlations, labeling errors, and undesirable behavior in out-of-distribution testing. Some conflict exists in the interpretation of the process of debugging. Some consider it to be purely the identification or uncovering of model errors [75, 37]. Others complement this definition by stating that, in addition to revealing causes of problems, debugging should also fix or mitigate them [52, 96]. When designing the debugging intervention, the former interpretation was adopted. Building on this, and following [58] we define EBHD to be the process of identifying or uncovering bugs in a trained model using human feedback given in response to explanations for the model.

2.3.3 EBHD as an Intervention

Debugging is an interactive process that incentivizes efforts to become familiar with a system. By subjecting humans to a debugging intervention, we would therefore expect their belief systems of the AI system to be updated. Specifically, we aim to make users aware of the limitations of AI systems - that neither explanations of the AI advice nor the advice itself are always reliable. This could cause them to get a better sense of when to trust and rely on a system, thereby increasing appropriate reliance.

2.4 Crowd Computing

As mentioned, computers have distinct qualities that allow them to outperform humans in specific areas. Because of that, they are often utilized for the automation of a specific subset of simple and complex tasks, which is very practical when dealing with large workloads. However, due to the nature of some tasks, automation in this manner might not be a viable

option and a different solution has to be come up with. One of these solutions is crowd computing. Crowd computing is a variation of the more common paradigm crowdsourcing, where a job is traditionally performed by a designated agent (usually an employee) and outsourced to an undefined, generally large group of people in the form of an open call [77]. In crowd computing, the job is often a problem or collection of problems to be solved. These problems have characteristics that make the use of machines for automation impossible or impractical. At the same time, they can be easily solved by humans. This low threshold causes the group to which the task can be outsourced to be very large. One of the areas where this can be capitalized on is research into human-AI interaction. Researchers create an AI system with which users can interact and present the system to users through a series of tasks, questions, or comparable approaches. Through dedicated platforms, they can reach a large audience which they can filter, based on the parameters of their research. After the selection, they will present their system to the participants, which will be appropriately compensated for their efforts. To ensure that the participants supply them with valid data for their research, additional filtering is applied to remove participants that do not engage seriously with the presented task. The most popular approach for this filtering is presenting users with attention checks. Attention checks are easy-to-answer questions that have to be answered correctly for the user to be allowed to continue their tasks. Incorrectly answering such a question often indicates a lack of attention from the participants and notifies the researcher of this. We also utilized crowd computing in our research, to analyze the effects of our intervention. These consisted of instructions disguised as normal tasks or questions. Figure 2.2 shows an example of one such attention check. The question in the middle gives instructions to the participant. If they are not paying attention, participants will ignore the instructions and select an answer at random, likely failing the attention check.

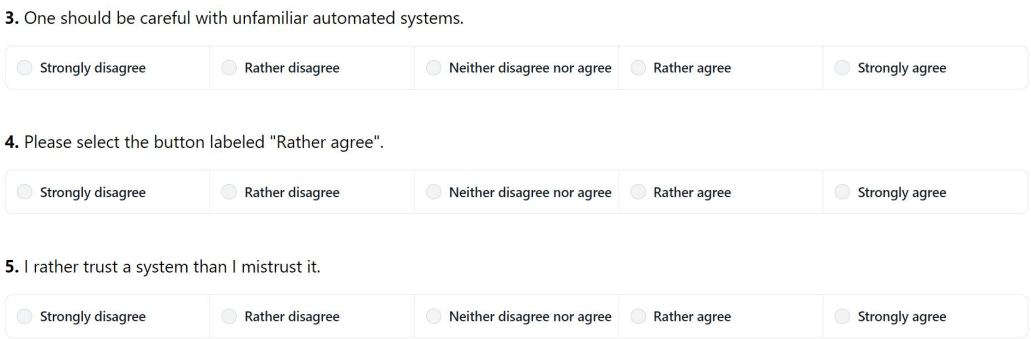

Figure 2.2: Several questions are shown. The first and the last are regular questions, while the question in the middle gives instructions to the participant. If they are not paying attention, participants will ignore the instructions and select an answer at random. When they do this, they will likely fail the attention check.

Chapter 3

Approach

This section outlines the hypotheses, describes the task that we use and presents the design of the debugging intervention.

3.1 Hypotheses

The experiment has been designed to reveal the impact of the proposed EBHD intervention on user estimation performance (i.e. how well the user is able to estimate what its own, as well as the AI system's performance, is on a set of tasks), as well as user reliance on AI systems. By having users challenge AI advice and explanations, and presenting them with feedback on their performance, it is expected that they will have an improved understanding of the system. This could improve their sense of the accuracy of the AI system, which impacts both estimation performance and reliance and increases the extent to which they appropriately rely on the AI system. We therefore hypothesize:

1. Encouraging users to critically evaluate the trustworthiness of AI advice at the instance level, in a debugging intervention, will improve their assessment of the AI system's performance at the instance and global levels.

2. Encouraging users to critically evaluate the trustworthiness of AI advice at the instance level in a debugging intervention will improve the extent to which users appropriately rely on the system.

As the internal task ordering of the intervention can affect its effectiveness, our experiment will consist of various conditions. These conditions will be based on the order in which the AI system's strengths and weaknesses are presented. Dependent on the ordering, users may show different learning effects. We, therefore, hypothesize that:

1. The perceived trustworthiness of AI advice at the instance level in a debugging intervention corresponds to an ordering effect with respect to appropriate reliance.

3.2 Task: Deceptive Review Detection

AI-assisted decision-making is typically used for tasks that are challenging for humans. AI systems often outperform humans in these tasks. Our experiment is based on one such task, called deceptive review detection, where AI advice could be a real need. In each task, participants are presented with a hotel review, which is either genuine or deceptive. Genuine reviews have been written by people that have actually visited the hotel and described their experience there. Deceptive reviews have been written by people that have not visited the hotel and are therefore writing about made-up, fake experiences. It is the task of the participant to decide whether the review they are presented with is genuine or deceptive.

3.2.1 Highlights as Explanations

During the task, users are not provided with any performance feedback, reflecting a real-world scenario. To support them in their decision-making, we provide them with predictions from an AI system, along with explanations in the form of highlights. The AI system is based on a fine-tuned BERT model and the highlights have been generated using LIME. The machine prediction only states whether the machine believes the review to be genuine or deceptive, without providing any additional feedback, like accuracy or confidence. Each highlight occurs on a single word and is either colored green or red, respectively stating that the machine believes a word to be more indicative of the review being more genuine or deceptive. Highlighted words that are more indicative will be highlighted with a more intense hue. Highlights do not occur in every word. Only the 10 most indicative words will be highlighted.

3.2.2 Two-stage Decision Making

Following existing empirical study design of human-AI decision-making [55], participants will complete each task in two stages of decision-making. In the first stage, they will be presented with the hotel review, without any AI assistance, meaning no predictions or explanations are shown. The decision made in the first stage - the initial decision - will therefore be completely dependent on the participant. In the second stage, participants are presented with the same review. Now, however, both machine prediction and explanations will be provided. Now, participants can make their decision using AI advice. As the task can be quite challenging, participants will be provided with guidelines to complete the tasks. These guidelines will be presented before each task batch. Additionally, there will be a button during the tasks through which they can access the guidelines.

3.2.3 Task UI

Figure 3.1 shows the task interface that users will initially be presented with. As can be seen, this version does not have any explanations. Users can select their answer by clicking the *genuine* or *deceptive* buttons. After they have provided their answer, participants are asked how confident they are in their decision. This can be indicated by 5 buttons presenting a

5-point Likert scale ranging between *Very unconfident*, *Rather unconfident*, *Neutral*, *Rather confident* and *Very confident*.

Figure 3.2 shows the task interface of the second stage. The review is the same, however, now the AI advice and explanations are also shown. Again, participants have to indicate their decision and confidence before moving on to the next task instance.

At any point in the process, participants can access the highlights using the button at the top of the interface.

