

Knowing Better Than the AI: How the Dunning-Kruger Effect Shapes Reliance on Human-AI Decision Making

Lucie A. Kuiper

Computer Science

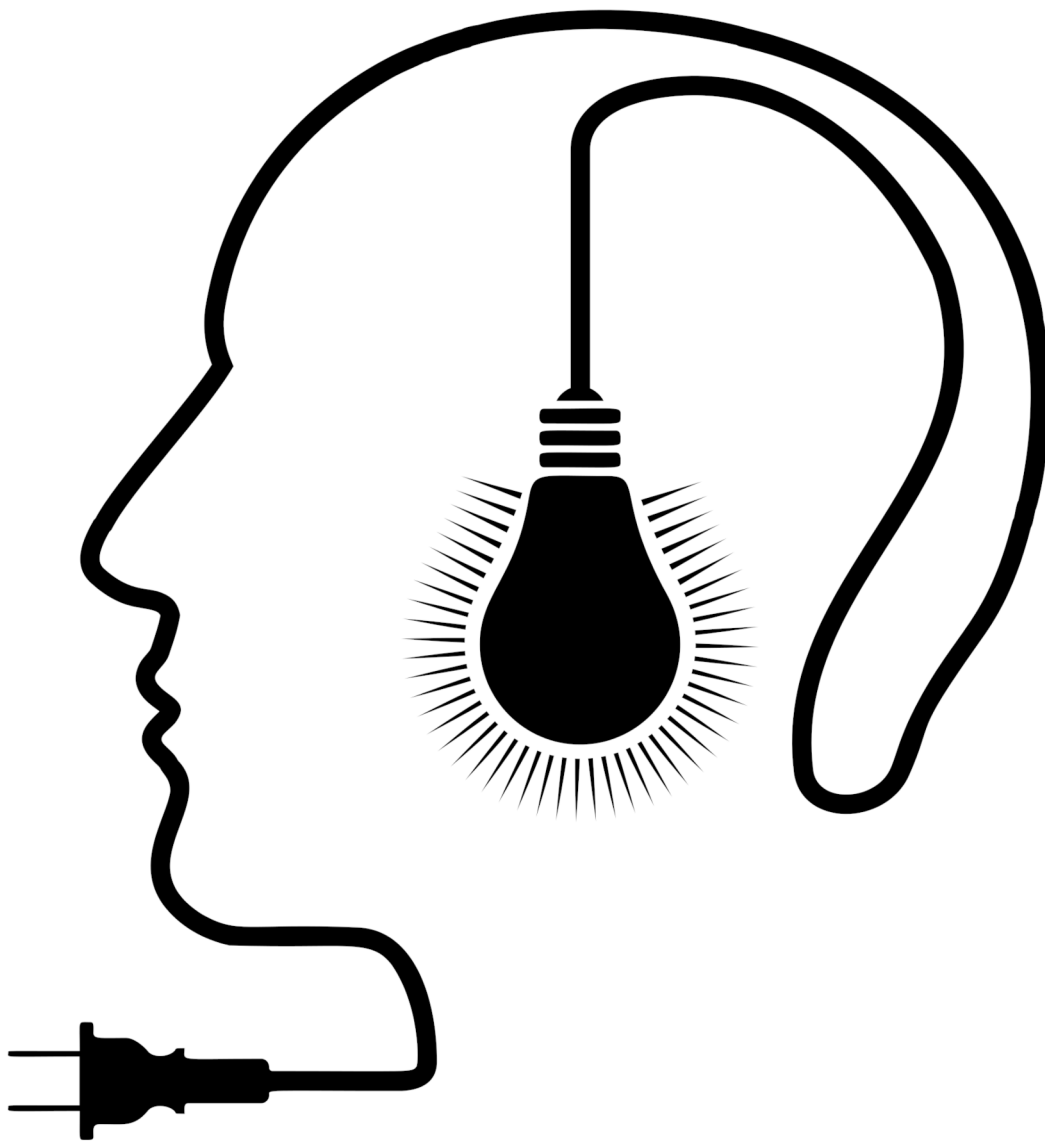


Image used for cover: image by Gordon Johnson, available from Pixabay¹

¹<https://pixabay.com/vectors/head-lightbulb-illumination-human-3893711/>

KNOWING BETTER THAN THE AI: HOW THE DUNNING-KRUGER EFFECT SHAPES RELIANCE ON HUMAN-AI DECISION MAKING

by

Lucie A. Kuiper

in partial fulfillment of the requirements for the degree of

Master of Science
in Computer Science

at the Delft University of Technology,
to be defended publicly on Wednesday November 30, 2022 at 17:15.

Student number: 4495403

Chair:	Prof. Dr. G.J.P.M. Houben	TU Delft
Supervisor:	Dr. U.K. Gadiraju	TU Delft
Thesis committee:	Dr. M.L. Tielman	TU Delft

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

ABSTRACT

Artificial Intelligence (AI) is increasingly helping people with all kinds of tasks, due to its promising capabilities. In some tasks, an AI system by itself will take over tasks, but in other tasks, an AI system making decisions on its own would be undesired due to ethical and legal reasons. In those cases, AI can still be of help by forming human-AI teams, in which humans get advice from the AI system helping them with making their final decisions. Human-AI teams are for instance used in the medical and legal fields. One problem arises, in which instances should one trust an AI system and in which not? Trusting the AI system when it is correct and trusting yourself when you are correct, results in a high appropriate reliance. If users appropriately rely on AI systems, it is possible to achieve complementary team performance, which is better than any single teammate. However, as known from previous literature, people struggle with assessing their performance and knowing how well they perform compared to peers. When one overestimates their performance this can be because of a dual burden, due to the lack of skill they also lack the skill to accurately estimate their performance. This phenomenon is called the Dunning-Kruger Effect (DKE). This raises questions about whether the inability to estimate their own capabilities would also reflect on their assessment of the AI system its performance.

In this thesis we look at how the DKE affects (appropriate) reliance on AI systems and if so, how such effects due to the DKE can be mitigated. The effects of the DKE and possible mitigation are being tested through an empirical study (N = 249). The attempt at mitigation is done by including a tutorial intervention, which has been proved in previous research to be useful in decreasing the DKE. The tutorial intervention is aimed at revealing the weaknesses of the participant and making them aware of their miscalibrated self-estimation. Furthermore, in this thesis, the effects of revealing how the AI system makes its decisions through explainable AI (XAI) are explored. The XAI consisted of highlights from logic unit-based explanations, it should allow participants to gain more understanding of the AI advice. This thesis shows how this will affect user self-assessment and reliance on the AI system.

We found that participants who overestimate themselves tend to rely less on the AI system, compared to participants that had an accurate or underestimation of their performance. After the tutorial participants have a better calibration of their self-assessment. While the designed tutorial intervention can help participants calibrate their self-assessment, it fails to promote (appropriate) reliance. Furthermore, the logic units-based explanations did not improve accurate self-assessing or increase user (appropriate) reliance on AI systems. This thesis shows the importance of considering cognitive biases when dealing with human-AI teams and invites more research on how to handle and mitigate the DKE in human-AI decision making.

PREFACE

This thesis project is the end of my time as a master's student in computer science at the Delft University of Technology. During this master, I have learned a tremendous amount, especially during the thesis. I would like to thank everyone who made that possible for me.

Completing this thesis would not have been possible without the help from others. First of all my sincere thanks to my supervisor Ujwal Gadiraju, for making this thesis possible. Thank you for proposing this thesis idea and guiding me through the process of the thesis. Also, I would like to thank my daily supervisor Gaole He, for the weekly meetings and for providing me with guidance, recommendations and feedback.

I would like to thank Geert-Jan Houben for being my thesis advisor and the chair of my thesis committee. Furthermore, I want to thank Myrthe Tielman, for taking the time to read and discuss my thesis as the final committee member.

Lastly, I want to thank my family, friends and boyfriend who always let me talk about my thesis and the problems I encountered along the way. And who always provided me with support and encouragement.

Lucie Kuiper
Delft, November 2022

CONTENTS

Acknowledgements	iii
List of Figures	3
List of Tables	5
1 Introduction	1
1.1 Research questions	2
1.2 Contributions	3
1.3 Outline	3
2 Background	5
2.1 Human-AI Decision making	5
2.2 Appropriate Reliance	5
2.3 Self-Assessment.	6
2.4 The Dunning-Kruger Effect (DKE)	6
2.5 Explainable AI (XAI)	7
3 Methods	9
3.1 Hypotheses	9
3.2 Tasks	10
3.2.1 Logical Reasoning Task.	10
3.2.2 Pilot Study	11
3.2.3 Task Selection	11
3.2.4 Two Stage Decision-Making	12
3.3 Quality Control	12
3.4 Tutorial	13
3.5 Logic Unit Based Explanations	14
4 Experimental Set-Up	17
4.1 Conditions	17
4.2 Measures and Variables	17
4.3 Participants	20
4.4 Procedure	20
5 Results	23
5.1 Cleaning Data	23
5.2 Descriptive Results	24
5.3 Type of Tests	24
5.4 Hypotheses	25
5.4.1 Hypothesis One	25
5.4.2 Hypothesis Two	26
5.4.3 Hypothesis Three	27
5.4.4 Hypothesis Four	29
5.4.5 Summary	30
5.5 Exploratory Results	31
5.5.1 XAI.	31
5.5.2 The DKE	31
5.5.3 Analysis of Trust	32

6 Discussion	35
6.1 Key Findings	35
6.2 Further Findings	36
6.3 Implications	36
6.4 Limitations	37
6.5 Future work.	37
7 Conclusions	39
A Interface	41
B Results	51
Bibliography	53

LIST OF FIGURES

1.1	Illustration of the process of human-AI decision making.	2
3.1	Question without AI advice (initial) Examples of questions used in experiment, both stages (with and without AI)	11
3.2	Second stage task with AI advice	12
3.3	Example of a tutorial page	14
3.4	Example of an AI advice page with XAI	15
4.1	Illustration of procedure followed in our study for a ✓ Tutorial, ✓ XAI condition participant. Blue boxes represent questionnaires, orange boxes represent parts involving tasks	21
5.1	Graphs on descriptive results shows the distribution of (a) participants over the nationalities, (b) the perceived helpfulness of logic units-based explanations, (c) overestimated self-estimation, accurate self-estimation and underestimated self-estimation across all experimental conditions over the first batch of tasks.	24
5.2	Distribution of agreement and switch fraction of participants	25
5.3	Estimation per group divided on performance. (a) Shows the participants divided on low, high and all (which includes low and high) on how they rated themselves on a scale of 0% to 100%, next to their actual average performance compared to others. (b) Shows the amount of participants that overestimated themselves per group, divided in low performers, high performers and average performers (the once that are not in either low or high are put in average). The overestimation is measured by comparing the amount of correct answers they estimated to have correct to the amount they actually had answered correctly.	32
A.1	Introduction page	42
A.2	ATI first out of three	42
A.3	ATI second out of three, includes attention check question	43
A.4	ATI third out of three	43
A.5	TiA-Propensity to trust questionnaire	44
A.6	Instructions	44
A.7	Task	45
A.8	Task with XAI advice	45
A.9	First attention check out of three	46
A.10	Second attention check out of three	46
A.11	Third attention check out of three	47
A.12	Survey question on self-assessment	47
A.13	Survey question on assessment of peers	47
A.14	Survey question on how participant would place themselves amongst peers based on performance	48
A.15	Tutorial instructions	48
A.16	Tutorial	48
A.17	Entering last batch of questions	49
A.18	Question on helpfulness	49
A.19	Disabled continue button	49
B.1	Accuracy distribution of participants	51
B.2	RAIR and RSR distribution of participants	52
B.3	Miscalibration distribution of participants	52

LIST OF TABLES

4.1	Appropriate reliance patterns considered by Schemmer <i>et al.</i>	19
4.2	Variables considered in our experimental study. “DV” refers to the dependent variable.	20
5.1	Wilcoxon-Mann Whitney test results for hypothesis 1 . Comparison between participants that did overestimate themselves and participants that did not. Significant results are shown in bold.	26
5.2	Wilcoxon-Mann Whitney test results for hypothesis 1a comparison between participants that did receive XAI and participants that did not. Compared to both participants that overestimated themselves and participants that did not. Significant results are shown in bold.	26
5.3	Wilcoxon-Mann Whitney test results for hypothesis 2 . Comparison between participants that had an accurate estimate of themselves and participants that did not. Significant results are shown in bold.	27
5.4	Wilcoxon-Mann Whitney test results for hypothesis 2a . Comparison between participants that did receive XAI and participants that did not. Compared to both participants that had an accurate estimation themselves and participants that did not. Significant results are shown in bold.	27
5.5	Wilcoxon signed ranks test results for hypothesis 3 . Comparison between the first and last batch of tasks the participant performed. Significant results are shown in bold.	28
5.6	Wilcoxon-Mann Whitney test results for hypothesis 3a . Comparison of miscalibration between participants that did receive XAI and participants that did not. Significant results are shown in bold	28
5.7	Wilcoxon-Mann Whitney test results to further explore hypothesis 3 . Comparison of miscalibration between participants that did receive a tutorial intervention and participants that did not. Significant results are shown in bold.	29
5.8	Wilcoxon signed ranks test results for hypothesis 4 . Comparison between the first and last batch of tasks the participant performed. Significant results are shown in bold.	29
5.9	Wilcoxon-Mann Whitney test results for hypothesis 4a . Comparison between participants that did receive XAI and participants that did not. Significant results are shown in bold.	30
5.10	Wilcoxon-Mann Whitney test results to further explore hypothesis 4 . Comparison between participants that did receive a tutorial intervention and participants that did not. Significant results are shown in bold.	30
5.11	Results of all hypotheses	30
5.12	Wilcoxon-Mann Whitney test. Comparison of participants that received XAI vs participants that did not receive XAI. Significant results are in bold.	31
5.13	Wilcoxon-Mann Whitney test. Comparison of participants that had a high accuracy versus participants that had a low accuracy. Significant results are in bold.	32
5.14	Wilcoxon signed ranks test results. Comparison between the TiA-trust after the first and last batch of tasks. The Likert scale is from 1-5. Significant results are shown in bold.	33
5.15	Wilcoxon-Mann Whitney test results. Comparison between participants that did receive a tutorial intervention and participants that did not. Comparison on trust on a Likert scale from 1-5. Significant results are shown in bold.	33

1

INTRODUCTION

In this time and age, when trying to navigate to a new destination, it is becoming second nature to consult artificial intelligence (AI) to find directions for the quickest route. As the performance of AI is increasing further and further, interaction with AI systems of all kinds is becoming omnipresent in a wide variety of functions. Often these AI solutions can be seen as advice given by the AI system. As a user, you do not need to oblige with this advice and can for instance still take another route, or choose among a set of routes. While making the sub-optimal choice for a route will only result in a longer journey, in other fields of work making a mistake of any kind can have bigger consequences. For instance, humans are also assisted by AI in the medical domain [1], in which a wrong call can have severe consequences. In these high-stakes domains, it is often not preferred to use fully automated AI systems largely due to ethical and legal reasons, therefore, human-AI teams can be used instead. In those domains, the AI system can help and advise the user, but the final decision is made by the human user. In the current situation, most of the time human-AI teams do outperform humans alone, but fail to outperform AI alone [2]. A big aim in current-day research is to find a complementary performance[2–5], in which a human is assisted by an AI system to come to a performance that could not be achieved by either of those on their own. As these solutions are often used in high-stakes domains with the risk of tremendous consequences, further studies on improving human-AI collaboration in these domains are crucial. An illustration of how human-AI decision making can take place is given in Figure 1.1.