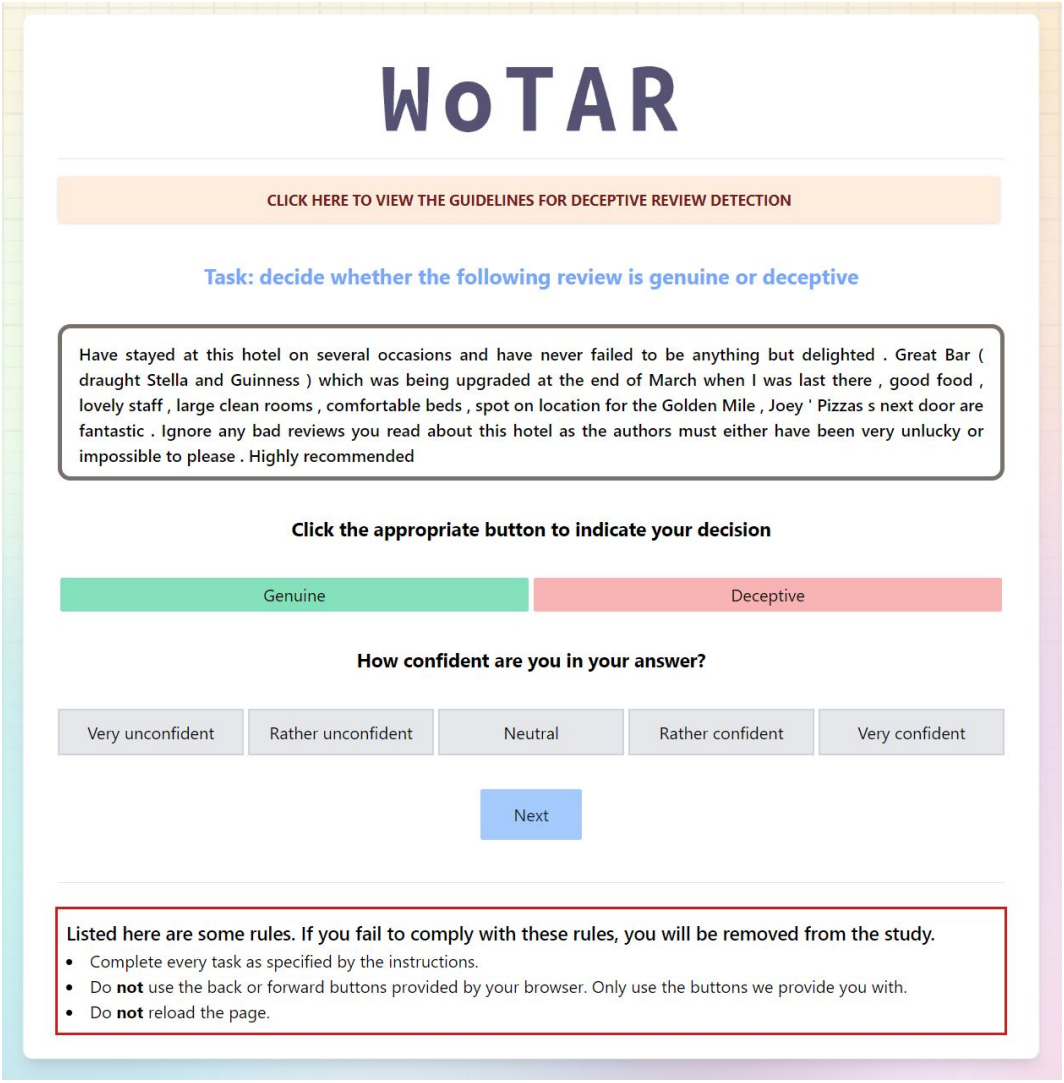

Figure 3.1: Task UI of the first stage of a trial case.

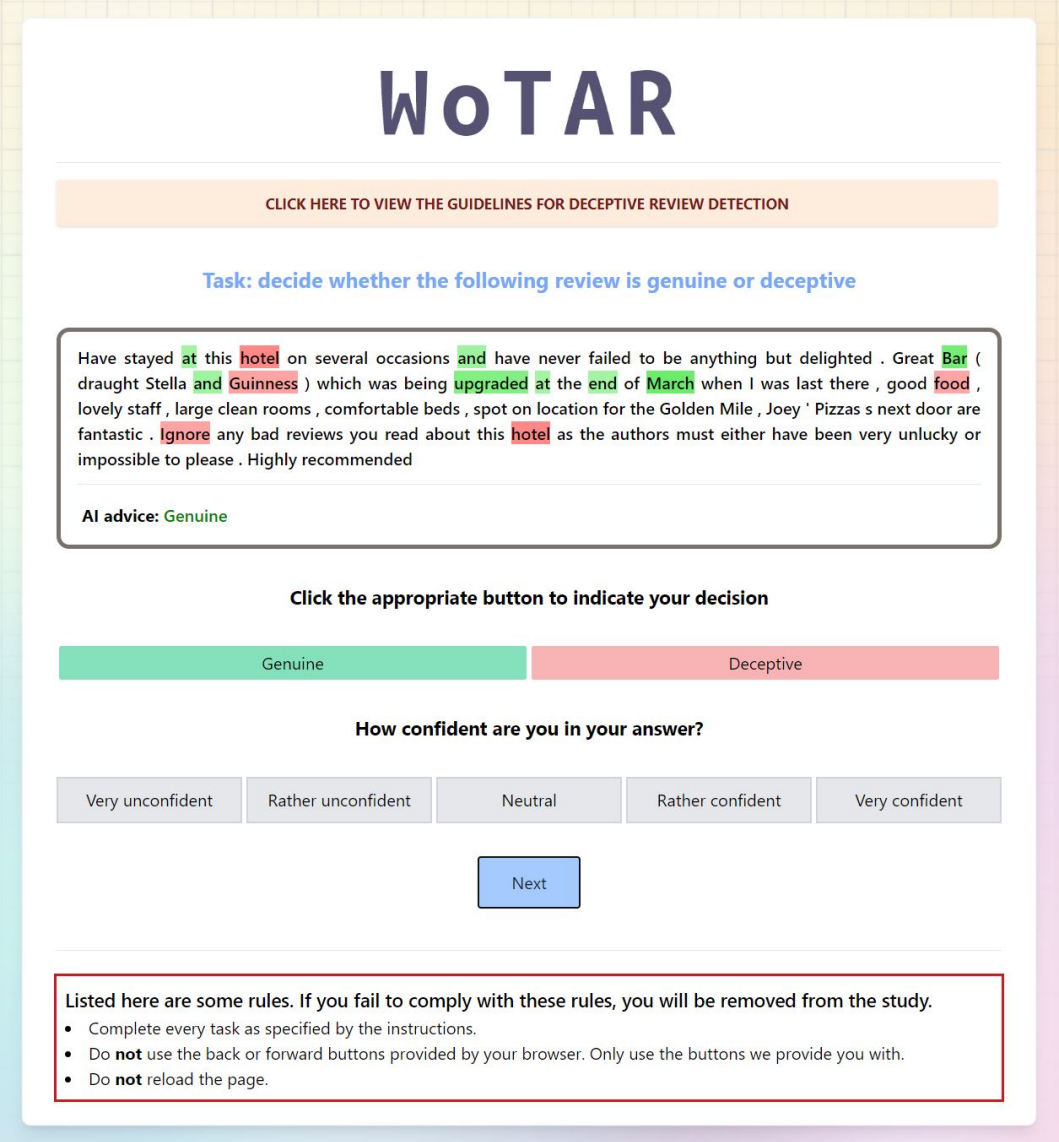

Figure 3.2: Task UI of the second stage of a trial case.

3.2.4 Task Selection

Using deceptive review detection as the task for our experiment was inspired by Lai, Liu, and Tan [55], which used the same task. To train the model and create the task batches, we also used the same dataset as they did, the Deceptive Opinion Spam Corpus v1.4 [72, 73].

To measure the effects of the debugging intervention, participants will be presented with task batch before and after the debugging intervention (more on this in section 4.4). To make sure the result is not influenced by the difficulty of the tasks in the different batches, we created two batches of tasks with equal difficulty. We did this by conducting a pilot

study where 10 participants were presented with 20 different deception detection tasks that were randomly sampled from the validation and test set of the dataset. Each task only had the initial stage (without any assistance). Based on the average performance of each task, we created two balanced batches of tasks.

3.3 Debugging Intervention

Our debugging intervention relies on EBHD to help participants more accurately assess the trustworthiness of AI advice at the instance level and calibrate their reliance on the AI system.

3.3.1 Setup

The aim of the intervention is to learn participants that (1) AI advice they are presented with is not necessarily correct and (2) explanations are not always informative and helpful for identifying the trustworthiness of AI advice. Thus, when selecting the task batch for the debugging intervention we considered two main factors: (1) the correctness of AI advice, and (2) whether an explanation is informative (i.e., whether or not such explanations combined with guidelines, can help participants easily identify the correct answer). Taking this into account, we selected 8 tasks for the debugging intervention.