To improve this human-AI decision making it is necessary to know what the aim should be. Often a performance improvement can be made by making humans align more with the AI system. This is due to the fact that AI systems often have a higher average performance. However, when aiming for this there would be no added benefit of keeping humans in the loop, as they would start copying the AI for optimal performance. Furthermore, by doing this it will not be possible to obtain a complementary performance when the AI system gets blindly relied upon. This results in a need to measure how appropriate human reliance on the AI system is, relying on it when it is correct and relying on themselves when the AI system is wrong. A method to measure appropriate reliance on AI systems is described by Schemmer *et al.* [6]. This measure shows when someone has a relative positive AI reliance (RAIR) and a relative positive self-reliance (RSR), which should both be considered for appropriate reliance. Appropriate reliance allows us to look at the performance of human-AI decision making in a different way, in which the focus is put on the trust dynamic between the human and AI. What factors can contribute to this idea of appropriate reliance? One thing that is being looked at is increasing the understanding of the users about the AI system, which can potentially increase appropriate trust [4].

One fruitful way to increase user understanding of the AI system is by supplying the user with explanations of the AI, known as explainable AI (XAI). One instance of XAI is by giving real-time support with inline highlights [7]. XAI can help users understand AI and therefore make better decisions which can then result in more appropriate reliance. However, it is important to take the human mental model into consideration while designing such XAI[4, 8], as explaining just how AI came to its decisions might not be the most helpful to users.

While performing tasks some people think they perform a lot better than they actually do, and overestimate their performance. It seems especially hard to calibrate the estimation of performance to match actual performance when underperforming, due to a dual burden in which lack of skill on a task also creates the

lack of meta-cognition to recognize this lack of skill. This effect in which the unskilled are also unaware is called the Dunning-Kruger Effect (DKE) [9]. Previous research has shown that crowd workers can suffer from the DKE [10]. This makes us wonder what implications the DKE has on human-AI decision making.

With human-AI decision making having an increasing effect on our lives, it is important to have a good understanding of this relationship between humans and AI systems. People who do not perform well generally overestimate their performance, this makes it interesting to take a look at how these people interact with AI systems. Ones that overestimate themselves have problems gaining a correct mental model of their own performance, therefore we are interested to see how they would place themselves compared to AI systems. Less awareness of own performance versus that of the AI system, could then result in less reliance on the AI and inflated unjustified self-reliance, which also could harm overall performance. In previous research by Schaffer *et al.* [11] the subject of the DKE in human-AI decision making has already been touched upon. Schaffer *et al.* showed that in a prisoner's dilemma-like task the self-assessment of a user affects the reliance on the AI system. This makes it interesting to take a look at how the effects of the DKE take further shape in human-AI reliance. This thesis focuses on how the (appropriate) reliance on AI in human-AI decision making is affected by the DKE. A subject which to the best of our knowledge has not been addressed in prior studies.

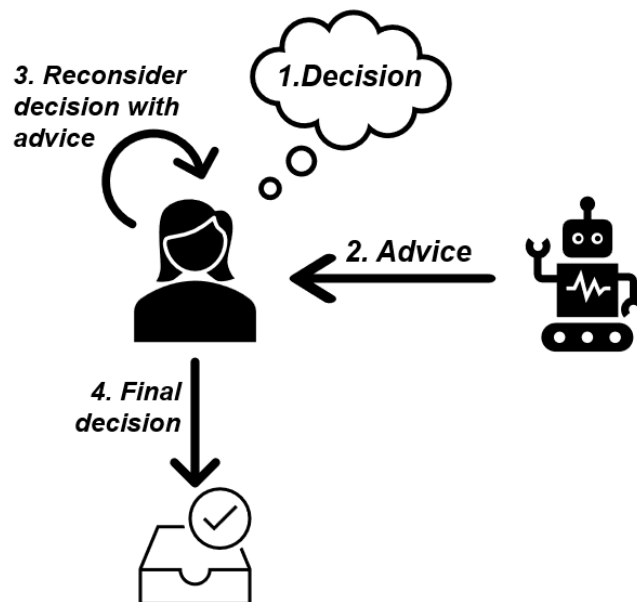


Figure 1.1: Illustration of the process of human-AI decision making.

1.1. RESEARCH QUESTIONS

We have established that the thesis should entail the effects of the DKE on reliance in human-AI decision making, more specifically how the users in human-AI decision making shape their reliance on AI systems. We expect that due to the DKE the reliance of these users might differ from their counterparts. Furthermore, it also raises the question, that if the DKE affects the (appropriate) reliance, and how this then can be improved. How can the effects of the DKE be minimized and mitigated? Therefore the two following research questions will be the focus of this thesis.

- **Research question 1: How does the Dunning-Kruger Effect shape reliance on AI systems?**
- **Research question 2: How can the Dunning-Kruger Effect be mitigated in human-AI decision making tasks?**

To find answers to these research questions we proposed four hypotheses on the effect of overestimation on reliance, the effect of accurate self-estimation on reliance, the effect of a tutorial intervention on self-assessment and the effect of a tutorial intervention on (appropriate) reliance. We expect to find under-reliance on AI systems from people that overestimate their performance, and thus are subject to the DKE, we

also want to create a way to mitigate this effect. This brings us to the problem of how to mitigate the DKE. In previous research, it has been found that a way to reduce this effect is by providing a way to increase knowledge as a way to also increase the meta-cognitive skills on self-awareness [9]. Furthermore, the paper by Lai *et al.* [7] shows that a tutorial intervention can be used to increase human-AI team performance. Therefore to find an answer to research question 2, we chose to look at what effect a tutorial could have on users that are subject to the DKE, specifically the effect on their (appropriate) reliance. Will tutorials make the users understand better how the AI is working and how their performance scales to the performance of AI? This inspired us to include a tutorial intervention in the research.

The experiment in this study was done by crowd-sourcing participants from Prolific¹ in which 249 participants participated. These participants were asked to go through our experiment, in which they were asked to answer multiple-choice questions with and without explanations when receiving AI advice. The implementation and additional material of the experiment can be found on the GitHub page². Note that the data and code will only be made public after our paper, which is under review at the moment of writing, is accepted.

1.2. CONTRIBUTIONS

This thesis aims to discover more about how the DKE shapes reliance on AI systems and how this can be mitigated. This gives us several challenges and while executing this research we have the following contributions:

- In this thesis a research gap is formulated in human-AI decision making. The effects of the DKE on (appropriate) reliance and ways to mitigate the DKE within the human-AI decision making context have not been explored enough.
- An experimental design including a tutorial consisting of manually created contrastive explanations is created for this thesis.
- We carry out a quantitative study with 249 participants. Some parts of the experiment are between-subject and others are within-subject. The experiment is about the effects of overestimating and having an accurate self-assessment on the (appropriate) reliance on AI systems in human-AI decision making. Furthermore, the experiment is about the effects of a tutorial intervention on the calibration of the self-assessment and the effects of a tutorial intervention on (appropriate) reliance on AI systems in human-AI decision making. Reliance is measured with two variables and appropriate reliance is measured with an additional two variables.

1.3. OUTLINE

The rest of the thesis is structured as follows. First in chapter 2 we give some background on human-AI decision making. We provide the necessary preliminary and background information for this work: self-assessment, the DKE, XAI and appropriate reliance. Next, we describe the methodology in chapter 3, it discusses the hypotheses and how those were found, logical reasoning tasks, task selection, two-stage decision-making, the pilot study, quality control, the tutorial, logic units-based explanations and some design choices. Following that we give the experimental setup in chapter 4, this chapter contains the conditions, measures, variables, participants and procedure. In chapter 5 we present the results of all hypotheses and some exploratory results. Next, in chapter 6 we discuss the findings, implications, and limitations and point out promising future directions. Finally, in chapter 7 we give a summary of the whole thesis and outline the most important findings.

¹prolific.co

²<https://github.com/LucieKuiper/Thesis>

2

BACKGROUND

This thesis is centred around human-AI decision making. More specifically, this thesis focuses on the effects of the Dunning-Kruger effect (DKE) on the (appropriate) reliance of humans on AI systems in human-AI teams. To gain a better understanding of current research a few areas need to be covered. First of all, it is important to discover more about the workings of Human-AI decision making. Additional information is given about the self-assessment of users in systems, specifically in human-AI interaction. The DKE will be discussed in more detail to build further on the self-assessment. As a means to mitigate the effect explainable AI (XAI) is being used in this thesis. This chapter describes how XAI has been used in previous studies. Finally, appropriate reliance and previous studies on appropriate reliance are presented.

2.1. HUMAN-AI DECISION MAKING

With the ever-growing influence of AI in our daily life, it is important to give our interaction with AI the right direction, as for instance discussed by Lee *et al.* [12]. Human-AI decision making is a form of human-AI interaction, in which humans and AI collaborate on certain tasks. Recent work features a lot of research on how users shape trust in AI systems and how the users' behaviour is affected by AI systems. Notable topics include risk perception [13, 14], performance feedback [4, 15] and confidence [16, 17] of machine learning models. Currently, human-AI teams often work on high-stakes tasks which involve a big risk when done wrong. For instance, in the legal, financial and medical field [18–20], in which mistakes can have big consequences. This comes as a logical consequence of the fact that complete automation in these areas is unwanted due to ethical and legal concerns. However, fully manual work in these areas can also be tedious and less effective than human-AI teams, which gives AI a place as an advisor. This gives the incentive to optimise these collaborations, as the concerns regarding legality and ethics will not change in the near future. Just AI often outperforms human-AI teams, however, it is still key to optimise for these teams. The desire for human-AI teams is to outperform both humans and AI systems. Currently, human-AI team performance often lies in between the solitary performance of humans and AI [2]. Therefore a lot of work should still be done in order to improve human-AI teams to reach the potential of complementary performance. In most of these human-AI teams, the last decision is made by the human that receives an estimation or advice from the AI system. Therefore it is important that the human interacting with the AI system knows how to interact best, but the AI should also supply the information in such a way that the person interacting with it can optimize their teamwork. This is backed by Bansal *et al.* [4], who state the importance of the mental model that humans create of the AI system, and its role in improving overall performance.

2.2. APPROPRIATE RELIANCE

For human-AI decision making to be successful people must trust the AI system when it is right, but trust themselves when they are right. The desire for human-AI teams is to have a complementary performance [3]. This can not be achieved when people always trust the AI system, as that would just result in a performance that is in line with the AI system. Furthermore, this would be an automation bias. In recent research, human-AI team performance is often better than human, but worse than AI on its own [2, 3, 21]. This is due to two causes. First of all, humans do not know when to trust AI and do not take advantage of the powers of

AI, which can be described as under-reliance. Secondly, humans trust AI systems too much when it is incorrect, instead of relying on themselves when they outperform the AI system, which can be described as over-reliance. Both can be a form of automation bias in which people do not really consider the answers anymore and just generally decide whether or not to follow the AI advice. While the increase of reliance on AI systems could increase performance as AI systems often still outperform human-AI teams, it will not be possible to achieve complementary performance in this way. Furthermore, it would just result in further automation bias and over-reliance which could defeat the whole purpose of human-AI teams. If someone just copies the AI advice, then there are no benefits of someone going over all tasks manually. The research by Lee and See [22] talks about the importance of appropriate trust for human-computer interaction. Increased appropriate reliance would help achieve complementary performance, as higher appropriate reliance should indicate someone relies on the AI when appropriate, but also relies on themselves when appropriate. It is important to use universal measures, so research on human-AI decision making can be compared to one another. However, often still different measures are used which complicates comparison [21]. The measures in this thesis used are described by Schemmer *et al.* [6] who presented a useful metric to measure appropriate reliance. In existing research, the focus is mainly on mitigating under-reliance and over-reliance. Several methods have been successful in mitigating the under-reliance and over-reliance, like for instance user tutorials [23, 24] and explanations [25] have been proven to be successful. Chiang and Yin [23] claim that tutorials, when used appropriately, can help participants rely on machine learning models more appropriately. A previous study Chiang and Yin [24], found that a short education session can be effective in reducing over-reliance on out-of-distribution data. Furthermore, such systems can educate people to further improve their capabilities. Inspired by this we explore whether the DKE in human-AI decision making can be mitigated by the use of a tutorial intervention.

2.3. SELF-ASSESSMENT

The evaluation of one their own performance is more commonly called "self-assessment". The level of meta-cognition one has determines how well one can assess themselves. It is a fundamental skill, as a better insight into how well you are performing can help improve your capabilities [26]. The use of self-assessment has been widely used in educational settings to improve student performance [27, 28]. While the benefits of self-assessment are clear as a way to self-improve, people do tend to struggle to get their self-assessment calibrated [29]. The true extent of usefulness of self-assessment has also been up to question in recent research [30]. Generally, people tend to overestimate their abilities, this overestimation does decrease by gaining knowledge [31]. The tendency of overestimation is partially described by the overestimation bias [32]. Often, even the estimation of peers can be better than the self-estimation [33]. This overestimation comes from several reasons, people tend to think they are above average and have an unrealistic optimistic view of themselves. In the field of Human-Computer Interaction (HCI) research has been done on several purposes of using self-assessment. For instance, in the research done by Gadiraju *et al.* [10], self-assessment is used to do pre-selection in crowdsourcing, based on competence. In this thesis, self-assessment will be measured as a way to indicate the DKE.

2.4. THE DUNNING-KRUGER EFFECT (DKE)

When someone has an inflated self-assessment that does not align with their actual performance, we can speak of the Dunning-Kruger effect. The DKE is first described and named by Kruger and Dunning [9]. The paper describes how people have a favourable outlook on their own performance. This could be due to the so-called "dual burden" which states that individuals who are unable to perform well in a task, also lack the meta-cognition to correctly assess their performance. In the same paper, Dunning and Kruger also suggested and tested a way to mitigate the found effect. By increasing the skill of the participants they successfully increased the ability of participants to assess their performance. Increasing the skill is done with a training. Besides increased skill, the training also reduces their estimation of their performance, resulting in a significantly reduced gap between actual and estimated performance. This suggests that training is a way to reduce DKE. While the paper does mention that some other effects take place at the same time, like the above-average effect, the emphasis is on the inability to recognize their own mistakes, not on being overconfident. In the paper by Kruger and Dunning [9] DKE was found in logical reasoning, humour and grammar. Future studies showed that DKE is also prevalent in HCI [10, 34], and more specifically in human-AI decision making [11]. The research done by Schaffer *et al.* [11] was a user study based on an adaption of the prisoner's dilemma-like game. they found that familiarity with the task domain correlated with trust in the AI advice, but reliance is

lower. With this study, they proved not only that DKE is present in human-AI decision making, but also that it has an impact on reliance. This gives the inspiration to take a further deep dive into the DKE in human-AI decision making by testing on another type of task to check for transferability. Furthermore, we perform more extensive tests on the (appropriate) reliance on the AI system, to build on the previous research and gain a better understanding of how the DKE affects the reliance on AI systems.