3.3.2 Task: Explanation-Based Human Debugging

During each task in the debugging intervention, participants are presented with a hotel review with explanatory elements consisting of AI advice in the form of a prediction and color-coded highlights on the 10 most important features. This can be seen in figure 3.3, where the interface of a debugging question is shown. Each highlight corresponds, based on its color, to the contribution of their token to the model prediction. This contribution is presented by a 5-point Likert scale ranging between *Deceptive*, *Somewhat deceptive*, *Neutral*, *Somewhat genuine*, and *Genuine*. Participants are instructed to read the text and, when deemed necessary, refine the explanations by adjusting the highlights using the panel on the right. Additionally, they have to indicate whether they think the AI is correct. When they have done this, they will be presented with a screen providing them with feedback on their performance, see figure 3.4. The performance feedback consists of every inconsistency between the participant's answers and the correct answers. The highlights that participants will initially be presented with are the highlights resulting from post-hoc XAI method LIME. The correct highlights have been created through manual correction of the researchers and serve as ground truth for the debugging intervention.

3.3.3 Ordering Effect

As mentioned, internal ordering based on the showing AI strengths and weaknesses can have an impact on the effectiveness of the debugging intervention. To balance the tasks, we manually selected four tasks with informative explanations (where explanations and

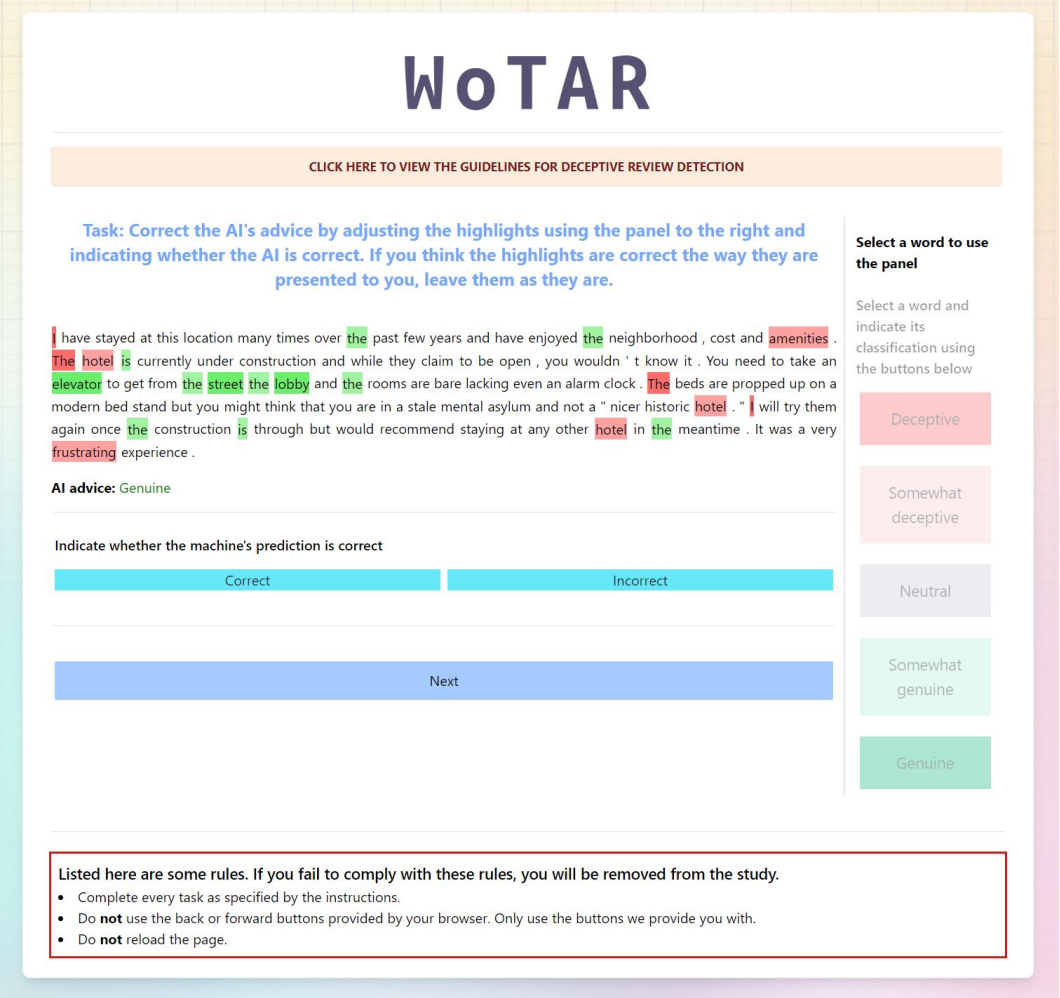

Figure 3.3: Task UI a trial case from the debugging intervention. Pressing a highlighted word will make the panel on the right-hand side of the interface accessible. This panel can then be used to indicate the color of the selected highlight.

guidelines can help participants easily identify the correct answer) and four tasks with uninformative explanations. The eight tasks presented in our debugging phase are:

- two tasks with correct AI advice and informative explanations
- two tasks with correct AI advice and uninformative explanations
- two tasks with incorrect AI advice and informative explanations
- two tasks with incorrect AI advice and uninformative explanations

The tasks are balanced based on the correctness of the AI advice and whether their explanations are informative. The (un)informative explanations have been manually selected,

whereas the correctness of the AI advice is determined randomly. Existing work suggests first impressions greatly affect human trust in AI systems (see section 2.1.1). With this in mind, we created three task batches consisting of the same tasks but ordered differently. The aim of the ordering is to induce positive and negative first impressions of the AI system. The task batches were ordered in the following way:

- Random order.
- Decreasing impression order (*i.e.* from good to bad): correct AI advice, informative explanation → correct AI advice, uninformative explanation → wrong AI advice, informative explanation → wrong AI advice, uninformative explanation.
- Increasing impression order (*i.e.* from bad to good): wrong AI advice, uninformative explanation → wrong AI advice, informative explanation → correct AI advice, uninformative explanation → correct AI advice, informative explanation.

In the decreasing order, participants will first be presented with correct advice and informative explanations, after which the quality of the advice and explanations will gradually deteriorate. It is expected that this will result in a positive first impression of the AI system on participants. In the increasing order, we expect the opposite to happen; presenting participants with incorrect advice and incorrect explanations first is expected to result in a negative first impression of the AI system.

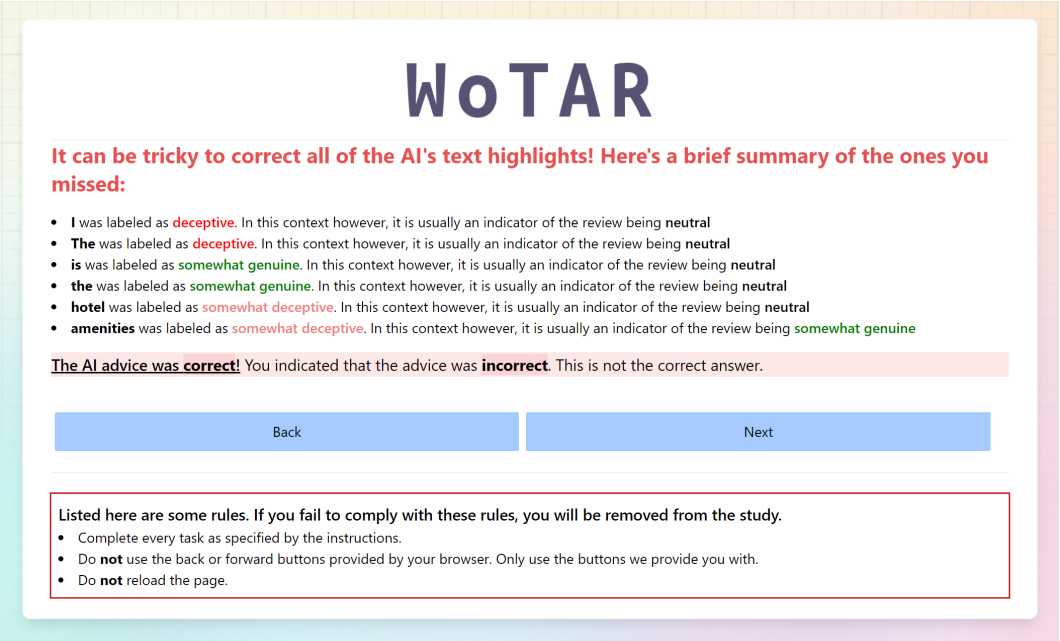

Figure 3.4: After adjusting the highlights, users are presented with feedback on their performance.

Chapter 4

Experimental Setup

This section will cover the experimental, measures, participants, and procedure of the study. This study was approved by the human research ethics committee of our institution.

4.1 Experimental Conditions

Throughout the study, every participant worked on two batches of the deceptive review detection task with the two-stage decision-making as described in 2. The difference between the conditions is whether the debugging condition is present and what its internal ordering is. The four experimental conditions resulting from this are:

1. **Control** No debugging intervention is present
2. **Debugging-D** A debugging intervention is present, with a decreasing impression order
3. **Debugging-I** A debugging intervention is present, with an increasing impression order
4. **Debugging-R** A debugging intervention is present, with a random impression order

In conditions with a debugging intervention present, the procedure described in section 4.4 took place for the 8 tasks in the debugging intervention. In the control condition, these tasks were presented similarly as in the regular task batches (section 4.4). The tasks were randomized within the limits that the balancing procedure (section 3.3.3) allowed for.

4.2 Measures

To verify our hypotheses, we selected several measures. This section will outline what these measures are and what they represent.

4.2.1 Measures Hypothesis 1

To verify hypothesis 1, we need to be able to assess the performance estimation of humans at the global and instance level.

To measure the performance estimation at the global level, participants were asked two questions after each regular task batch (non-intervention batch): "From the previous 10 tasks, on how many tasks do you estimate the AI advice to be correct?" and "From the previous 10 tasks, how many questions do you estimate to have been answered correctly?". By comparing the actual performance of the AI and participant to the estimates of the participant, we can calculate the miscalibration between the two. The actual performance of the AI has been set to 0.8, meaning that out of 10 tasks, it provides the correct advice on 8 of them. The actual performance of the participant is measured throughout the experiment. Using this information, we calculate the Miscalibration of AI Performance (MAP) and the Miscalibration of Team Performance (MTP).

To measure the performance at the instance level, we use the participant's indicated confidence. When making an AI-assisted decision with high confidence, a participant implicitly indicates whether it believes the machine to be correct or incorrect. We, therefore, measure the estimation performance at the instance level by filtering out the AI-assisted tasks where participants indicated they were "Very confident" and calculating the fraction of questions that were answered correctly. We refer to this measure as Correct Confident Answers (CCD)

4.2.2 Measures Hypothesis 2

To verify hypotheses 2 and 3, both reliance and appropriate reliance of the participants on the AI system was measured.

Measuring reliance was done through the Agreement Fraction and Switch Fraction. These metrics are widely adopted in existing literature [95, 97, 64]. The former is the extent to which participants agree with the given advice. The latter represents how often participants adopt AI advice when it is in disagreement with their initial answer.

d_i	**AI advice**	d_f	**Reliance**
0	1	1	Positive AI reliance
0	1	0	Negative self-reliance
1	0	1	Positive self-reliance
1	0	0	Negative AI reliance

Table 4.1: The different appropriate reliance patterns considered in [82]. d_i and d_f refer to the initial human decision and final human decision respectively. 1 and 0 refer to correct and incorrect respectively.

Appropriate reliance was measured through the measures proposed in Schemmer et al. [82]. They base their measures on four reliance patterns: negative AI reliance, positive AI

reliance, negative self-reliance, and positive self-reliance. These refer to the cases where users rely on or don't rely on either themselves or the AI system. Figure 4.2.2 shows an overview of these four patterns. The proposed measures are calculated in the following way:

$$\textbf{Relative positive AI reliance (RAIR)} = \frac{\text{Positive AI reliance}}{\text{Positive AI reliance + Negative self-reliance}}$$

$$\textbf{Relative positive self-reliance (RSR)} = \frac{\text{Positive self-reliance}}{\text{Positive self-reliance + Negative AI reliance}}$$

RAIR is the extent to which a participant relies on correct AI advice when the participant's initial decision is incorrect. RSR is the extent to which a participant relies on their own, correct, initial decision when the AI system provides incorrect advice. Both are presented with a number between 0 and 1. The closer these measures are to 1 for a participant, the more that participant appropriately relies on the AI system. Also, for each batch, AI-assisted accuracy is considered to measure participants' performance.

For a deeper analysis of our results, several additional measures were considered based on observations from existing literature [60, 83, 87]:

- The Trust in Automation (TiA) questionnaire [49], a validated instrument to measure (subjective) trust [87] consisting of 6 subscales: *Reliability/Competence* (TiA-R/C), *Understanding/Predictability* (TiA-U/P), *Propensity to Trust* (TiA-PtT), *Familiarity* (TiA-Familiarity), *Intention of Developers* (TiA-IoD), and *Trust in Automation* (TiA-Trust).
- The Affinity for Technology Interaction Scale (ATI) [33], administered in the pre-task questionnaire. Thus, we account for the effect of participants' affinity with technology on their reliance on systems [87].
- The NASA-TLX questionnaire [42] for the working load assessment of the debugging intervention.

4.3 Participants

4.3.1 Sample Size Estimation

To estimate the number of participants needed for the experiment, we did a power analysis for a Between-Subjects ANOVA using G*Power [32]. We used a significance level of 0.05, on which we applied a Bonferroni correction. Because the experiment was used to test 3 hypotheses, the resulting alpha level was $\frac{0.05}{3} = 0.017$. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a statistical power of $(1-\beta) = 0.8$, and that we will investigate 4 different experimental conditions. The resulting required sample size was 230. To accommodate potential exclusion, we recruited 324 participants from the crowdsourcing platform Prolific [footnote].

4.3.2 Compensation

The estimated completion time for the experiment was 30 minutes. We maintained an hourly wage of £7.6. The participants were therefore rewarded with £3.8 for their efforts. Additional bonuses were handed out based on their performance during the 20 trial cases. Every correct (final) decision was rewarded with £0.05. This bonus was handed out to incentivize the crowd workers to try their best on each task. This approach is widely adopted in existing research [18, 55].

4.3.3 Filter Criteria

We selected proficient English-speaking participants, all above the age of 18. Each of them had an approval rate of at least 90%, and more than 80 successful submissions on the Prolific platform. At the start of the experiment, participants were required to read the basic introduction and guidelines about the deceptive review detection task. After that, they were presented with two easy questions that followed directly from the introduction. Additionally, during the following two steps, before participants had made a significant time commitment, participants were given two simple instructions (select a specific answer to a multiple-choice question). Failing to answer the questions correctly or complete the tasks lead to a direct removal from the study. The questions and tasks were designed to be sufficiently easy that any participant should have been able to answer them correctly if they had read the introductions carefully. Answering any of them wrong or failing to complete the tasks correctly, therefore, indicates to us that they have not read the instructions carefully and, therefore, were not paying attention. This means their data could not be used for our study, which is why their removal was justified. 90 participants were filtered out in this step. The resulting sample of 234 participants had an average age of 39 ($SD = 13$) and a gender distribution of (48.7% female, 49.6% male, 1.7% other).

4.4 Procedure

A visual representation of the procedure can be found in figure 4.1. As mentioned, at the start of the experiment participants are presented with a basic introduction to the deceptive review detection task. According to Lai, Liu, and Tan [55], guidelines about how to identify deceptive reviews are highly useful in improving user performance on this task, which is why these were also provided to the users in the introduction. Following the introduction, the participants were presented with the qualification check, where they had to answer two questions about the introduction correctly to ensure they are paying attention and understand the task. Participants that passed this test were presented with the (1) TiA-PtT and TiA-Familiarity part of the TiA questionnaire, and (2) the ATI questionnaire. Following these questionnaires was the first of two batches of tasks (10 tasks per batch). Which of the two batches was presented to the participant was randomly decided. Additionally, the task order within each batch was also randomized. After finishing their first batch of tasks, the participants had to answer two questions related to the measurement of their performance estimation (see section 4.4) and answer a post-task questionnaire consisting of the remaining