2.5. EXPLAINABLE AI (XAI)

Explainable AI (XAI) gives a user of AI more insight into the AI by showing how the AI system has come to its conclusions. When people are interacting with AI systems they generally just receive advice and need to decide whether they want to follow that advice or not. Without being given any other context this can be troublesome, as the trust in the AI must just be built from the advice. Furthermore, it is in line with the GDPR (in effect in the European Union) [35], which states that AI and ML systems should give meaningful information about the system. This can be done with XAI and creates an incentive to test the workings of these explanations in research. The current use of XAI often puts its usefulness to enhanced performance into question. In recent research, Jacovi *et al.* [36], questions were put around the function and comprehension of humans of XAI. These problems call for a more human-centred approach. Buçinca *et al.* [37] stretches the importance of good explanations as misleading explanations could lead to a reduction of human-AI team performance, which again shows the importance of designing XAI with the human in mind. Furthermore, other research found that in human-AI decision making the use of XAI is not as effective as expected and could lead to automation bias [11]. This automation bias in AI has also been found in other research, specifically when strengths were observed early on [8]. In another paper, by Wang and Yin [25], the conclusion was made that the helpfulness of XAI is determined by the expertise of the person in that domain. More expertise meant the XAI could provide more help, as the people might not be able to interpret the XAI without domain knowledge in the field the decision making is about. As XAI is more helpful in domains people have some expertise in, it is best to find a task that is not too foreign to laymen. Therefore, we chose logical reasoning tasks in this thesis. The paper by Wang and Yin [25] also mentioned three desired abilities of XAI 1) making users understand the AI model, 2) making people see the uncertainty of an AI prediction, and 3) calibrating the trust in the model. This thesis will most specifically focus on the third point mentioned, as trust in AI is measured with and without XAI.

3

METHODS

This section describes the hypotheses that are derived from the research questions. Then it builds further on that, by describing the elements that are needed to answer these hypotheses, such as which tasks were used, why they were used and how these tasks were selected. Furthermore, a pilot study was done to gain more information prior to the experiments. Moreover, the tutorial intervention and explanation of AI used in the explanations are explained.

3.1. HYPOTHESES

In the introduction, the following two research questions were formulated, to study the research gap found. **Research question 1: How does the Dunning-Kruger effect shape reliance on AI systems?** and **Research question 2: How can the Dunning-Kruger effect when working with AI systems be mitigated?**. To answer these research questions the experiment has been designed around the Dunning-Kruger effect (DKE) and AI reliance. The DKE states, that people that are less competent find more difficulty in estimating their performance, due to the dual burden of incompetence[9]. As a consequence of the DKE, a participant could wrongly overestimate their own capabilities compared to AI, making them less likely to rely on AI advice. As explainable AI(XAI) gives more insight into the AI system, people who still overestimate themselves might even have a lower reliance on AI.

Therefore the first hypothesis is formulated as follows.

- **Hypothesis 1: Users overestimating their own performance will demonstrate relatively less reliance on AI systems.**
- *Hypothesis 1a: giving explanations during the AI-advice stage can reduce this effect.*

Appropriate reliance on AI is key to increased human-AI decision making performance. Therefore, appropriate reliance on AI is also to be studied in addition to the general reliance on AI. Besides overestimation of themselves, users can also underestimate themselves, additionally, some users might show an accurate estimation of their performance. We expect, that an accurate self-estimation will benefit one in knowing when to rely on AI. In this case, once more the influence of supplying participants with explanations is expected to amplify this effect, as it could increase understanding of the AI. Which might benefit participants that estimate their own performance right more. Thus the second hypothesis is formulated as follows.

- **Hypothesis 2: Users demonstrating accurate self-assessments will exhibit relatively appropriate reliance on AI systems.**
- *Hypothesis 2a: giving explanations during the AI-advice stage can amplify this effect.*

In the second research question, the desire is to mitigate the effect of the DKE. It has previously been found that feedback can help improve the perception of the complexity of questions and therefore decrease the DKE [38]. Feedback can be used to make users aware of their mistakes and help them recalibrate their self-assessment. With the right way of supplying feedback, we claim that it should be possible to mitigate the DKE. When users are supplied with explanations it is expected that the effect is increased as they are equipped with more information on the AI and its inner workings. This leads us to the third hypothesis, formulated as follows:

- **Hypothesis 3: Making users aware of their miscalibrated self-assessment, will help them improve their self-assessment.**
- *Hypothesis 3a: giving explanations during the AI-advice stage can amplify this effect.*

By obtaining a better understanding of themselves users' appropriate reliance on AI could also be improved, besides obtaining better self-assessment. Better awareness of their own capabilities due to being supplied with feedback can result in a better understanding of when they are not capable and should rely on AI instead of on themselves, which can lead to more appropriate reliance on AI. In the case where participants get explanations on the AI advice it is anticipated, that due to increased access to information on the AI, the effect will be amplified. Therefore, the following hypothesis was formulated.

- **Hypothesis 4: Making users aware of their miscalibrated self-assessment will result in relatively more appropriate reliance on AI systems.**
- *Hypothesis 4a: giving explanations during the AI-advice stage can amplify this effect.*

3.2. TASKS

To verify the DKE in an empirical study, we need a task, which complies with a few criteria. First of all, the tasks to be performed must not be too easy as the participants should make some mistakes to make them have to rely on AI and make them prone to the DKE. Furthermore, when participants already have a high accuracy giving them training might have a lesser effect. Something else to take into consideration is using a layman-friendly task, as we will not be able to gather enough experts in a specific field. While being layman-friendly, the task should also have some risk in putting trust into the AI, as we can not put the participants into situations with real urgency for a correct answer. To create some risk in giving an incorrect answer, we supplied them with monetary bonuses for correct answers, which should motivate participants to try their best. Another important criterion is that there has to be a ground truth answer for the tasks, as we must be able to pinpoint correct and incorrect answers to gather results. Furthermore, it is desirable for the task to be a realistic task to be performed as a human-AI decision making task, tasks that are arbitrary for AI are not preferable, as participants might notice and copy the answers, an example of an arbitrary task for AI would be calculating mathematical questions.

3.2.1. LOGICAL REASONING TASK

To satisfy most of the criteria mentioned previously the ReClor dataset has been chosen [39]. ReClor is a reading comprehension dataset that requires logical reasoning to answer the questions. The dataset is suitable for this study, as the human performance is not too high (which is tested in the pilot study (section 3.2.2)), and the test would not be arbitrary for AI, the current highest scores obtained by AI on the ReClor dataset can be found on their leaderbord¹. Furthermore, the ReClor dataset holds multiple-choice tasks, with a ground truth answer, which allows for easy checking of the answers. Logical reasoning tasks have been used in human-AI team research previously [2], other research suggest that on similar logical reasoning tasks the ceiling of human performance could outperform AI-based systems [40], mainly due to the complexity of reading comprehension to AI systems. It tests the language processing and logical reasoning skills of the participants, which is a layman-friendly task as all kinds of topics come across and it is not focused on a specific domain. Another reason to go with a logical reasoning task is that the DKE has been previously found in such tasks[9, 10], which makes it interesting to check if adding the advice from AI changes anything.

An example of one of the tasks can be seen in Figure 3.1, in this figure the three elements of the tasks are shown. (1) The context, and necessary information to be able to answer the question. (2) The question (called "task" in the figure), is a question asked about the context that needs to be correctly answered. (3) Answers, four different potential answers are displayed in multiple-choice style, one of these is the most correct in answering the question, the ground truth. Participants are asked to answer the question based on the context. The idea is that AI can extract information from the context, question and four possible answers to perform these same tasks. Currently, AI systems are not capable of performing these tasks without making mistakes.

¹<https://eval.ai/web/challenges/challenge-page/503/leaderboard/1347/Test>

Tasks

Context

Physician: In comparing our country with two other countries of roughly the same population size, I found that even though we face the same dietary, bacterial, and stress-related causes of ulcers as they do, prescriptions for ulcer medicines in all socioeconomic strata are much rarer here than in those two countries. It's clear that we suffer significantly fewer ulcers, per capita, than they do.

Task 1/16

Which one of the following, if true, most strengthens the physician's argument?

A The two countries that were compared with the physician's country had approximately the same ulcer rates as each other.

B The physician's country has a much better system for reporting the number of prescriptions of a given type that are obtained each year than is present in either of the other two countries.

C A person in the physician's country who is suffering from ulcers is just as likely to obtain a prescription for the ailment as is a person suffering from ulcers in one of the other two countries.

D Several other countries not covered in the physician's comparisons have more prescriptions for ulcer medication than does the physician's country.

Confirm and continue

Figure 3.1: Question without AI advice (initial) Examples of questions used in experiment, both stages (with and without AI)

3.2.2. PILOT STUDY

To gain more insight into the difficulty of specific questions of the Reclor data set, a pilot study was set up. Furthermore, the pilot helped gain information on the average amount of time spent by Prolific² crowd workers on these tasks. A supplementary benefit was that the pilot study allowed us to have a test run with participants to get rid of any last bugs and problems in the system. The information on task difficulty was needed to create two batches of questions of approximately the same difficulty. These batches of similar difficulty are needed to make fair comparisons on the tasks done before the intervention and after. For the intervention, which will be a tutorial, it is also needed to have questions with adequate difficulty, as the tutorial will have more effect if participants need to receive feedback, due to wrong answers. Therefore it is desirable that the questions are not too easy. Task difficulty was also necessary to confirm that the experiment would be of adequate level and not too easy or too hard.

For this pilot study, ten participants were sourced from Prolific. That number of participants should be enough to meet our goals for this pilot and get a rough estimate of performance and time spent on tasks. In the pilot study, there were only plain logical reasoning questions, without AI advice, as the aim was to gain insight into question difficulty to humans. The experiment included thirty questions, which were randomly selected from the validation set of ReClor. The initially estimated time per question was around one minute, with the addition of the introduction, attention checks and other elements. In total the expected time spent on this pilot study was 40 minutes. In reality, the participants spend 33 minutes on average on the tasks. Therefore we could conclude that the time spent per question would be a bit below one minute. The participants received an hourly rate of £7.50 (the pay rate was £5.00 with the expected time of 40 minutes) and a bonus of £0.05 per correct answer. Of the ten participants, six failed no attention checks and were deemed usable for further analysis.

3.2.3. TASK SELECTION

The number of times a question was answered correctly by these six participants was used as a difficulty measure. The more times the question was answered correctly, the lower the difficulty. The desire was to create two batches of questions of equal difficulty. This was done by randomly selecting a question of six difficulty levels for each batch. The difficulty levels used were: zero times answered correctly, one time answered correctly etc. up until five times answered correctly. The batches of questions included no questions that were answered correctly by all six participants. In this way, two batches of six questions were created, while making sure there was no overlap in questions. For the tutorial, it was decided to only take four tasks. That was done to make sure that the tutorial would not be tiresome, as the experiment would become quite long. The difficulty levels chosen were, one time answered correctly until four times answered correctly. The

²www.prolific.co

decision on six questions per batch was also made to make the experiment somewhat shorter than the pilot study. In total, the main experiments include sixteen tasks rather than the thirty of the pilot.

3.2.4. TWO STAGE DECISION-MAKING

The experiment has a two-stage decision-making design. In Figure 3.1 and 3.2, the two stages are pictured. Figure 3.1 shows the first stage, the task is presented without AI advice. While in the second stage, shown in figure 3.2, AI advice is given and the own previous choice is shown under the header "Your choice". The AI advice given was simulated at random with an accuracy of 67% for the normal sets of questions and 50% in the tutorial stage. The accuracy of 67% is set such that AI outperforms the average participant, but still makes a significant amount of mistakes, making blindly relying on AI a bad strategy. To keep more control over the experiment the AI advice was simulated, the questions that would have incorrect AI advice were randomly selected and the advised answer was also selected at random.

Tasks

Context
 Physician: In comparing our country with two other countries of roughly the same population size, I found that even though we face the same dietary, bacterial, and stress-related causes of ulcers as they do, prescriptions for ulcer medicines in all socioeconomic strata are much rarer here than in those two countries. It's clear that we suffer significantly fewer ulcers, per capita, than they do.

Task 1/16
 Which one of the following, if true, most strengthens the physician's argument?

A The two countries that were compared with the physician's country had approximately the same ulcer rates as each other.

B The physician's country has a much better system for reporting the number of prescriptions of a given type that are obtained each year than is present in either of the other two countries.

C A person in the physician's country who is suffering from ulcers is just as likely to obtain a prescription for the ailment as is a person suffering from ulcers in one of the other two countries.

D Several other countries not covered in the physician's comparisons have more prescriptions for ulcer medication than does the physician's country.

AI advice
 D

Your choice
 A

Confirm and continue

Figure 3.2: Second stage task with AI advice

In this experiment, we obligated participants to choose without AI advice at first and only after that they received the AI advice. The two-stage method was used to allow us to collect data from the initial decisions that users had made. If we only showed the task with AI advice, we would not be had been able to tell whether they agreed with AI, because it could have been that they took the advice from the AI or that it was their own choice without considering the AI advice. The two stages allow us to calculate things like relative appropriate reliance on AI, which would not be possible without knowing the initial decision. Another thing to consider is that a single-step approach could reduce appropriate reliance on AI [41], therefore a two-step approach seems more suitable in this case. Furthermore, in this way participants are forced to first think for themselves and then reconsider with AI advice. Especially since considering the reconsideration with AI advice is important for the research, it is pragmatic to make this a clear step. While research shows people make these steps in their heads as well [42], this design imposes the reconsideration a bit more onto the participants.

3.3. QUALITY CONTROL

Several measures have been taken to ensure the quality of the experiment and the answers from the participants and to make sure they did put genuine consideration into their answers.

First of all, attention checks were included in which the participants were asked to select specific answers. Attention checks are mock-up questions, in which the participant is asked to select a specific answer, this is done to check whether they were reading the tasks and paying attention. Attention checks have been proven to be a useful way to increase the quality of data produced [43] and have been used in a wide variety

of human-computer interaction user studies [3, 15, 44, 45]. An attention check was placed once in a pre-task questionnaire, in which the participants were asked to select a specific number, the number that had to be selected was mentioned in the introduction, to make sure the introduction was read before starting the experiment. Other attention checks were in the tasks, these attention checks gave instructions to select a specific answer, while the rest of the task was exactly like the other questions. The instructions to select a specific answer were in the context paragraphs (once) and the questions (twice), this ensured that participants were reading the whole context and question before selecting an answer.