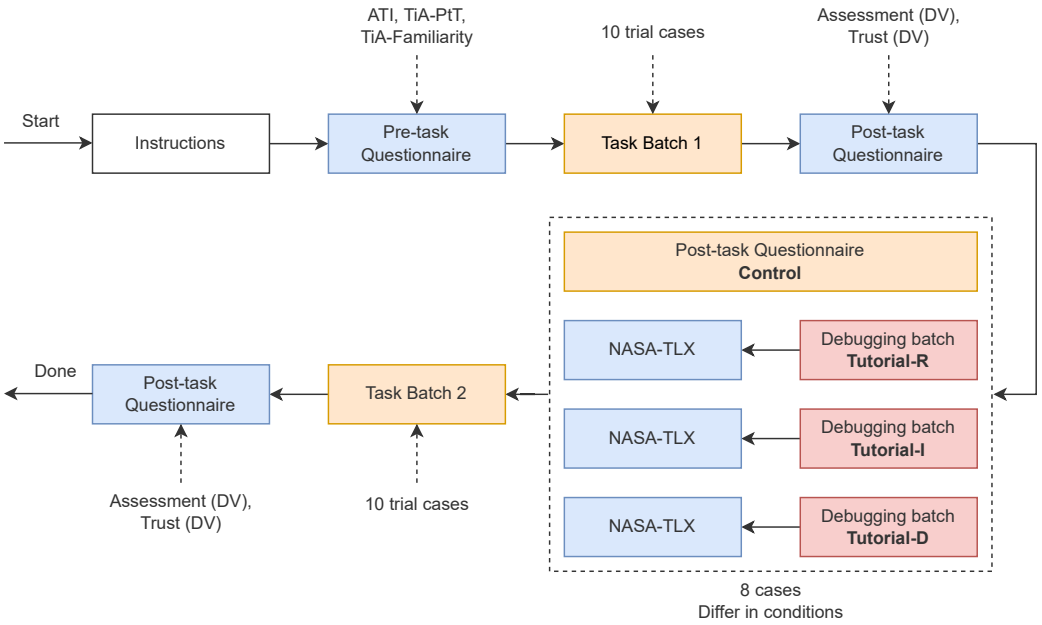

Figure 4.1: Illustration of the procedure that participants followed within our study. The blue boxes represent questionnaire phases, the orange boxes represent task phases, and the red box represents the debugging intervention.

sections of the TiA questionnaire. Following this was the debugging intervention. For participants in **Control**, this meant they had to complete another 8 tasks in the same fashion. Participants not in **Control** were subjected to the debugging intervention as described in section 3.3, after which they had to answer a NASA-TLX questionnaire assessing the workload they experienced during the intervention. After the intervention, participants have to complete the second batch followed by the same questions and post-task questionnaire as the ones after the first batch.

Chapter 5

Results and Analysis

This section will present the main results of the study. It starts by stating some descriptive statistics. These are followed by an analysis of the hypotheses and their measures (see section 4.2. Finally, an explorative analysis of the data will be covered.

5.1 Descriptive Statistics

As a measure of participant reliability [34], we only consider participants who passed all attention checks. Participants were distributed in a balanced fashion over the four experimental conditions in the following way: 57 (Control), 59 (Debugging-R), 60 (Debugging-D), and 58 (Debugging-I). On average, they spent 51 minutes (SD = 14) on our study.

Variable Distribution. The distribution of the covariates is as follows: *ATI* ($M = 3.91$, $SD = 0.94$, 6-point Likert scale, and *1: low, 6: high*), *TiA-PtT* ($M = 2.89$, $SD = 0.61$, 5-point Likert scale, *1: tend to distrust*, *5: tend to trust*), *TiA-Familiarity* ($M = 2.29$, $SD = 1.09$, 5-point Likert scale, *1: unfamiliar with AI system used in study*, *5: familiar with AI system used in study*).

The working load of the debugging intervention is measured with the NASA-TLX questionnaire (on a scale of [-7, 7]). For all dimensions except "Performance", a higher value indicates a higher working load. In the dimension "Performance", a smaller value indicates a higher estimated performance on tasks. The dimensions have been visualized in figure 5.1. In general, participants think the debugging intervention requires a high amount of "Mental Demand" and "Effort", but a low amount of "Physical Demand" and "Temporal Demand". Most participants do not show high expectations in achieved "Performance". They also don't show much "Frustration".

Performance Overview. The average accuracy achieved by the participants was 0.64 (SD = 0.11) over the two task batches. This is lower than the aforementioned AI accuracy of 0.8. The agreement fraction is 0.66 (SD = 0.13), while the switching fraction is 0.31 (SD = 0.22). This means that participants did not always switch to the AI advice in cases of initial disagreement. Participants, therefore, did not blindly rely on the AI system. In the two batches of tasks (10 for each batch), the average estimated AI performance is 5.81 (SD = 1.91) and 5.79 (SD = 1.71) respectively; the average estimated team performance is

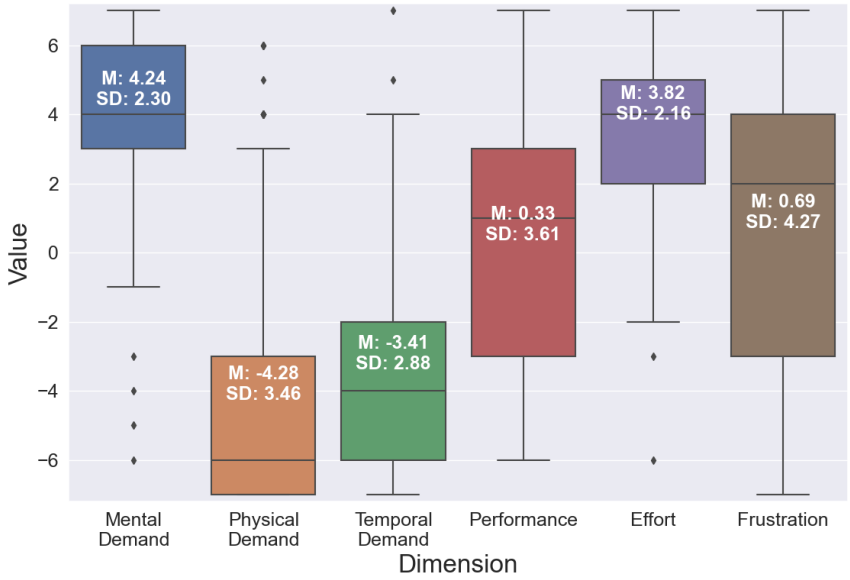

Figure 5.1: Box plot illustrating the distribution of the different dimensions in NASA-TLX questionnaire. *M* and *SD* represent mean and standard deviation respectively.

6.64 (SD = 1.74) and 6.44 (SD = 1.87) respectively. Overall, participants underestimated the performance of the AI system and believed they could outperform the AI system on this task after receiving AI advice.

5.2 Hypothesis Tests

This section will check for each hypothesis whether it is supported.

5.2.1 H1: The effect of a critical evaluation setting on AI performance estimation

To verify H1, the Wilcoxon signed-rank test was used to compare all assessment-based dependent variables of participants before and after the debugging intervention. Participants in **Control** are not considered in this comparison. The results are shown in table 5.1. Only in condition **Debugging-D**, participants showed a significant difference in team-performance estimation. Using Post-hoc Mann-Whitney tests to make pairwise comparisons of performance revealed no significant differences. Thus, H1 is **not** supported.

5.2.2 H2: the effect of a critical evaluation setting on appropriate reliance

To verify H2, we also used the Wilcoxon signed rank test to compare all reliance-based dependent variables of participants before and after the debugging intervention. The results

Condition	Debugging		Debugging-R		Debugging-D		Debugging-I	
DV	*T*	*p*	*T*	*p*	*T*	*p*	*T*	*p*
MAP	3833	.677	363	.516	358	.475	457	.347
MTP	4006	.085	512	.215	324	**.016**†	375	.320
CCD	3261	.493	388	.566	325	.680	391	.794

Table 5.1: Wilcoxon signed ranks test results for **H1** on AI performance estimation. "†" indicates the effect of the variable is significant at the level of 0.017 (adjusted alpha).

are shown in figure 5.2. Overall, no statistical difference in reliance is observed when comparing all conditions with the debugging intervention. By performing the post-hoc Mann-Whitney test on the accuracy, we found that after the debugging intervention, the accuracy drops significantly. For a more fine-grained analysis, we further conducted the Wilcoxon signed rank tests on each condition with the debugging intervention. As can be seen, participants in **Debugging-I** show a significant difference in RAIR, while no significant difference is found with the post-hoc Mann-Whitney test. The observed results do **not** support H2.

Condition	Debugging		Debugging-R		Debugging-D		Debugging-I	
Dependent Variables	*T*	*p*	*T*	*p*	*T*	*p*	*T*	*p*
Accuracy	3659	**.004**†	487	.096	403	.130	351	.059
Agreement Fraction	4993	.207	676	.722	512	.950	482	.058
Switch Fraction	4969	.093	573	.530	724	.870	409	.042
RAIR	4246	.039	454	.341	656	.764	321	**.010**†
RSR	2162	.528	241	.650	155	.095	292	.921

Table 5.2: Wilcoxon signed ranks test results for **H2** on reliance-based dependent variables. "†" indicates the effect of the variable is significant at the level of 0.017 (adjusted alpha).