As the estimated time to answer a question was around one minute, it would be out of the ordinary if participants answered questions within a few seconds. To prevent that from happening a timer was set on the questions for 30 seconds, participants were not able to continue before this timer was done. For the AI advice stage, a similar timer was present, however, this timer was only set to be 5 seconds as the AI advice needs to be reconsidered but a total reevaluation of the task might not be needed. The timer on the attention checks was also shortened to 5 seconds as those questions did not need consideration after finding the check, but still needed a timer to prevent the question from being obviously different to participants that do not pay attention, as that might alert them that an attention check is taking place. The feedback stage has a 5-second timer as well, to make sure that it is not just clicked away without reading. Previous studies in human-AI decision making have used timers, to make sure participants at least spend some time on the task [7, 46]

Another piece of quality control was present in the order of the questions. The questions were shuffled To prevent a learning effect or any other undesired effect to happen. The questions are randomly shuffled within the batches. As the batches were balanced on the difficulty it was important to only shuffle within the batch to not undo the balancing. The order of batches was also randomly assigned, making sure that even if there was still some difficulty difference it was further balanced out. Furthermore, a bonus was given to participants for correct answers to give incentive for giving better answers.

3.4. TUTORIAL

For hypothesis 3 and 4 an intervention is needed to make participants aware of their miscalibration. We argue that making users aware of their miscalibration can be a way to successfully mitigate the DKE, as similar results have been found in previous research [9, 47]. The most prevalent way to increase the meta-cognitive skills, and therefore reduce the DKE, is by actually increasing the skill on the task overall. A way of doing this is in the form of a tutorial. The importance of feedback to reduce the DKE also stretches beyond human-computer interaction [38]. Tutorials have already been successfully used as an intervention in human-AI interaction research to increase the skill of participants [23, 48]. Besides increasing the skill of the participants, the tutorial is also meant to show why the given answer was incorrect, which gives insight into the complexity of the given task. This type of feedback should allow participants to better access how well they are performing [47]. The tutorial was done in a feedback style: the participants would first complete a task, thus first answering on their own without AI, after that being supplied with AI advice when they can make a new choice. Then they receive feedback if their answer (after AI advice) was incorrect if they gave a correct answer they could proceed to the next task after being notified that their previous answer was correct. The feedback did consist of a contrastive explanation, a contrastive explanation tells why the given answer was worse than the correct answer. It is a double-edging sword, not only telling why the correct answer is accurate, but also why the given answer was incorrect. Furthermore, the explanation needed to reveal their mistakes, to that purpose we made the explanations to be persuasive. The explanations were written manually by the authors as there were no off-the-shelf tools, known to us, strong enough that could give these persuasive explanations. . An example of a tutorial feedback page is given in Figure 3.3. The tutorial page has quite a similar layout to the AI advice page, with "AI advice" and "Your choice" being displayed on the right. The difference is that the "Your choice" in this task reflects the choice of the participant after receiving AI advice, it is to let the user reflect on the previous step. The correct answer is displayed on right in light blue, to make sure users can quickly see what the right answer was. The other elements of the task page are left in as previously such that participants can use the question, context and answers to better understand the feedback they receive. The explanation is located at the bottom of the page right above the continue button.

i

Tasks

Context

When doctors vaccinate a patient, their intention is to expose him or her to a weakened form of a disease-causing pathogen and thus to make the patient better able to resist the pathogen and less likely to develop a severe form of that disease later.

Correct answer
D

Task 8/16

Which one of the following best illustrates the principle that the passage illustrates?

A In some circumstances, firefighters use fire to fight fire by creating an intense explosion very close to an uncontrollable blaze that they wish to extinguish, thus momentarily depriving it of the oxygen it needs to continue burning.

B Some police departments energetically pursue those who commit minor crimes; in doing so they intend to provide examples to deter people who might be tempted to commit more-serious crimes.

C In some cases, a business will close down some of its operations, its intention being to position the company to be more profitable later even though this involves expenses in the current period.

D Some parents read their children fairy tales containing allegorical treatments of treachery and cruelty, with the intention of making them less emotionally vulnerable to these phenomena when they encounter them later in life.

AI advice
A

Your choice
C

The answer is D rather than C. In answer C the problem is solved by cutting down profit in the beginning but growing it later on. However, in this context, the problem is solved by giving exposure to a weakened version. In answer D the fairy tales also act as a weakened version of treachery and cruelty to which the children are exposed, thus making the principles align.

Confirm and continue

Figure 3.3: Example of a tutorial page

3.5. LOGIC UNIT BASED EXPLANATIONS

In the hypotheses section **hypotheses 1a, 2a, 3a and 4a**, all considered the addition of explanations to the AI. The idea behind this is that if participants have an explanation of the AI they can better understand the AI and therefore have a better estimation of the AI performance per task. While feature attribution methods like text highlighting of input are popular, research shows it is hard for humans to interpret this kind of information [49–51]. However, as highlighted sections in logical reasoning tasks have the potential for logical reasoning similar to human understanding, therefore, explanations on text spans of logic units are a good choice to reveal the way the AI system came to a decision. The selection of these logic units was done with the LogiFormer, which is presented by Xu *et al.* [52], their implementation can be found on their GitHub³. LogiFormer allows one to create explanations by creating logic units based on pre-trained language models. The LogiFormer collects the logic units from a graph transformer network for logical reasoning. The logic units in the network are text spans connected with causal relations. Furthermore, we relied on the self-attention matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ (n indicates the number of logic units) from the last layer in the graph transformer network, which followed from this interpretability design. Therefore, we used the following formula, in which E is the top- k logic units that received the most attention from other logic units, to discover the most important logic units:


$$E = \mathit{Argmax}_k \left(\sum_{j=1}^{j=n} \mathbf{A}_{ij} \right), \quad (3.1)$$

Our implementation and the extracted logic unit-based explanations can be found on the GitHub⁴ The LogiFormer was trained on the Reclor dataset, to get it accustomed to the specific dataset. The most important spans were then collected following Equation 3.1. In Figure 3.4 an example is given of how the logic units

³<https://github.com/xufangzhi/Logiformer>

⁴<https://github.com/LucieKuiper/Thesis>

are displayed to the participants. The highlighted spans can only appear in the context and answers and do not include the question. The k was set to five, to not overcrowd the tasks with too many parts highlighted. The spans can be created for each different answer, so for each task, four different sets of five spans exist. This allowed us to use it in our implementation as the AI advice answers were randomly altered to create our desired performance of the AI.



Tasks

Context

Physiologist: The likelihood of developing osteoporosis is greatly increased by a deficiency of calcium in the diet. Dairy products usually contain more calcium per serving than do fruits and vegetables. Yet in countries where dairy products are rare, and fruits and vegetables are the main source of calcium, the incidence of osteoporosis is much lower than in countries where people consume a great deal of calcium from dairy products.

Task 3/16

Which one of the following, if true, would most help to resolve the apparent discrepancy described by the physiologist?

<p>A A healthy human body eventually loses the excess calcium that it takes in.</p>	<p>B There are more people who have a calcium deficiency than there are who have developed osteoporosis.</p>	<p>AI advice</p> <p>C</p>
<p>C The fats in dairy products tend to inhibit the body's calcium absorption.</p>	<p>D Many people who eat large quantities of fruits and vegetables also consume dairy products.</p>	<p>Your choice</p> <p>D</p>

Confirm and continue

Figure 3.4: Example of an AI advice page with XAI

4

EXPERIMENTAL SET-UP

This chapter outlines the experiment, this is done by first talking about the four conditions that are being studied in the experiments. Next, the variables and the measures needed are described. After that, the participants of the experiment are characterised. Finally, the complete procedure of the experiments is illustrated.

4.1. CONDITIONS

To verify the hypos, we considered four experimental conditions in this study. The difference between the conditions is whether participants receive a Tutorial and/or explainable AI (XAI). Therefore, the four conditions were as follows:

- **Condition 1:** no tutorial and no explanations during AI advice. Depicted as: × Tutorial, × XAI
- **Condition 2:** with tutorial and no explanations during AI advice. Depicted as: ✓ Tutorial, × XAI
- **Condition 3:** no tutorial and with explanations during AI advice. Depicted as: × Tutorial, ✓ XAI
- **Condition 4:** with tutorial and with explanations during AI advice. Depicted as: ✓ Tutorial, ✓ XAI

The Analysis of the four experimental conditions enabled us to be more exact about what factor contributed to the results. The four conditions were needed to cover all the hypotheses. Hypothesis 1 and hypothesis 2 could be covered by the first condition. However, we argued that the other three conditions could also be used to prove these hypotheses as for the first batch of questions the tutorial has not been shown yet and thus has no effect. The conditions with XAI are also taken into account overall, as the hypothesis is only about overestimating and does not mention XAI. For hypotheses 1a, 2a, 3a and 4a, it was necessary to have conditions 3 and 4 as the difference between with and without XAI needs to be measured. For hypothesis 3 and hypothesis 4 condition 2 and condition 4, the ones with a tutorial, were important as these hypotheses argue certain effects when users are made aware of their mistakes. This making aware was done by giving the participants a tutorial.

The main difference between the with or without tutorial was whether participants received a tutorial during the experiment. This tutorial consists of four questions with feedback. In order to keep the conditions as similar as possible, participants in the without tutorial conditions received the same four questions, the only difference being that they did not receive any feedback on their answers. Which was done to circumvent results that would happen due to the learning effect. Two conditions had XAI, this consisted of five highlighted logic unit spans that are shown during the AI advice stage. At the end of the experiment, These participants were asked whether they found the explanations helpful. This was done at the very end of the experiment and thus should not affect the other parts. Therefore, it is fair to say that the explanations were the reason for this difference in case these conditions have significantly different results from the conditions without explanations.

4.2. MEASURES AND VARIABLES

To answer the hypotheses some measures and variables were essential. Before deploying the experiment these measures and variables had to be known, such that they could all be collected during the experiment.

For all hypotheses, it is important to know the measurement of self-estimation. The self-estimation is determined during the self-assessment questionnaire. The participants were asked, directly after both batches of questions, how many questions they thought they answered correctly.

$$\text{Self-estimation} = \text{Number of correct answers user estimates to have given} \quad (4.1)$$

This answer could then be used to tell whether participants had an accurate self-estimation or had an under or over-estimation of themselves. To make things less complex the degree of miscalibration was calculated, which took the self-estimation and the number of right answers into consideration. The formula to calculate the degree of miscalibration can be found in Equation 4.2. When the degree of miscalibration > 0 , one overestimates themselves, a degree of miscalibration < 0 indicates underestimation and a degree of miscalibration $= 0$ shows an accurate estimation of one's own performance.

$$\text{Degree of Miscalibration} = \text{Self-estimation} - \text{Number of actual correct answers} \quad (4.2)$$

Besides the measurement of the estimation of themselves, the participants in the experiment were also asked, during the self-assessment questionnaire, to answer how well they thought that other participants had performed and how they would rate themselves compared to other participants. Asking them how they rated themselves, was done by asking "*How do you estimate your own performance (after AI advice) compared to your peers from 0-100%. 0% being the worst, 50% exactly average and 100% the best one*".

$$\text{Estimation of others} = \text{Number of correct answers user estimates that others have given on average} \quad (4.3)$$

$$\text{Comparison of Self to Others} = \text{Percentage input 0\% being the worst 50\% average and 100\% the best} \quad (4.4)$$

To answer hypothesis 1 a measure was needed to determine the reliance on the AI system. In this case, the reliance was determined with the agreement fraction and switch fraction. These measures have previously been used in other human-computer interaction research [15, 53]. The agreement fraction calculation is shown in Equation 4.5. Where the number of decisions where the user agreed with the AI advice is divided by the total amount of decisions made, this gave us the fraction of how often the user agreed with the AI advice that was given.

$$\text{Agreement fraction} = \frac{\text{Number of decisions same as the AI system}}{\text{Total number of decisions}} \quad (4.5)$$

The calculation of the switch fraction is shown in Equation 4.6. In which the number of decisions where the user switched is divided by the number of decisions with initial disagreement. It is similar to the agreement fraction in the sense that it looks at how often the participant agrees with the AI. However, the big discrepancy is that the questions in which the participant chose the same answer as AI before getting AI advice are not taken into account. To be able to measure this it was important to collect the answers that participants gave before being given AI advice as a helping tool, this gave us more insight into the reliance relation of the participant to the AI.

$$\text{Switch fraction} = \frac{\text{Number of decisions where the user switched to agree with the AI system}}{\text{Total number of decisions with initial disagreement}} \quad (4.6)$$

These equations were separate variables and would not be added together, it is expected that both these variables are lower for participants that overestimate themselves, as a lower score on the agreement and switch fraction, shows lower reliance on the AI system. When only one of these variables would be significantly lower no definite conclusions could be drawn.

In hypothesis 2 a difference in appropriate reliance was looked for, as we were not only interested in reliance, but also whether that reliance was appropriate. Inspired by the paper by Schemmer *et al.* [6] we used their proposed measures to calculate appropriate reliance on AI, the appropriate reliance was determined with two variables the relative positive AI reliance (RAIR) and the relative positive self-reliance (RSR). In Table 4.1 the four different types of reliance are shown, consisting of different combinations of the correctness of the user their initial decision, the correctness of the AI advice, and the correctness of the final answer from the user.

Table 4.1: Appropriate reliance patterns considered by Schemmer *et al.*

Initial user decision	AI advice	Final decision after AI advice	Reliance
Incorrect	Correct	Correct	Positive AI reliance
Incorrect	Correct	Incorrect	Negative self reliance
Correct	Incorrect	Correct	Positive self reliance
Correct	Incorrect	Incorrect	Negative AI reliance

The calculation for the RAIR can be seen in Equation 4.7. The positive AI reliance, when the AI advice is correct and the initial decision was incorrect and the user switched to be correct, is divided by positive AI reliance plus negative self-reliance. The negative self-reliance can be described as when the initial decision was incorrect and the AI advice was correct but the user did not switch to be correct. In essence, positive AI reliance plus negative self-reliance are all cases in which the initial decision was wrong and the AI advice was correct. By dividing the positive AI reliance over all cases in which the AI advice was correct and the initial human decision is incorrect, we get the fraction of how likely it is that one accurately reconsiders their answer after receiving AI advice. This gives us insight into whether the participants made an accurate choice when choosing to switch to the AI advice, the outcome will reflect how appropriate the reliance is, and how often the right decision was made.

$$\text{Relative Positive AI Reliance (RAIR)} = \frac{\text{Positive AI reliance}}{\text{Positive AI reliance} + \text{Negative self reliance}} \quad (4.7)$$

The RSR calculation is shown in Equation 4.8. The positive self-reliance, when the initial answer is correct and the AI advice was incorrect and the user keeps their answer, is divided by positive self-reliance plus negative AI reliance. Negative AI reliance is when the initial decision was correct the AI advice was incorrect and the user switched to an incorrect answer. By doing this calculation we are left with a fraction that shows how likely it is given that the AI advice was incorrect and the initial decision was correct for the user to stay with their own decision. A higher fraction shows appropriate reliance as it indicates that the user is stuck with their answer when correct, not trusting the AI when it should not be trusted. A higher fraction, therefore, shows a better appropriate reliance on AI. The RSR could thus increase our insight into the appropriate reliance relation of the user to the AI system.

$$\text{Relative Positive Self Reliance (RSR)} = \frac{\text{Positive self reliance}}{\text{Positive self reliance} + \text{Negative AI reliance}} \quad (4.8)$$

These equations are separate variables and were not added together, it was expected that both these variables were higher for participants that have an accurate self-assessment. When only one of these variables would be significantly higher no definite conclusions could be drawn.

While the four hypotheses did not require an accuracy measure, the accuracy of the participants has also been included as a variable. The accuracy shows us how well the human-AI team performed and gave further insight.

$$\text{Accuracy} = \frac{\text{Number of questions answered correctly after receiving AI advice}}{\text{Total number of questions}} \quad (4.9)$$

Besides these variables, some more measurements were taken. For instance, the participants that received explanations during AI advice were asked whether this was perceived as helpful with the question: "*To what extent was the explanation (i.e., the highlighted words/phrases) helpful in making your final decision?*". The participants could then answer this question on a Likert scale ranging from 1: not helpful to 5: very helpful.

Furthermore, a few surveys on trust were done. First of all, from the Trust in Automation questionnaire [54], the Propensity to Trust (TiA-PtT) and the Trust in Automation (TiA trust) were put in the study. These surveys were done after both batches of questions. By including these surveys we acknowledged and could measure the part that trust could play in the effects found. Besides trust, we also considered the participants' affinity with technology, as this could affect their reliance on AI systems [55]. The survey chosen to measure this affinity was the Affinity for Technology Interaction Scale (ATI) [56], this survey was done at the beginning of the experiment before the tasks

Table 4.2: Variables considered in our experimental study. “DV” refers to the dependent variable.

Variable Type	Variable Name	Value Type	Value Scale
Performance (DV)	Accuracy	Continuous, Interval	[0.0, 1.0]
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]
Assessment (DV)	Degree of Miscalibration	Continuous, Interval	[-6,6]
	Self-Estimation	Continuous, Interval	[0,6]
	Estimation of Others	Continuous, Interval	[0,6]
	Comparison of Self to Others	Continuous, Interval	[0,100]
Trust (DV)	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: tend to trust
Other	Helpfulness of Explanation	Likert	5-point, 1: not helpful, 5: very helpful

4.3. PARTICIPANTS

The participants used in the experiments were sourced from Prolific¹ To establish the number of participants needed in the experiment, a power analysis for a Between-Subject ANOVA using G*power was computed [57]. As we test multiple hypotheses we must correct the α , this is done with the Bonferroni correction since we test four hypotheses which resulted in $\alpha = \frac{0.05}{4} = 0.0125$. The effect size was set on a moderate effect and that gave $f = 0.25$. Furthermore, considering the statistical power of $(1 - \beta = 0.8)$ and taking into account the four different experimental conditions. Combining these factors resulted in a required size of 244 participants. When recruiting participants we had to take into account that some exclusions would be made. Therefore the number of participants was increased per condition until enough participants for each condition were collected.

To ensure the quality of the data received some constraints were set on the suitable participants. In prolific we requested the participants to be proficient English speakers to assure the logical reasoning tasks were not hindered by a lacking understanding of the language. Another constraint was set on an acceptance rate of 90% or higher, to only allow the top performers, which is an important factor in receiving higher quality results [43]. Furthermore, the gender was balanced, meaning prolific would even the group out to be almost a fifty-fifty split of male and female. The balancing of gender is important as previous research suggests that self-assessment and the overconfidence bias can be different based on gender [58, 59].

As in the pilot, it was determined that the questions took around one minute the experiment was set to be around 20 minutes with 16 tasks. The desired base rate payout was of **£7.50 an hour**, as the participants were expected to spend around 20 minutes on the tasks the payout was **£2.50** per accepted participant. To give the incentive to perform well on the tasks a bonus was included of **£0.10** per correctly answered question after receiving AI advice. This vulnerability of the participant is needed to give a risk factor to relying on AI. Therefore the maximum payout per participant was **£4.10** if they answered all questions correctly.

4.4. PROCEDURE

With all elements of the experiment described the only thing left is the procedure of the experiment. The experiment was slightly different depending on the condition, but the overall procedure was rather similar. In Figure 4.1 the procedure of the \checkmark Tutorial, \checkmark XAI is shown. A walkthrough of the complete experiment is as follows:

First of all, the participants were shown an introduction text that also included the informed consent. The timer only allowed the participant to proceed after 5 seconds to make sure they had read the information. Following the introduction were some questionnaires first ATI, in which the participants were asked to answer nine questions about their affinity with technology (on a scale of 1 = completely disagree, 2 = largely disagree, 3 = slightly disagree, 4 = slightly agree, 5 = largely agree and 6 = strongly agree). The questionnaire also included an attention check. This attention check asked how many tasks there are in the experiment, this should be known from the introduction text. The options were 1, 2, 3, 16, 80 and 240 to make sure the other answers were so far off that they would not be confusing. The continue button was enabled after 30 seconds. The next questionnaire was TiA-PtT, this questionnaire consisted of three questions (on a scale of: 1

¹www.prolific.co

= Strongly disagree, 2 = Rather disagree, 3 = Neither disagree nor agree, 4 = Rather agree, 5 = Strongly agree). The participants could proceed after six seconds.

After the introduction and questionnaires, the participants could start the tasks next. The participants were given a task in which they had to answer a question on the context, they got four possible answers and had to choose the best fitting answer. During this batch of questions, the participants were given one attention check, that looked similar to the other tasks. The continue button was enabled after 30 seconds. following the task is the same task but this time with AI advice, participants have to reconsider their answer now being aware of what choice the AI recommends. For the \times Tutorial, \checkmark XAI and the \checkmark Tutorial, \checkmark XAI conditions, the participant also got highlighted spans to reveal the inner workings of the AI system. The timer on the AI advice stage was set to five seconds. From the AI advice, the participants go to the next task until they completed the six tasks from the first batch.

When the complete batch of tasks was done the participants had to complete a self-assessment. The assessment consisted of three questions. In the first question, the participants were asked how many questions they estimated to have answered correctly, in the second question they were asked how many they estimate that others had correct, the last question asked how they would rate themselves against the other participants on a 0-100% basis. For each question, it took two seconds before the continue button was enabled. Directly after the self-assessment, there was a questionnaire on TiA-trust. The participants were asked to answer two questions on trust (on a scale of: 1 = Strongly disagree, 2 = Rather disagree, 3 = Neither disagree nor agree, 4 = Rather agree, 5 = Strongly agree). Participants could proceed after four seconds.

Following the self-assessment and questionnaire was the tutorial. The tutorial phase was different depending on the condition, the \times Tutorial, \times XAI and \times Tutorial, \checkmark XAI groups were just given tasks in the same way as the first batch of tasks, only this time consisting of only four tasks and one attention check. The \checkmark Tutorial, \times XAI and \checkmark Tutorial, \checkmark XAI were likewise supplied with four tasks and one attention check. However, the participants in these groups were given feedback after submitting their answers with AI advice, this feedback consisted of either an explanation of why the correct answer was better if their answer was wrong or a screen telling them their previous answer was correct. To make sure the participants read the feedback a timer of five seconds was put in.

Once the tutorial was done participants received an additional batch of tasks in the same way as the first batch of tasks. Thus six tasks each task first by themselves and then with AI advice. After these tasks, another self-assessment and TiA-trust questionnaire followed. The two separate self-assessments and questionnaires allowed us to measure the difference between the two batches, and thus also between before and after the tutorial. Lastly, at the end of the experiment participants from the \times Tutorial, \checkmark XAI and \checkmark Tutorial, \checkmark XAI were asked how helpful they perceived the high lighting explanations on AI. This was asked (on a scale of: 1 = Not at all helpful, 2 = Slightly helpful, 3 = Rather helpful, 4 = Helpful, 5 = Very Helpful). This question had a timer of two seconds before participants could finish the experiment.

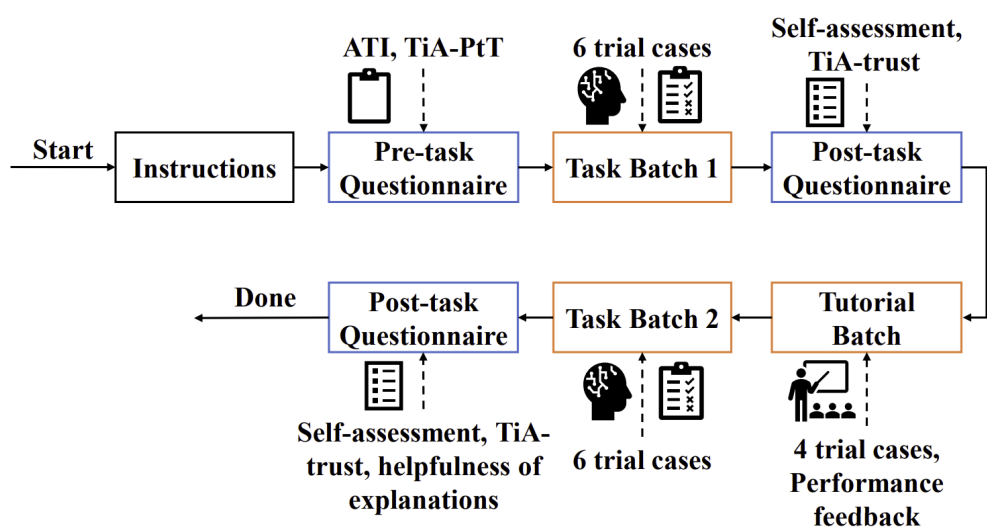


Figure 4.1: Illustration of procedure followed in our study for a \checkmark Tutorial, \checkmark XAI condition participant. Blue boxes represent questionnaires, orange boxes represent parts involving tasks

5

RESULTS

In this chapter, we will show the results found by carrying out the experimental set-up (chapter 4). The participants for these experiments were gathered from the crowdsourcing platform Prolific¹. Not all submissions could be used for analysis due to various reasons, and those were cleared from the study. The focus was foremost on finding answers to the hypotheses mentioned in the experimental setup. Furthermore, exploratory analysis has been done to gain a better insight into the results and their implications.

5.1. CLEANING DATA

Not all the data that has been collected from Prolific could be used. To be left with genuine data only, all data that could be less reliable was removed. In total, we had 375 entries in the database after all experiments were completed. The experiment was run one by one per condition, to ensure enough reliable entries were collected for each condition. The explanation of why entries were deleted and by how many are given below.

- In total 44 entries did start on the experiment but did not finish all logical reasoning questions, without being removed due to failed attention checks. Most of them (37) completed less than five questions, and sixteen of those did not even complete one question. This might indicate some participants decided to opt out, this could, for example, be due to question difficulty or time taken for the initial first few questions. Incomplete entries were not used and the participants were not paid for those efforts.
- Sixteen entries missed two attention checks. After missing the second attention check these participants were automatically removed from the study. These entries have not been used in further study as by missing two attention checks there is a sizeable chance the participant was not paying attention to the non attention check questions and surveys. They were not compensated for these efforts.
- Another, 61 entries failed one attention check and were therefore deemed unreliable. These entries have been removed from the data. The participants that only missed one attention check were allowed to continue the research and were paid for their efforts.
- Two entries miss the ATI and TiA-PtT questionnaires, this was possible if one reloaded the entry page. For conditions tested later on this problem was fixed, reloading the page would not result in going to the questions immediately anymore. To keep consistency these entries were removed. The participants were paid for their efforts.
- One entry has no self-assessment or TiA questionnaire completed. This participant is removed from the study, as without these answers their data is obsolete. This participant did not validly finish the experiment and therefore did not get any compensation.
- Another entry was removed for missing just the last Self-assessment and TiA questionnaire. This participant was manually approved and did get compensation.
- Lastly an entry was removed as it was not registered in prolific and thus we could not track the origin.

¹<https://www.prolific.co/>

In total this resulted in 126 entries that needed to be removed from the dataset, leaving us with 249 entries. These entries were distributed evenly over the four experimental conditions as follows: 63 (\times Tutorial, \times XAI), 62 (\checkmark Tutorial, \times XAI), 62 (\times Tutorial, \checkmark XAI), 62 (\checkmark Tutorial, \checkmark XAI).

5.2. DESCRIPTIVE RESULTS

As mentioned in the previous section 249 participants were used for the analysis. To get a broad overview of these participants some overall analysis has been done. The participants spend around 32 minutes with a standard deviation of 13 minutes on the tasks that we provided them. As only 16 tasks were in the experiment, this time was a bit longer than expected. The distribution of covariates was as follows: ATI had a median of 3.73 with a standard deviation of 0.99 and TiA-Propensity to Trust had a median of 2.95 and a standard deviation of 0.60. The participants had an overall performance of 56.9% with a standard deviation of 16.0 % after they received help from AI advice. This is lower than the performance of the AI which was 66.7% for these tasks. However, it is higher than the initial performance before receiving AI advice of 41.4% with a standard deviation of 18.9 %. Thus we can already see that the overall performance increases by getting AI advice, but that it still is not as high as the AI performance. Prolific was set to hire an equal amount of female and male participants to have this factor balanced. As some participants got removed during the cleaning stages this ratio could have become unbalanced. The distribution of gender was as follows: 128 male participants and 121 female participants. Another variable at is the age which had a mean of 38.0 with a standard deviation of 12.9. Additionally, we looked at the nationalities as cultural differences can affect the confidence of participants [59]. The distribution of the nationalities of the participants is shown in Figure 5.1a. The United Kingdom is by far the most represented among the participants. Furthermore, the helpfulness of XAI was assessed only for the participants that received XAI, by questioning them on their experience with the XAI. The participants were not positive about the helpfulness of XAI, 18 participants answered that the XAI was not helpful at all, 49 that it was slightly helpful, 22 rather helpful, 27 helpful and only 8 very helpful, this distribution is shown in Figure 5.1b. Lastly, the distribution of the self-estimation over the conditions in Figure 5.1c. These distributions are for the first batch of questions as those are used for the first two hypotheses (sections 5.4.1 and 5.4.2). This means that the conditions with a tutorial have not completed the tutorial yet. Correct estimation is defined as having a miscalibration of zero, thus having the same estimated performance as actual performance.

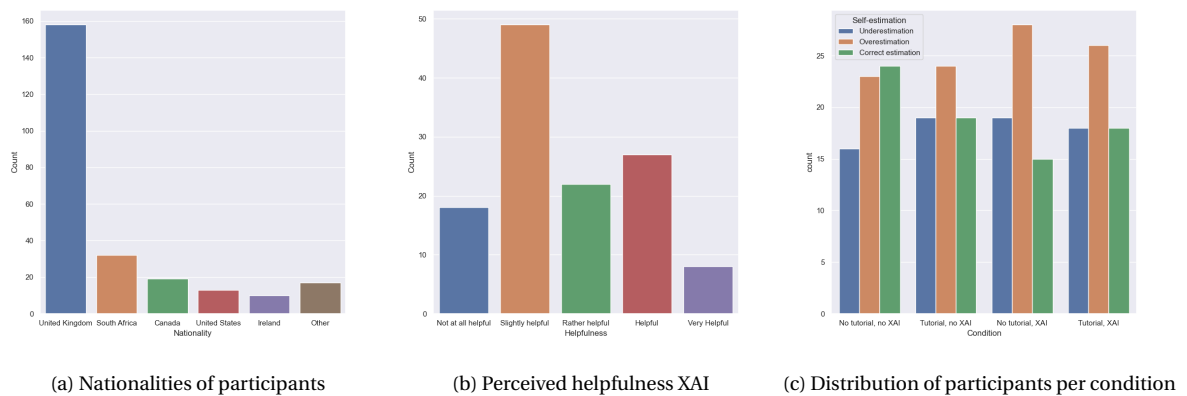


Figure 5.1: Graphs on descriptive results shows the distribution of (a) participants over the nationalities, (b) the perceived helpfulness of logic units-based explanations, (c) overestimated self-estimation, accurate self-estimation and underestimated self-estimation across all experimental conditions over the first batch of tasks.