While no significant difference is found in the reliance-based measures as a consequence of the debugging intervention, in general, we did witness a drop in reliance-based measures: Accuracy (0.67 $\rightarrow$ 0.63), Agreement Fraction (0.68 $\rightarrow$ 0.66), Switch Fraction (0.34 $\rightarrow$ 0.28), RAIR (0.38 $\rightarrow$ 0.30), RSR (0.64 $\rightarrow$ 0.61). This is very obvious on condition Debugging-I: Accuracy (0.68 $\rightarrow$ 0.63), Agreement Fraction (0.71 $\rightarrow$ 0.66), Switch Fraction (0.39 $\rightarrow$ 0.29), RAIR (0.43 $\rightarrow$ 0.29), RSR (0.59 $\rightarrow$ 0.61). When AI advice disagrees with users' initial decisions, users tend to rely on themselves more than they should. This results in decreased (appropriate) reliance and accuracy. In the deceptive review detection tasks, the AI system performs generally better than the participants. The reduced reliance may help explain why, on average, we found a decrease in accuracy.

5.2.3 H3: ordering effects of debugging tasks

To verify H3, we compared (1) the difference of reliance-based dependent variables between batch 1 and (2) the user reliance on the second batch with participants of all conditions using

the Kruskal Wallis test. No significant difference was found. The task working load was compared by conducting the Kruskal-Wallis H-test on the six measures in the NASA-TLX questionnaire. Again, no significant difference is found. H3, therefore, is **not** supported.

In order to take a deeper look at the ordering effect of the debugging tasks and see how it affects the final performance of participants, some additional analysis was performed. We looked at the participants who achieved an accuracy level of over 80% in the second task batch and filtered out participants that blindly relied on the AI system. Among these participants, we found that the number of participants in condition **Debugging-D** (14) is clearly more than in condition **Debugging-R** (9) and **Debugging-I** (9). In comparison, the number of participants achieving this level of accuracy in **Control** is 11. Although the ordering effect does not show a significant statistical difference, such an observation lends partial support to **H3**.

5.3 Explorative Study

In addition to the hypothesis testing, we also performed an explorative study.

5.3.1 Trust analysis

To explore the effects of the debugging intervention on user trust in the AI system, we compared the trust before and after the intervention using the Wilcoxon signed rank test. No significant difference was found in the test results. This suggests that the debugging intervention can calibrate user reliance and estimation of AI performance without directly shaping their trust.

5.3.2 Covariates Impact on Trust and Reliance

To analyze the impact of covariates on user trust and reliance, Spearman rank-order tests were conducted with the covariates and average trust and reliance-based dependent variables on the two batches of tasks. The results show that propensity to trust (**TiA-PtT**) shows significant positive correlations with the following trust-based measures: **TiA-R/C** ($r(234) = 0.270, p = .000$), **TiA-U/P** ($r(234) = 0.165, p = .011$), **TiA-IoD** ($r(234) = 0.234, p = .000$), **TiA-Trust** ($r(234) = 0.303, p = .000$). Additionally, **ATI** shows a significant positive correlation with the reliance-based measure **Agreement Fraction** ($r(234) = 0.159, p = .015$).

5.3.3 Confidence Analysis

The confidence dynamics of the four different conditions are shown in figure 5.2. On average, participants show positive confidence (above neutral) in their final decisions. Conditions **Debugging-I** and **Debugging-R** show an initial decrease in confidence, but this level comes back to average soon, after which it roughly stays at that level. In contrast, participants in condition **Debugging-D** showed increased confidence after the debugging intervention and keeps relatively stable compared to the other conditions.

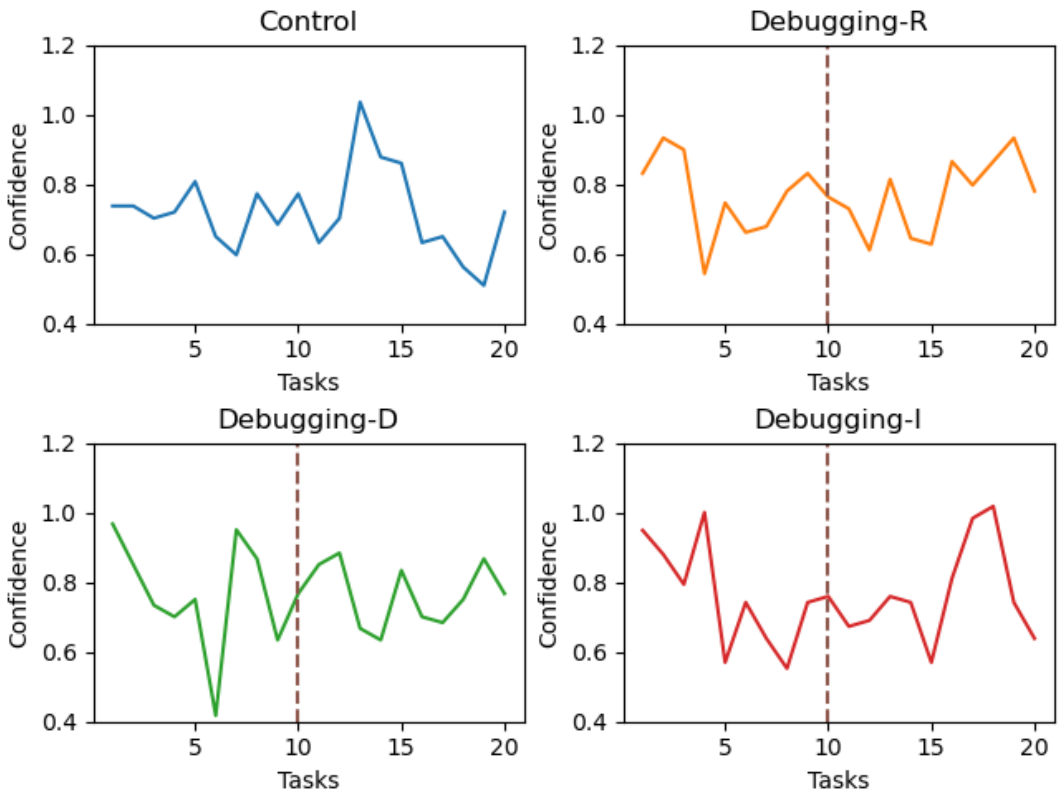

Figure 5.2: Illustration of dynamics of confidence change in the 20 tasks of each condition. The purple dashed line represents the debugging intervention.

We calculated the confidence change after receiving AI advice based on nine different reliance patterns: whether the initial decision agrees with AI advice, whether the final decision agrees with AI advice, switch behavior, and the four reliance patterns considered in calculating appropriate reliance (section 4.2.2). The results are shown in 5.3. They show that, generally, participants exhibit increased confidence when there was an agreement between their initial decision and the AI advice. When their initial decision did not agree with the AI advice, they exhibited a decrease in confidence. Even when switching to the AI advice, given initial disagreement, participants tend to show decreased confidence in their final decision. Considering the four reliance patterns, participants generally show a confidence drop. This drop becomes more severe when participants insist on their own decision, instead of adopting AI advice.

Pattern	**Dependent Varibles**	*M*	*SD*
Reliance	**Initial agreement**	0.37	0.75
	Initial disagreement	-0.41	0.99
	Final agreement	0.23	0.90
	Final disagreement	-0.44	0.90
	Switch behavior	-0.32	1.18
Appropriate Reliance	**Positive AI reliance**	-0.34	1.17
	Negative AI reliance	-0.23	1.18
	Positive self-reliance	-0.41	0.88
	Negative self-reliance	-0.47	0.89

Table 5.3: Reliance and confidence correlation.

Chapter 6

Discussion

6.1 Key Findings

We proposed a debugging intervention to promote appropriate reliance on AI systems. We expected it to improve user estimation performance and calibrate user reliance so that they would more appropriately rely on the AI system. The way the debugging intervention intended to do this was by showing participants that AI systems are not always reliable, and their explanations are not always informative. The results from the experiment don't provide support for this; such a debugging intervention fails to calibrate participants' estimation for AI performance at both the global and local levels. The results also show that, after the debugging intervention, participants tend to rely less on the AI system.

Although almost no significant difference was found between the different ordering of debugging tasks, early exposure to AI weakness caused participants to show a more obvious tendency to disuse the AI system. This under-reliance was found to result in sub-optimal team performance. This is in line with recent work, which found that a bad first impression of an AI system can lead to an underestimation of AI competence and reduced reliance on the system [36].