5.3. TYPE OF TESTS

To determine what type of hypotheses tests can be used, we must know the distributions, as normal and non-normal distributions use different tests. Tests for normal distributions hold more power but should be used on normal distributions only. Tests for non-normal distributions can be used for all types of distributions, however, they are less powerful. An example of distributions found during the research of this thesis can be found in Figure 5.2. More graphs of distributions can be found in the appendix (appendix B). From a visual inspection, we can tell that the histograms in Figure 5.2 are most likely not normally distributed. To confirm

that they are not normally distributed a *Shapiro-Wilk* test is used, which calculates the likelihood that a distribution is normally distributed. With all *Shapiro-Wilk* tests having a $p < 0.0001$ we can confirm with great certainty that the distributions are non-normally distributed. Therefore, the tests used in this thesis are those suitable for non-normal distributed data. When comparing two distributions dealing with between-subject the *Wilcoxon-Mann Whitney* test is used. In the cases of a within-subject, the *Wilcoxon signed ranks* test is used to compare the distributions.

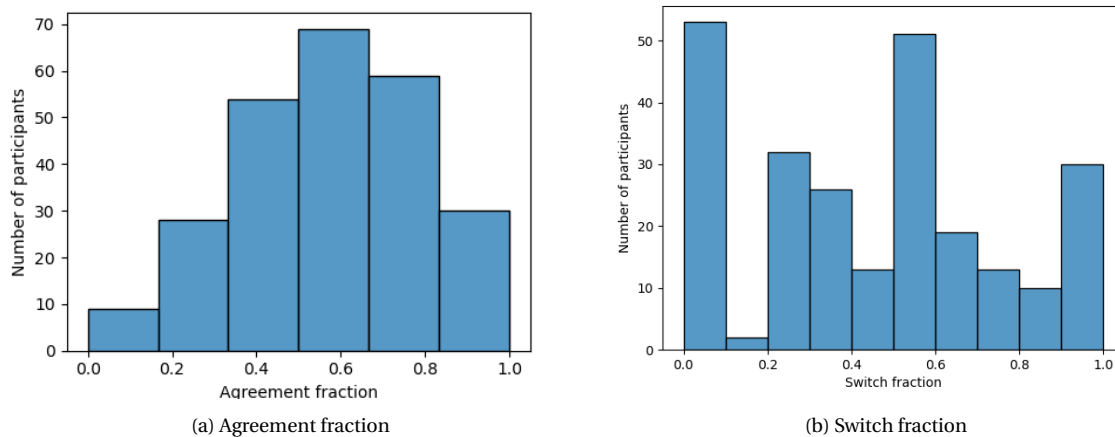


Figure 5.2: Distribution of agreement and switch fraction of participants

5.4. HYPOTHESES

This section describes the results found from the hypothesis tests and some extra tests done per hypothesis. In the end, a summary is given to shortly sum up the results found in this section, and thus the conclusions we found for each hypothesis.

5.4.1. HYPOTHESIS ONE

Before conducting any hypothesis test it is important to know what we are trying to test for. **Hypothesis 1** states the following: *Participants overestimating their own performance rely less on AI systems*. To give conclusions to this hypothesis we must establish the interpretation of overestimation and reliance on AI.

- **Overestimating:** Degree of miscalibration < 0 .
- **Reliance on AI:** Is determined with the agreement fraction and switch fraction definitions can be found in the measures and variables (section 4.2). The agreement fraction shows how often the final answer complied with the advice given. Switch fraction looks at how often one switches to agree with the AI when a participant's initial answer did not correspond. For both of these fractions, a higher fraction indicates that the participant relies more on AI.

In these tests, only the first six tasks (first batch) the participant answered are considered and the first self-assessment. The choice to only use the first six questions is to prevent having an effect from the tutorial stage. In the first six questions, none of the participants has seen the tutorial. Therefore, we can include participants from all conditions.

The group of participants will be split in two to be able to give answers to the hypothesis. The first group exists out of all participants that overestimated themselves, the second group consists of all participants that had a right or underestimation of themselves. This division resulted in **101** participants that overestimated themselves, and **148** that had a correct or underestimation of themselves. As all tests done in this hypothesis are between subjects, the *Wilcoxon-Mann Whitney* test is used to determine the results.

HYPOTHESIS TEST

The results from the *Wilcoxon-Mann Whitney* tests can be found in Table 5.1. In addition to the previously discussed agreement fraction and switch fraction, the results also give accuracy, relative positive AI reliance

(RAIR) and relative positive self-reliance (RSR). The other metrics make a more complete story but are not necessary for proof of the hypothesis. With the Bonferonni correction an α smaller or equal to $0.0125 (\frac{0.05}{4})$ is deemed significant. As the p-value of both the agreement and the switch fraction is below this threshold of 0.0125, there is a significant difference in the two groups compared, with the median being higher for the participants that did not overestimate themselves. Therefore, we can say that **hypothesis 1 is supported** by the results. Users that overestimate their own performance also have a significantly lower reliance on the AI system.

Furthermore, the accuracy, RAIR and RSR are also significantly higher for the participants that did not overestimate themselves. Indicating that besides having a lower reliance, participants that overestimate themselves also have lower accuracy and appropriate reliance.

Table 5.1: Wilcoxon-Mann Whitney test results for **hypothesis 1**. Comparison between participants that did overestimate themselves and participants that did not. Significant results are shown in bold.

Dependent Variables	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (Overestimation)	<i>M</i> ± <i>SD</i> (Other)
Accuracy	3229.0	<.0001	0.449 ± 0.190	0.660 ± 0.168
Agreement Fraction	5683.0	.0010	0.594 ± 0.237	0.695 ± 0.198
Switch Fraction	4809.5	<.0001	0.325 ± 0.323	0.515 ± 0.298
RAIR	3811.0	<.0001	0.273 ± 0.332	0.613 ± 0.372
RSR	5241.0	<.0001	0.267 ± 0.433	0.544 ± 0.477

EXPLAINABLE AI (XAI)

Hypothesis 1 can be supported by the results, however, besides that hypothesis **hypothesis 1a** was formulated which states that supplying users with XAI during the AI-advice stage would increase this effect. To test this hypothesis a *Wilcoxon-Mann Whitney* test is used. We therefore must split the group into a group that received XAI and a group that did not receive XAI. This comparison was done for the group that overestimated themselves and the group that did not overestimate themselves separately, resulting in four different groups. The participants are distributed as follows: no XAI overestimation of themselves: 47, no XAI underestimation or correct estimation of themselves: 78, XAI overestimation of themselves: 54, no XAI underestimation or correct estimation of themselves: 70. As the p-values exceed the α of 0.0125 we can not support the hypothesis that XAI would reduce the effect found in **hypothesis 1**. Therefore, there is **no support for hypothesis 1a**.

In the groups, without overestimation, we did find a significant difference in RAIR for participants with and without XAI. This indicates that participants that do not overestimate themselves might benefit from the XAI in gaining appropriate trust. However, as RSR has no significant difference, this shows only partially improved appropriate reliance.

Table 5.2: Wilcoxon-Mann Whitney test results for **hypothesis 1a** comparison between participants that did receive XAI and participants that did not. Compared to both participants that overestimated themselves and participants that did not. Significant results are shown in bold.

Participants Dependent Variables	With Overestimation				Without Overestimation			
	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (XAI)	<i>M</i> ± <i>SD</i> (no XAI)	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (XAI)	<i>M</i> ± <i>SD</i> (No XAI)
Accuracy	1129.5	.3270	0.429 ± 0.206	0.472 ± 0.168	2537.5	.4402	0.648 ± 0.164	0.671 ± 0.172
Agreement Fraction	1071.5	.1710	0.562 ± 0.243	0.631 ± 0.228	2624.5	.6778	0.688 ± 0.198	0.701 ± 0.198
Switch Fraction	1087.5	.2071	0.285 ± 0.300	0.371 ± 0.345	2471.5	.3184	0.489 ± 0.308	0.538 ± 0.289
RAIR	1141.5	.3549	0.247 ± 0.321	0.303 ± 0.346	2321.5	.0001	0.581 ± 0.379	0.642 ± 0.364
RSR	1139.5	.2657	0.222 ± 0.408	0.319 ± 0.460	2681.0	.8355	0.536 ± 0.476	0.551 ± 0.481

5.4.2. HYPOTHESIS TWO

Hypothesis 2 states: Accurate self-assessment will result in more appropriate reliance on AI systems. To establish this we must establish the interpretation of accurate self-assessment and appropriate reliance.

- **Accurate self-assessment:** Degree of miscalibration = 0.
- **Appropriate reliance:** Is determined with RAIR and RSRS. The formulas used to calculate them can be found in the measures and variables (section 4.2). RAIR shows how well a participant trusts the AI system when it is appropriate. On the other hand, RSR shows how appropriate their reliance on themselves is. For both these measures higher fraction indicates a more appropriate trust.

Just as with the previous hypothesis, only the first batch is being considered to prevent having an effect from the tutorial stage. The group is split into two, participants with an accurate self-estimation and participants with an inaccurate self-estimation. In total **76** participants had an accurate self-estimation and **173** had an inaccurate self-estimation. The tests done for this hypothesis are all between-subject tests and therefore the *Wilcoxon-Mann Whitney* test is used.

HYPOTHESIS TEST

The results from the *Wilcoxon-Mann Whitney* tests are shown in Table 5.10. Besides RAIR and RSR, the accuracy, agreement fraction and switch fraction are included. An α of 0.0125 needs to be reached to prove a significant difference. Only the switch fraction and RAIR show a significant difference, being higher for participants with an accurate self-estimation. For this hypothesis, we are interested in RAIR and RSR and hypothesised that both would be significantly higher for the participants with accurate self-estimation. Interestingly RSR is lower for participants with accurate self-estimation, however, this is not a significant result. As can be seen in Figure B.2 the data points were mostly clustered at zero and one, with a few points being at 0.5. With RAIR being significant and RSR not supporting the hypothesis, there is only **partial support for hypothesis 2**. To prove or disprove hypothesis 2, more experiments need to be done. The accuracy and agreement fraction seem to be higher for the accurate group, but this is not significant.

Table 5.3: Wilcoxon-Mann Whitney test results for **hypothesis 2**. Comparison between participants that had an accurate estimate of themselves and participants that did not. Significant results are shown in bold.

Dependent Variables	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (Accurate)	<i>M</i> ± <i>SD</i> (Inaccurate)
Accuracy	7269.0	.1703	0.605 ± 0.153	0.561 ± 0.223
Agreement Fraction	7393.0	.1090	0.691 ± 0.214	0.638 ± 0.221
Switch Fraction	8087.5	.0035	0.526 ± 0.311	0.399 ± 0.319
RAIR	7966.5	.0062	0.577 ± 0.375	0.431 ± 0.393
RSR	6473.0	.8284	0.421 ± 0.476	0.436 ± 0.481

XAI

Hypothesis 2 is only partially supported by the results. Additionally, there is another hypothesis involved (**hypothesis 2a**) that claimed that the effect would be significantly bigger with XAI being provided. The group is divided similarly to **hypothesis 2**. However, these groups are looked at separately and within these groups, they are split on whether they did receive XAI or not. The primary focus will be on the participants with accurate self-estimation, as we expect an increase in effect in this group. The distribution of the groups is as follows: no XAI accurate self-estimation 43, no XAI inaccurate self-estimation 82, XAI accurate self-estimation 33, and XAI inaccurate self-estimation 91. Both the accurate and inaccurate groups are divided roughly by half in each group, indicating that obtaining an accurate self-estimation is not significantly influenced by being provided XAI. If the *p*-values exceed the α of 0.0125 we can not support the hypothesis that XAI would increase the effects of the hypothesis. That is in this case true for all but RAIR for inaccurate self-estimation. This can indicate that participants that did not have an accurate self-estimation, trusted the AI more appropriately due to XAI while not significantly trusting themselves more appropriately. Therefore, there is **no support for hypothesis 2a**.

Table 5.4: Wilcoxon-Mann Whitney test results for **hypothesis 2a**. Comparison between participants that did receive XAI and participants that did not. Compared to both participants that had an accurate estimation themselves and participants that did not. Significant results are shown in bold.

Participants Dependent Variables	Accurate self-estimation				Inaccurate self-estimation			
	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (XAI)	<i>M</i> ± <i>SD</i> (no XAI)	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (XAI)	<i>M</i> ± <i>SD</i> (No XAI)
Accuracy	616.5	.3038	0.581 ± 0.157	0.624 ± 0.150	34449.0	.3800	0.5421 ± 0.230	0.581 ± 0.215
Agreement Fraction	596.0	.2248	0.657 ± 0.216	0.717 ± 0.211	3527.5	.5268	0.625 ± 0.231	0.652 ± 0.209
Switch Fraction	555.0	.1040	0.457 ± 0.329	0.578 ± 0.290	3440.0	.3718	0.379 ± 0.316	0.421 ± 0.323
RAIR	568.0	.1282	0.500 ± 0.384	0.636 ± 0.361	1330.5	.0022	0.412 ± 0.393	0.451 ± 0.396
RSR	712.0	.9812	0.424 ± 0.486	0.419 ± 0.475	3350.0	.1912	0.390 ± 0.470	0.489 ± 0.490

5.4.3. HYPOTHESIS THREE

Hypothesis 3 proposes that: *Making users aware of their miscalibrated self-assessment, will improve their self-assessment*. To test this hypothesis a way of making users aware of miscalibrated self-assessment must be

established. Furthermore, it is also necessary to have a clear understanding of what is deemed an improved self-assessment.

- **Tutorial:** a tutorial has been used to make the users aware of their miscalibrated self-assessment. This tutorial has only been provided to two conditions (\checkmark **Tutorial**, \times **XAI** and \checkmark **Tutorial**, \checkmark **XAI**).
- **Improved self-assessment:** an improved self-assessment is defined as a reduction of miscalibration of the self-assessment. For this hypothesis participants who had no miscalibration in the first self-assessment (e.g. their self-assessment had the same score as their actual result) were excluded. They could not be made aware of their miscalibrated self-assessment, as there was no miscalibration to be made aware of.

For the third hypothesis, the miscalibration of the first six tasks (first batch) will be compared to the miscalibration of the last six tasks (last batch). After all, only two conditions are used and some participants are excluded due to correct self-assessment after the first batch. Considering these exclusions **87** participants are left in total. The comparison that is being made is within the subjects, their first batch compared to the last batch. Therefore, the *Wilcoxon signed ranks* test is used for the hypothesis testing. To compare the influence of the tutorial and XAI a between-subject test is performed, this is done with a *Wilcoxon-Mann Whitney* test.

HYPOTHESIS TEST

The results from the *Wilcoxon signed ranks* test are given in Table 5.5. From the median, we can see that the last batch has a lower miscalibration than the first batch. For the median and standard deviation, it is important to note that these values could range between 0-1, and for the first six between 0.167-1 (as participants without initial miscalibration were excluded). The change in the miscalibration of self-assessment between the first batch and the last batch is significant. The p-value is below the Bonferonni corrected α of 0.0125. Thus, **hypothesis 3 is supported**. Participants with initial miscalibration have a reduction of their miscalibration after receiving a tutorial.

Table 5.5: Wilcoxon signed ranks test results for **hypothesis 3**. Comparison between the first and last batch of tasks the participant performed. Significant results are shown in bold.

Dependent Variable	T	p	$M \pm SD$ (First batch)	$M \pm SD$ (Last batch)
Mis-calibration	365.0	.0005	0.278 \pm 0.152	0.190 \pm 0.175

XAI

Likewise to the other hypotheses, we have to consider **hypothesis 3a**, which states that the effect would be bigger for participants that were supplied with XAI. To prove this, we expect the mis-calibration for the last batch to be significantly lower for the participants that received XAI. As the XAI and no XAI group are independent these tests are done with the *Wilcoxon-Mann Whitney* test. The group size for the \checkmark **Tutorial**, \times **XAI** condition excluding participants that had no miscalibration after the first batch is **44**. The group size for the \checkmark **Tutorial**, \checkmark **XAI** condition is **43**, again excluding participants without initial miscalibration. The results from these tests are found in Table 5.6. The p-values found in the results are both well above the α of 0.0125. Therefore, these results are not significant. It is not possible to conclude whether explainable AI increases the effect of reduction of miscalibration. Therefore, there is **no support for hypothesis 3a**.

However, the reduction of miscalibration for participants that received XAI is somewhat lower than that of participants who had no XAI.

Table 5.6: Wilcoxon-Mann Whitney test results for **hypothesis 3a**. Comparison of miscalibration between participants that did receive XAI and participants that did not. Significant results are shown in bold

Dependent Variable	U	p	$M \pm SD$ (XAI)	$M \pm SD$ (No XAI)
Mis-calibration last batch of tasks	868.5	.4803	0.167 \pm 0.120	0.213 \pm 0.210
Difference in mis-calibration	862.0	.4601	-0.114 \pm 0.203	-0.062 \pm 0.236

COMPARISON TO NO TUTORIAL

While support is found for **hypothesis 3**, it does not ensure that this improvement is a result of the tutorial. In this section, we further explore the source of this reduced mis-calibration. To test this the tutorial group will

be compared with the non-tutorial group in a between-subject test, therefore the *Wilcoxon-Mann Whitney* test is being used. It is assumed that the first batch should result in similar outcomes, as the conditions are the same for the first batch for both with and without the tutorial, which makes the second batch with miscalibration rather interesting. In this case, groups are independent of one another (tutorial vs no tutorial). The group size of the participants that received a tutorial is **87**, and the group of participants that did not receive a tutorial is of size **86**, with participants excluded that had no miscalibration on their first batch. Both the miscalibration for the last batch and the difference in miscalibration between the first and last batch (the change) are looked at. From the values depicted in Table 5.7 no conclusions can be drawn, as the p-values are not significant. A small difference can be seen in the means between the with and without tutorial conditions, which shows that the participants with tutorial have a lower mis-calibration. However, this effect seems to be minimal and is not significant. More research needs to be done to draw conclusions. No support can be found that the tutorial is a factor in reduced miscalibration between the first and last batch of questions.

Table 5.7: Wilcoxon-Mann Whitney test results to further explore **hypothesis 3**. Comparison of miscalibration between participants that did receive a tutorial intervention and participants that did not. Significant results are shown in bold.

Dependent Variable	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (Tutorial)	<i>M</i> ± <i>SD</i> (No tutorial)
Mis-calibration last batch of tasks	3386.5	.252	0.190 ± 0.175	0.203 ± 0.150
Difference in mis-calibration	3644.0	.7615	-0.088 ± 0.220	-0.079 ± 0.202

5.4.4. HYPOTHESIS FOUR

Hypothesis 4 states the following: *Making users aware of their miscalibrated self-assessment will result in more appropriate reliance on AI systems.* To test this hypothesis, the participants must be made aware of their miscalibrated self-assessment, this is done with a tutorial as previously mentioned. The hypothesis is very similar to **hypothesis 3** where the groups were the same (before the tutorial and after the tutorial from participants that received a tutorial). Appropriate reliance has already been discussed for **hypothesis 2**, in short, to find proof for appropriate reliance RAIR and RSR are calculated. The higher RAIR and RSR are the more appropriate the reliance is. In this case, participants without initial miscalibration are excluded from the experiments. This results in **87** participants included in the hypothesis test. The hypothesis test is a within-subject experiment, thus the *Wilcoxon signed ranks* test is used. However, the tests done for the comparisons on XAI to no XAI and tutorial to no tutorial were done with the *Wilcoxon Mann Whitney* test as these are between-subject tests.

HYPOTHESIS TEST

The results found with the *Wilcoxon signed ranks* test are given in Table 5.8. Besides the values for RAIR and RSR, also the accuracy, agreement and switch fraction are given to gain more insight. For both RAIR and RSR there is no significant difference between the first batch and the last batch, as the p-values are all well above the Bonferonni correct α of 0.0125. The values even seem to go slightly down, contrary to our expectations. Furthermore, accuracy and agreement also go down both but are not significant. The switch fraction does go up but is also not significant. With all these facts we can state that **hypothesis 4 is not supported** by the data from the experiments.

Table 5.8: Wilcoxon signed ranks test results for **hypothesis 4**. Comparison between the first and last batch of tasks the participant performed. Significant results are shown in bold.

Dependent Variables	<i>T</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (First six)	<i>M</i> ± <i>SD</i> (Last six)
Accuracy	942.5	.1066	0.575 ± 0.223	0.527 ± 0.217
Agreement Fraction	1222.0	.9041	0.630 ± 0.222	0.625 ± 0.240
Switch Fraction	1251.5	.3591	0.378 ± 0.327	0.414 ± 0.337
RAIR	899.0	.5860	0.431 ± 0.405	0.400 ± 0.373
RSR	579.0	.3967	0.477 ± 0.494	0.425 ± 0.479

XAI

Hypothesis 4a states that the effects of **hypothesis 4** would be amplified. In this case thus that RAIR and RSR would increase for the ✓ **Tutorial**, ✓ **XAI** condition. Therefore we test whether RAIR and RSR are increased more for the last batch, after receiving the tutorial, for participants that received explanations.

Participants without mis-calibration in the first batch were excluded from the test. Taking this exclusion into consideration the group sizes are as follows: the \checkmark Tutorial, \times XAI group had 44 participants and the \checkmark Tutorial, \checkmark XAI group had 43 participants left. The test done in this case is the *Wilcoxon-Mann Whitney* test, as the groups are independent of one another. The results can be found in Table 5.9. There are no significant results found, the differences in the last batch of questions even seem to be smaller. The RAIR increases from XAI to no XAI, which is opposing the hypothesis and RSR does decrease. With these contradictory effects, it is not expected that increasing group size would show an effect. The conclusion can be drawn that there is **no support for hypothesis 4a**

Table 5.9: Wilcoxon-Mann Whitney test results for **hypothesis 4a**. Comparison between participants that did receive XAI and participants that did not. Significant results are shown in bold.

Participants Dependent Variables	First Batch of Questions				Last Batch of Questions			
	<i>U</i>	<i>p</i>	$M \pm SD(XAI)$	$M \pm SD(no XAI)$	<i>H</i>	<i>p</i>	$M \pm SD(XAI)$	$M \pm SD(No XAI)$
Accuracy	830.0	.3133	0.545 \pm 0.231	0.605 \pm 0.212	924.5	.8548	0.527 \pm 0.200	0.527 \pm 0.236
Agreement Fraction	851.0	.4101	0.610 \pm 0.244	0.651 \pm 0.199	932.0	.9069	0.617 \pm 0.251	0.632 \pm 0.232
Switch Fraction	896.0	.6696	0.369 \pm 0.340	0.387 \pm 0.316	883.0	.5920	0.391 \pm 0.333	0.436 \pm 0.343
RAIR	904.5	.7179	0.419 \pm 0.412	0.444 \pm 0.402	915.5	.7931	0.386 \pm 0.365	0.415 \pm 0.385
RSR	788.5	.1296	0.398 \pm 0.492	0.558 \pm 0.470	1002.5	.5909	0.455 \pm 0.492	0.395 \pm 0.470

COMPARISON TO NO TUTORIAL

There is no support found for **hypothesis 4**. However, it is still of interest to take a look at how participants that had the tutorial to participants that had no tutorial compared to one another. It is assumed that the first batch of questions will give the same distribution, as the tutorial group, that did not yet receive a tutorial. In this case, the groups compared are again independent of one another, thus the *Wilcoxon-Mann Whitney* test is being used. The participants without mis-calibration for the first set of questions are again excluded from these tests. The group size of the with the tutorial group is 87 the group without the tutorial has a size of 86 participants. While higher values for the RAIR and RSR were expected for the participants that received a tutorial, the value is only higher for the RSR. All these values are not significant, therefore there is no support that the tutorial increases RAIR and RSR compared to participants without a tutorial stage.

Table 5.10: Wilcoxon-Mann Whitney test results to further explore **hypothesis 4**. Comparison between participants that did receive a tutorial intervention and participants that did not. Significant results are shown in bold.

Dependent Variables	<i>U</i>	<i>p</i>	$M \pm SD(Tutorial)$	$M \pm SD(No Tutorial)$
Accuracy	3370.0	.2473	0.527 \pm 0.217	0.568 \pm 0.199
Agreement Fraction	3302.0	.1741	0.625 \pm 0.240	0.678 \pm 0.228
Switch Fraction	3260.5	.1412	0.414 \pm 0.337	0.490 \pm 0.351
RAIR	3094.0	.0443	0.400 \pm 0.373	0.516 \pm 0.382
RSR	3847.0	.7143	0.421 \pm 0.476	0.401 \pm 0.484

5.4.5. SUMMARY

In Table 5.11 a short overview of the results for all hypotheses and sub-hypotheses can be found. This overview shows that an increased effect when participants received XAI has not been found for any of the hypotheses. Hypotheses one and three were supported with significant results. Hypothesis two had two elements that contributed to the hypothesis RAIR and RSR. Only RAIR was significantly higher for participants with an accurate self-estimation and RSR was surprisingly lower, but this effect was far from significant. Therefore, only partial support for hypothesis two is found. For hypothesis four no significant results were found.

Table 5.11: Results of all hypotheses

Hypotheses	Support found for hypothesis	Increased effect with XAI
One	Yes	No
Two	Partially (only RAIR)	No
Three	Yes	No
Four	No	No

5.5. EXPLORATORY RESULTS

In the hypotheses (section 5.4) some exploratory results on the tutorial were found. This section describes another exploratory finding on both XAI and the Dunning-Kruger Effect (DKE).

5.5.1. XAI

The influence of XAI is tested in the context of some of the hypotheses. However, this test is tested in solitary to see whether only supplying participants with XAI could increase accuracy or (appropriate) reliance on the AI system. The test is only done on the first batch of questions to make sure an interaction factor between the tutorial and XAI does not influence the outcome. As this is a between-subject test the *Wilcoxon-Mann Whitney* test is used. The results can be found in Table 5.12. None of the results is significant. However, all variables seem to be slightly higher for the no XAI conditions. This would suggest that XAI is counterproductive in helping users understand the AI system better. Nonetheless, no conclusions can be drawn due to the insignificance of the results. Further research would be needed to confirm any results.

Table 5.12: Wilcoxon-Mann Whitney test. Comparison of participants that received XAI vs participants that did not receive XAI. Significant results are in bold.

Dependent Variables	<i>U</i>	<i>p</i>	$M \pm SD(XAI)$	$M \pm SD(No XAI)$
Accuracy	7016.0	.1823	0.552 ± 0.213	0.596 ± 0.195
Agreement Fraction	7005.0	.1794	0.633 ± 0.227	0.675 ± 0.212
Switch Fraction	6681.0	.0579	0.400 ± 0.320	0.471 ± 0.320
RAIR	6873.0	.1126	0.435 ± 0.391	0.515 ± 0.392
RSR	7218.0	.2910	0.399 ± 0.473	0.464 ± 0.484

5.5.2. THE DKE

To find out more about the DKE, we took a look at two groups, the top and bottom performers. The aim is to split the participants in quarterlies in line with previous research [10], the bottom performers being the lowest 25%, and the top performers the highest 25%. The performance over only the first batch of questions was taken into consideration. To make sure that participants with the same accuracy did not end up in different groups, the participants with up to two questions right were put into the low accuracy group, which resulted in the bottom 20.1%. The top consisted of the participants with five or six questions correct this was the top 18.9%.

In the survey participants were asked how many questions they thought they had answered correctly and also how they would rate themselves among others on a scale of 0% to 100%. In Figure 5.3a the percentages are shown. This graph shows three groups, the low accuracy group, the high accuracy group and the all group, which includes all participants (so also the low and high accuracy groups). In this graph we can see that even though the performances highly differ, all the groups of participants rate themselves between 53% and 60%. So while the lower accuracy group does put themselves lower compared to other groups. This difference is minimal to the actual difference. The graph shows how the low performers highly overestimate themselves compared to other participants, this indicates that the group suffers from the DKE and that the DKE is present in human-AI team logic tasks. To further establish the DKE, the overestimation on the amount of correct answers is also checked for, the results can be seen in Figure 5.3b. The high performance group consisted of 47 (18.9%) participants, while the low performing group consists of 50 participants (20%) and the average performance group consists of 152 (61%) of participants. Therefore, direct comparisons between the groups of bars are harder to make as they do not consist of the same amount of participants. However, we can draw some conclusions from the ratios. The average group has more correct and underestimating participants than it has overestimating participants. The low accuracy group has 40 participants that overestimated themselves and only 10 that had no overestimation, which results in 80.0% of the participants overestimating their performance in this group. However, in the high accuracy group only 3 participants overestimated their performance and 44 had no overestimation, this results in 6.4% of the participants overestimating. This difference aligns with the findings by Kruger and Dunning [9]. Therefore, we can use the low accuracy group as the group with DKE and the high accuracy group as the group without DKE. These findings validate our motivation to design a tutorial intervention to mitigate the DKE in human-AI teams when doing logical reasoning tasks.