In further analysis of covariates, we found that, in general, the sub-scale propensity to trust shows a positive correlation with all other trust sub-scales. However, no significant correlations were found between the propensity to trust and reliance, which indicates that the increased trust due to the propensity to trust does not translate to reliance behaviors. Meanwhile, the confidence dynamics in different reliance patterns showed that AI advice may amplify the confidence of user decisions when in agreement, and decrease user confidence when in disagreement. Under disagreement, users appear to rely more on themselves (indicated by a confidence decrease), as opposed to adopting AI advice.

6.2 Implications

The findings suggest that the debugging intervention and similar interventions with training purposes (e.g., user tutorials) may suffer from the cognitive bias brought by the ordering effect within such interventions. To show users both the strengths and weaknesses of AI

systems, these should be presented to them in a balanced fashion. Too much exposure to AI weakness could leave users with a bad expression, causing them to disuse the AI system, and should be avoided. At the same time, participants in our study tend to be optimistic about the team performance while underestimating the AI performance. This is possibly caused by meta-cognitive bias — the Dunning-Kruger effect [50]. To promote appropriate reliance, helping participants understand AI systems' strengths and weaknesses is not enough. Participants also need to be made aware of their own strengths and weaknesses. These findings can form guidelines for future designs of training interventions intended to promote appropriate reliance.

The study also shows that the reliance patterns (e.g. agreement, disagreement) have a clear correlation with the confidence dynamics of its participants. When there is disagreement between participants' initial decision and the AI advice, a decrease in confidence is observed. And compared to insisting on their own decision, participants have more confidence when giving agency to AI advice. Such observations may be a dangerous signal for appropriate reliance, as they indicate users' and AI systems' predictions are not independent. Further research is required to explore how to cope with confidence dynamics that emerge when users are exposed to AI disagreement.

6.3 Limitations

This section will cover the limitations and threats to the validity of the study. The three main threats and limitations that were present are task selection, bias, and internal factors.

6.3.1 Task selection

As mentioned, the task that was used to conduct the experiment was *deceptive review detection*. This has previously been identified as a particularly hard task, especially when comparing it to others like spam detection, topic classification, and sentiment analysis. While using a hard task has its merits, there are some potential downsides. It can discourage them from carefully completing the tasks and increases the chances of having participants submit random guesses as an answer. Additionally, the debugging intervention might be less effective. The intervention is focused on increasing the understanding of the AI system using highlights as explanations. As tasks get more challenging, the patterns recognized by the machine can become more obscure, potentially resulting in highlights that are unclear or counterintuitive for participants. This is very harmful to the effectiveness of the debugging intervention.

6.3.2 Bias

As pointed out by Draws et al. [27], cognitive biases introduced by task design and workflow may have a negative impact on crowdsourcing experiments. With the help of the Cognitive Biases Checklist introduced [27], we analyzed potential bias in our study.

Self-interest bias

is possible because the crowd workers we recruited from the Prolific platform were motivated through monetary compensation. Thus, it would be challenging to keep participants engaged in the debugging intervention and highly motivated to learn from the weakness of the AI system. For their efforts, participants were compensated at a rate of £7.6, minimum wage. This compensation might not offer them enough incentive to commit fully to the tasks. Especially when working with a hard task, which might be mentally demanding, they might feel like they are not compensated well enough. If they do not fully commit to completing the tasks successfully, they might not experience the full extent of the intervention's benefits. To alleviate any participants with low effort results, we put attention checks to remove ineligible participants from our study. The observation of reduced reliance brought by bad first impressions also happens with *Anchoring Effect*. Meanwhile, the participants generally underestimate the AI performance and believe they can outperform the AI system, which also may fall into *Overconfidence or Optimism Bias*.

6.3.3 Internal factors

During the debugging intervention, the highlights presented to participants were generated by the AI system. These reviews with their accompanying explanations were handpicked from the training data, as to make sure they fulfilled the requirements posed by the experiment design. This selection was carried out by two of the researchers. While we consider this selection to be completed successfully, there is no guarantee they were exempt from any mistakes or misinterpretations of the researchers. In addition to that, the ground-truth values of the explanations have also been decided upon by the researchers. Not only introduces this possibility of bias but when explanations become sufficiently obscure, the researchers might also be affected by the phenomenon described in section 6.3.1. The extent to which the determination of the ground truth is correct is therefore dependent on the researchers' knowledge of the task and AI system. This, also, could harm the effectiveness of the debugging intervention.

Chapter 7

Conclusions and Future Work

To conclude this thesis, we will summarize the purpose and findings of this research. Finally, some directions for possible future work will be provided.

7.1 Summary

In this thesis, we present an empirical study to understand the impact of a debugging intervention on the estimation of AI performance and user reliance on an AI system. Through our experiment, we measured its effect on humans' estimation performance and (appropriate) reliance. Additionally, we explore the effects of internal ordering on these measures. While our experimental results do not provide support to our original hypotheses, we can not determine with certainty that a debugging intervention does not help facilitate appropriate reliance on AI systems. Our results suggest that we should be careful when presenting users with AI weakness, to avoid any anchoring effects that may result in under-reliance. Our exploratory study suggests that user reliance and estimation performance can be calibrated without directly shaping trust, and confidence is amplified or decreased when there is agreement or disagreement with the AI system, respectively.

7.2 Future work

As mentioned, there are some limitations to this research. Future work could focus on these and explore the effects of a debugging intervention when the tasks presented to participants are not as challenging, and there is a more clear notion of ground truth.

Another potential area that may be explored is how to mitigate potential bias of users. Our results showed that users tend to overestimate self-performance while underestimating AI performance. Mitigating these effects could result in a better judgment of AI trustworthiness, thereby improving appropriate reliance.

Finally, further research is required to explore how to cope with confidence dynamics that emerge when users are exposed to AI disagreement. When there is disagreement between participants' initial decision and the AI advice, a decrease in confidence is observed. Additionally, when compared to insisting on their own decision, participants have more

confidence when giving agency to AI advice. Such observations may be a dangerous signal for appropriate reliance and research is required to find ways of mitigating these effects.

Bibliography

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[2] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

[3] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275–285, 2020.

[4] Suresh Kumar Annappindi. System and method for predicting consumer credit risk using income risk based credit score, 2014. US Patent 8,799,150.

[5] Sharon Bailin. Critical thinking and science education. *Science & Education*, 11(4): 361–375, 2002.

[6] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. How can explainability methods be used to support bug identification in computer vision models? In *CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437, 2019.

[9] Richard A Berk and Justin Bleich. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Pub. Pol'y*, 12:513, 2013.

[10] Dianne C Berry and Donald E Broadbent. On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology Section A*, 36(2):209–231, 1984.

[11] Gilles Brassard, Claude Crépeau, and Jean-Marc Robert. All-or-nothing disclosure of secrets. In *Conference on the Theory and Application of Cryptographic Techniques*, pages 234–238. Springer, 1986.

[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[13] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019.

[14] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[15] Eric T Chancey, James P Bliss, Yusuke Yamani, and Holly AH Handley. Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human factors*, 59(3):333–345, 2017.

[16] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. In *AMIA annual symposium proceedings*, volume 2016, page 371. American Medical Informatics Association, 2016.

[17] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[18] Chun-Wei Chiang and Ming Yin. You'd better stop! understanding human reliance on machine learning models under covariate shift. In *13th ACM Web Science Conference 2021*, WebSci '21, page 120–129, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383301. doi: 10.1145/3447535.3462487. URL https://doi-org.tudelft.idm.oclc.org/10.1145/3447535.3462487.

[19] Chun-Wei Chiang and Ming Yin. Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In Giulio Jacucci, Samuel Kaski, Cristina Conati, Simone Stumpf, Tuukka Ruotsalo, and Krzysztof Gajos, editors, *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, pages 148–161. ACM, 2022.

[20] Chun-Wei Chiang and Ming Yin. Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In *27th International Conference on Intelligent User Interfaces*, pages 148–161, 2022.

[21] Ewart J De Visser, Samuel S Monfort, Kimberly Goodyear, Li Lu, Martin O'Hara, Mary R Lee, Raja Parasuraman, and Frank Krueger. A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. *Human factors*, 59(1):116–133, 2017.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[23] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[24] Murat Dikmen and Catherine Burns. The effects of domain knowledge on trust in explainable ai and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162:102792, 2022.

[25] Steven E Dilsizian and Eliot L Siegel. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports*, 16(1):1–8, 2014.

[26] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[27] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 48–59, 2021.

[28] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. Misuse and disuse of automated aids. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 43, pages 339–339. SAGE Publications Sage CA: Los Angeles, CA, 1999.

[29] Robert H Ennis. Critical thinking and subject specificity: Clarification and needed research. *Educational researcher*, 18(3):4–10, 1989.

[30] Noreen C Facione and Peter A Facione. *Critical thinking assessment in nursing education programs: An aggregate data analysis*. California Academic Press, 2000.

[31] Peter Facione. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (the delphi report). 1990.

[32] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

[33] Thomas Franke, Christiane Attig, and Daniel Wessel. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467, 2019.

[34] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1631–1640, 2015.

[35] Tim Van Gelder. Teaching critical thinking: Some lessons from cognitive science. *College teaching*, 53(1):41–48, 2005.

[36] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020.

[37] Filip Gralinski, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski, and Tomasz Górecki. Geval: Tool for debugging nlp datasets and models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, 2019.

[38] Ben Green and Yiling Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *arXiv preprint arXiv:2012.05370*, 2020.

[39] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[40] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

[41] Diane F Halpern. Assessing the effectiveness of critical thinking instruction. *The journal of general education*, 50(4):270–286, 2001.

[42] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988. doi: https://doi.org/10.1016/S0166-4115(08)62386-9. URL https://www.sciencedirect.com/science/article/pii/S0166411508623869.

[43] Andreas Henelius, Kai Puolamäki, and Antti Ukkonen. Interpreting classifiers through attribute interactions in datasets. *arXiv preprint arXiv:1707.07576*, 2017.

[44] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.

[45] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[46] Ayanna Howard, Cha Zhang, and Eric Horvitz. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 1–7. IEEE, 2017.

[47] Gajendra Jung Katuwal and Robert Chen. Machine learning model interpretability for precision medicine. *arXiv preprint arXiv:1610.09045*, 2016.

[48] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[49] Moritz Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*, pages 13–30. Springer, 2018.

[50] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[51] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 41–48. IEEE, 2010.

[52] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[53] Emily R Lai. Critical thinking: A literature review. *Pearson's Research Reports*, 6(1): 40–41, 2011.

[54] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38, 2019.

[55] Vivian Lai, Han Liu, and Chenhao Tan. "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM, 2020.

[56] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

[57] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[58] Piyawat Lertvittayakumjorn and Francesca Toni. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528, 2021.

[59] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[60] Mengyao Li, Brittany E Holthausen, Rachel E Stuck, and Bruce N Walker. No risk no trust: Investigating perceived risk in highly automated driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 177–185, 2019.

[61] Matthew Lipman. Critical thinking and the use of criteria. *Inquiry: Critical Thinking across the Disciplines*, 1(2):2–2, 1988.

[62] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[63] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[64] Zhuoran Lu and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 78:1–78:16. ACM, 2021.

[65] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[66] Dietrich Manzey, Juliane Reichenbach, and Linda Onnasch. Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6(1):57–87, 2012.

[67] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[68] John E McPeck. Critical thinking and subject specificity: A reply to ennis. *Educational researcher*, 19(4):10–12, 1990.

[69] Stephanie M Merritt. Affective processes in human–automation interactions. *Human Factors*, 53(4):356–370, 2011.

[70] Mahsan Nourani, Joanie T. King, and Eric D. Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. 2020.

[71] Institute of Electrical and Electronics Engineers. Ansi/ieee standard glossary of software engineering terminology. 1991.

[72] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.

[73] Myle Ott, Claire Cardie, and Jeffrey T Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 497–501, 2013.

[74] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[75] Devi Parikh and C Zitnick. Human-debugging of machines. *NIPS WCSSWC*, 2(7):3, 2011.

[76] Richard Paul and Gerald M Nosich. A model for the national assessment of higher order thinking. 1992.

[77] Alexander J Quinn and Benjamin B Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1403–1412, 2011.

[78] Amy Rechkemmer and Ming Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani, editors, *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 535:1–535:14. ACM, 2022.

[79] Byron Reeves and Clifford Nass. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10:236605, 1996.

[80] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[81] Julian Sanchez, Wendy A Rogers, Arthur D Fisk, and Ericka Rovira. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical issues in ergonomics science*, 15(2):134–160, 2014.

[82] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. Should i follow ai-based advice? measuring appropriate reliance in human-ai decision-making. In *ACM Conference on Human Factors in Computing Systems (CHI'22), Workshop on Trust and Reliance in AI-Human Teams (trAIt)*, 2022.

[83] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

[84] Andrew Selbst and Julia Powles. "meaningful information" and the right to explanation. In Sorelle A. Friedler and Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, page 48. PMLR, 2018.

[85] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[86] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310, 2018.

[87] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. Second chance for a first impression? trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*, pages 77–87, 2021.

[88] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[89] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

[90] Xinxi Wang and Ye Wang. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 627–636, 2014.

[91] Daniel T Willingham. Critical thinking: Why it is so hard to teach? *American federation of teachers summer 2007, p. 8-19*, 2007.

[92] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 189–201, 2020.

[93] Yi Yang, Wei Qian, and Hui Zou. Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. *Journal of Business & Economic Statistics*, 36(3):456–470, 2018.

[94] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[95] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[96] Roozbeh Yousefzadeh and Dianne P O'Leary. Debugging trained machine learning models using flip points. In *ICLR 2019 Debugging Machine Learning Models Workshop*, 2019.

[97] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

Appendix A

Appendix

A.1 Additional Experimental Details

Guidelines. Following Lai et al. [55], we provided the following guidelines in the user study:

- Deceptive reviews tend to focus on aspects that are external to the hotel being reviewed, (e.g. husband, business, vacation).
- Deceptive reviews tend to contain more emotional terms; positive deceptive reviews are generally more positive and negative deceptive reviews are more negative than genuine reviews.
- Genuine reviews tend to include more sensorial and concrete language, in particular, genuine reviews are more specific about spatial configurations, (e.g. small, bathroom, on, location).
- Deceptive reviews tend to contain more verbs, (e.g. eat, sleep, stay).
- Deceptive reviews tend to contain more superlatives, (e.g. cleanest, worst, best).
- Deceptive reviews tend to contain more pre-determiners, which are normally placed before an indefinite article + adjective + noun, (e.g. what a lovely day!).

Timer. Besides attention checks, we also added a timer to ensure each participant spent enough time on the questionnaires, task instructions, and decision tasks. A conservative estimate through trial runs reflected that participants would take at least 25 seconds to complete each decision task and 30 seconds to complete each debugging task. We reduced the time for the decision-making in the second stage to 15 seconds. Since attention check pages do not require deliberation, we reduced that time to 0 seconds, and participants were allowed to leave this question open. This, however, was tracked and when a participant left an attention check open or answered one incorrectly a second time, they were removed from the study.

Qualification Test. To ensure participants carefully read the task instruction and understand the task, we used two questions for the qualification test.

- In this study, the deceptive reviews written by? Option 1: An AI system, option 2: People without actual experience.
- Indicate whether the following statement is true or false: "Guidelines are provided for finding deceptive reviews". Option 1: True, option 2: False.

Variable Type	Variable Name	Value Type	Value Sale
Assessment (DV)	MAP	Continuous, Interval	[0, 10]
	MTP	Continuous, Interval	[0, 10]
	CCD	Continuous, Interval	[0, 10]
Reliance (DV)	Agreement Fraction	Continuous, Interval	[0.0, 1.0]
	Switch Fraction	Continuous, Interval	[0.0, 1.0]
	RAIR	Continuous, Interval	[0.0, 1.0]
	RSR	Continuous, Interval	[0.0, 1.0]
Performance (DV)	Accuracy	Continuous, Interval	[0.0, 1.0]
Trust (DV)	TiA-R/C	Likert	5-point, 1: poor, 5: very good
	TiA-U/P	Likert	5-point, 1: poor, 5: very good
	TiA-IoD	Likert	5-point, 1: poor, 5: very good
	TiA-Trust	Likert	5-point, 1: strong distrust, 5: strong trust
Covariates	ATI	Likert	6-point, 1: low, 6: high
	TiA-PtT	Likert	5-point, 1: tend to distrust, 5: tent to trust
	TiA-Familiarity	Likert	5-point, 1: not familiar, 5: very familiar

Table A.1: The different variables considered in our experimental study. "DV" refers to the dependent variable.

Variables. To have a more comprehensive view of variables used in our experimental analysis, we listed the main variables in table A.1. Notice that we did not add the confidence and dimensions from the NASA-TLX questionnaire into it.