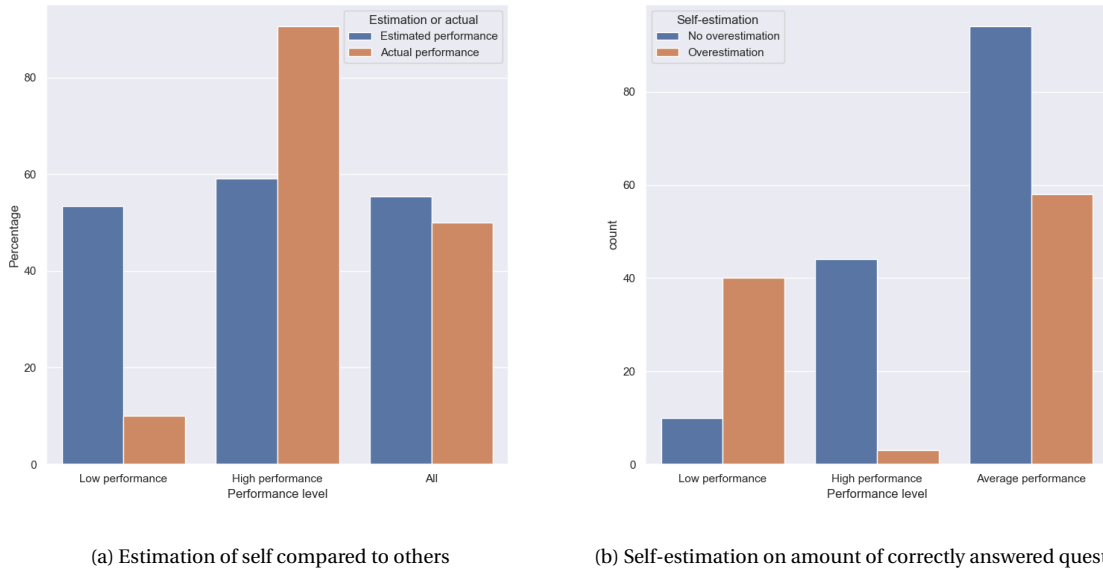


Figure 5.3: Estimation per group divided on performance. (a) Shows the participants divided on low, high and all (which includes low and high) on how they rated themselves on a scale of 0% to 100%, next to their actual average performance compared to others. (b) Shows the amount of participants that overestimated themselves per group, divided in low performers, high performers and average performers (the once that are not in either low or high are put in average). The overestimation is measured by comparing the amount of correct answers they estimated to have correct to the amount they actually had answered correctly.

THE IMPACT OF THE DKE ON (APPROPRIATE) RELIANCE

With the DKE effect established, we are interested in whether the DKE affects the (appropriate) reliance on AI. In Table 5.13 the measures between the low and high accuracy participants are calculated with a *Wilcoxon-Mann Whitney* test, as this is a comparison between-subject. These tests are done on the first batch of tasks. The participants with a lower accuracy achieve a significantly lower agreement fraction, RAIR and RSR. The switch fraction is also lower for participants with a low accuracy, although this is not significant. Therefore, participants in the lower accuracy group achieve a significantly poorer appropriate reliance and partially significantly lower reliance on the AI system. The less (appropriate) reliance on the AI system, reflects that the under-reliance is (partially) to blame for the lower performance.

Table 5.13: Wilcoxon-Mann Whitney test. Comparison of participants that had a high accuracy versus participants that had a low accuracy. Significant results are in bold.

Dependent Variables	<i>U</i>	<i>p</i>	$M \pm SD(\text{Low})$	$M \pm SD(\text{High})$
Agreement Fraction	221.0	<.0001	0.443 ± 0.160	0.727 ± 0.123
Switch Fraction	870.5	.0263	0.270 ± 0.187	0.286 ± 0.286
RAIR	638.0	<.0001	0.197 ± 0.183	0.631 ± 0.471
RSR	164.5	<.0001	0.120 ± 0.312	0.915 ± 0.217

5.5.3. ANALYSIS OF TRUST

Besides reliance, we are also interested in how the tutorial would affect trust. The trust was measured with a TiA-Trust questionnaire after the first and last batch of tasks. With this questionnaire being asked before and after we can test whether the trust changes after having received the tutorial intervention. This test is done with a *Wilcoxon signed ranks* test as it is a within-subject comparison. The results are shown in Table 5.14. The results suggest that the tutorial did not help on increasing participants' trust in the AI system. This suggests that the intervention helps to recalibrate their self-assessment without increasing their trust in the AI system.

Table 5.14: Wilcoxon signed ranks test results. Comparison between the TiA-trust after the first and last batch of tasks. The Likert scale is from 1-5. Significant results are shown in bold.

Dependent Variable	<i>T</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (First batch)	<i>M</i> ± <i>SD</i> (Last batch)
TiA-trust	1063.5	.9521	2.996 ± 0.880	3.02 ± 0.819

The trust does not seem to increase due to the tutorial intervention. However, it is still interesting to look at the difference in trust for the last batch comparing participants with a tutorial intervention to those that had no tutorial intervention. As this is a between-subject test the *Wilcoxon-Mann Whitney* test is used. The results can be found in Table 5.15. Interestingly it seems that participants that did not receive a tutorial intervention have a significantly higher trust in the AI system after the last batch of questions. A reason for this difference could be that the tutorial intervention reduced trust due to the process being too tedious.

Table 5.15: Wilcoxon-Mann Whitney test results. Comparison between participants that did receive a tutorial intervention and participants that did not. Comparison on trust on a Likert scale from 1-5. Significant results are shown in bold.

Dependent Variables	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (Tutorial)	<i>M</i> ± <i>SD</i> (No Tutorial)
TiA-trust	6301.0	.0085	3.02 ± 0.819	3.324 ± 0.799

6

DISCUSSION

In this chapter, the findings of this thesis will be discussed and placed into context. The first section states the key findings found in this thesis, specifically the answers to the research questions. In the second section, further findings are discussed that were mainly found in the exploratory results. These results are put into context in the implications section. After that, the limitations are discussed, which might have influenced the research and the results. In the future work, possibilities for future research are proposed.

6.1. KEY FINDINGS

The thesis began with two research questions. Research question 1 being: **How does the Dunning-Kruger Effect shape reliance on AI systems?** With the results of the experiments it was found that as expected people that overestimate themselves have less reliance and also less appropriate reliance on AI. This indicates that the Dunning-Kruger effect (DKE) results in participants being less capable to evaluate the AI system its performance compared to their own. It was however less clear whether participants that had an inaccurate self-estimation also had less (appropriate) reliance, thus **hypothesis 2** is only partially supported. More specifically participants that overestimated themselves showed under-reliance on the AI system. To answer the research question, the DKE results in less (appropriate) reliance on AI systems.

With the first research question in mind research question 2 was formulated: **How can the Dunning-Kruger Effect be mitigated in human-AI decision making tasks?** In this research question the goal is to mitigate the effects found in research question 1. Two ways to mitigate this effect were tested in this thesis. First of all, a tutorial was used to mitigate the effect. The results show that the miscalibration of self-estimation was significantly reduced after doing the tutorial, indicating that the effect was mitigated. However, results also showed that it was unclear whether this improvement was due to the tutorial intervention or a learning effect, as there was no significant difference found between the groups with and without the tutorial. Furthermore, while the miscalibration of self-estimation was reduced, the (appropriate) reliance on AI was not. Although the tutorial reduces miscalibration, it does not have a significant effect on reducing the negative effects of this miscalibration (e.g. the less (appropriate) reliance on the AI system).

Secondly, the effects of explainable AI (XAI) were looked at to see whether the conditions with XAI resulted in more extreme results. For none of the ways, the XAI has been tested increased effects in the XAI group were found. Therefore, the participants that were in the group with a tutorial and XAI did not improve miscalibration more than the ones that did not receive XAI. Furthermore, the accuracy and appropriate reliance also showed no difference between the groups. This indicates that the XAI used in this thesis had little to no effect in mitigating the DKE. Furthermore, the participants in the XAI conditions did not show a different reliance on AI systems due to the DKE, compared to those without XAI. Summarising, no successful intervention has been found to significantly increase the (appropriate) reliance on AI. The tutorial intervention showed a reduction of the miscalibration, however, it is unclear whether this is a result of the tutorial or a general learning effect.

6.2. FURTHER FINDINGS

In the key findings, the results of the research questions were discussed, in this section, the more exploratory results are briefly summarised and discussed. In this study, the human-AI teams performed better than humans on their own but were outperformed by AI. This is in accordance with previous research [4]. This means that in this research no complementary performance was found. This was also not expected as the aim of the study was to look at the relationship between people who overestimate themselves with AI advice, rather than optimising for complementary performance.

Furthermore, in this research, we looked at how helpful participants perceived the explainable AI to be. The results from the survey suggest that the explanations on AI were not perceived as very helpful. Furthermore, the results found for the hypotheses on XAI suggest that providing the participants with explanations on AI had little to no influence. Additionally in the exploratory results, we exposed the lack of change in results due to the XAI in general. Participants with XAI did not perform significantly better and did not have a higher (appropriate) reliance on the AI system. The participants with XAI even seemed to perform slightly worse and have a lower (appropriate) reliance on XAI systems, having the opposing effect of what is desired. This leaves us wondering whether the way the explanations were presented was the right approach. The highlighted text was aimed at providing spans that would be understandable for the participants and were not directly from the AI system. However, participants might still have had problems with interpreting the explanations on the AI in the form of just highlights.

Additionally, we looked at the DKE by dividing the participants and comparing the low to the high performers. With the findings, we could establish that our experiment does involve the DKE, thus in human-AI decision making working on logical reasoning tasks we can establish the DKE. We also found that the DKE affects the (appropriate) reliance, being lower for participants that have lower accuracy and thus suffer from the DKE. This shows that the under-reliance is (partially) to blame for the lower performance and therefore gives motivation to mitigate the DKE to increase performance and (appropriate) reliance.

Lastly, we found that the TiA-trust is not increased by a tutorial intervention. Participants do not show increased trust after receiving a tutorial intervention. However, we did find that participants that did not receive a tutorial intervention did show a significantly higher trust in the AI system. This suggests that the tutorial intervention might even work counteracting. This can be due to the tutorial intervention being perceived as tedious. While the trust after the last batch is higher for the participants who did not get a tutorial, the trust in participants who did get a tutorial intervention did increase non-significantly.

6.3. IMPLICATIONS

As discussed in the key findings people who overestimate themselves have less (appropriate) reliance on the AI system and this effect can be reduced with a tutorial intervention, but what value do these findings hold?

The fact that participants who overestimate themselves have a lower (appropriate) reliance on AI indicates that this group does not use the AI advice to its full potential, as this group also has a lower accuracy. Therefore, the effect of participants that overestimate themselves should be taken into consideration when researching human-AI decision making. We can also state that the DKE affects the (appropriate) reliance on AI, as the participants with overestimation also had lower performance in general. This aligns with the original findings by Kruger and Dunning [9]. By showing that this occurs in human-AI teams with logical reasoning tasks we build on the paper by Schaffer *et al.* [11] who already showed that the DKE can affect human-AI teams. In this study, we suggested a way to decrease the effect by creating a tutorial to re-calibrate the participant's self-estimation. The re-calibrating of the self-estimation was successful, which is in line with the original paper on the DKE [9]. However, the effects of the tutorial on the performance are still ambiguous. This is in accordance with the research done by Lu and Yin [15], who also had trouble improving performance by providing tutorials. Furthermore, the (appropriate) reliance was not improved by the tutorial intervention, which indicates that even though the participants gained better self-awareness the tutorial currently fails to give the participants a better perception of the AI system. Additionally, the tutorial also failed to increase trust of the participants in the AI system. This is still an interesting take as the participants reduced their miscalibration, thus decreasing the DKE, but still failed to improve themselves in other parts. The role a tutorial intervention can play is therefore still a bit ambiguous. Moreover, the effects were not increased with the XAI and participants also claimed they did not find the XAI particularly helpful. This suggests that our current take on XAI, might not have been the most helpful and other approaches should be looked at.

6.4. LIMITATIONS

The findings in this study are interesting, nonetheless, they come with a few limitations, which will be discussed in this section. First of all, it is not completely clear what and how biases have played a role in the experiments. Which is in accordance with Draws *et al.* [60] who emphasises the importance of acknowledging and preventing cognitive biases. To the best of our capabilities, we tried to reduce cognitive biases. Although we put in attention checks it is still possible for participants to only pay attention to whether a task is an attention check, and not put in the effort for the other tasks. The participants participate in the study for monetary compensation and most likely do not have a lot of motivation beyond that. They could therefore easily fall prey to the *self-interest bias*. Although the ReClor dataset was chosen to cater to laymen, some participants might be familiar with the subjects of some of the context texts. This could result in the participants having a *familiarity bias* and/or an *availability bias*, which could influence the results.

Secondly, in the study, the distributions of especially the Relative Positive Self Reliance (RSR) were off due to there only being two questions per batch in which the AI was wrong. To gain a better distribution of the RSR (and in less matter the RAIR) more data points would have been needed. With more data points there would have been more questions in which the AI was wrong. Therefore, RSR could only be 0, 0.5 or 1, with only a very limited amount of participants being in the 0.5 occasion. This can be explained by the fact that only instances where the initial answer of the participant was correct are taken into consideration. A participant answering both questions initially correctly has an even lower chance, this results in even fewer data points that can be used (for many thus zero or only one). Thus, the number of tasks in which the AI system gave wrong advice per batch is not sufficient to get a good distribution of RSR.

Furthermore, the tutorial needed to be created manually, which makes it subject to inconsistency. By being created manually the explanations can be biased by the writer or be unclear. While the same can be true for automatic feedback, it is still something to consider. The explanations were written in a contrastive way to best fit human thinking, as suggested by Miller [61]. However, it is hard to judge the quality of the feedback in an objective matter.

Another limitation is that some hypotheses could not be significantly proven but did show some differences. The results can be insignificant due to setting the impact factor too high. The insignificance of some of the results leads to the research questions still being partially unanswered. The results were nonetheless included to show small insignificant differences and results that probably can not be proven no matter the number of participants included in the study. There might be a significant difference in some more results if the number of participants is increased.

Lastly, it is unsure how these results will transfer to other tasks and domains. In this thesis the tasks used were logical reasoning questions, this was done to cater to layman-friendliness. However, as established earlier, most human-AI decision making is in domains with experts, such as in the medical and legal domains. The results in those domains could differ from the ones found here, as the level of expertise could play a further role. Another reason the results could be different is that it is unclear how many in those domains would still suffer from the DKE. The people in question are experts in their domain, so it would be expected that they are at least aware of what they are not aware of. As found by Mellinger [62] self-assessment does differ depending on the expertise in the medical domain. However, they could still suffer from the DKE as their lack of expertise with AI could be of hindrance and introduce a new potential risk of overestimation. This also translates to the tutorial intervention, this might need to be more focused on understanding the AI rather than concentrating on revealing the participants' weaknesses. Furthermore, we need to consider the fact that most participants were from the United Kingdom as shown in Figure 5.1a. Rachmatullah and Ha [59] suggest that confidence can depend on cultural differences, therefore we need to take into consideration that the results found might not transfer directly to participants with a different distribution of nationalities of participants. The results of this thesis still hold value in showing the effects of the DKE in human-AI teams but give no definitive conclusions about how these effects transfer to domains in which we deal with experts.

6.5. FUTURE WORK

The previous section discussed some limitations and the current implications, from those discussions some ideas have been put together to give some recommendations for potential future research.

In this research, the focus was on the implicit reliance of participants, by first answering the question on their own and only later on receiving AI advice. As done in the research by Liu *et al.* [3], it might be interesting to see how people explicitly perceive the AI system, does a tutorial make them believe they trust AI more? This will further help in understanding how the DKE effect shapes reliance on AI. For the XAI the participants

were asked about helpfulness, which aligned with the results. Both showed that the participants were not influenced a lot by the XAI. For the tutorial, the participants were not asked what their perception of it was, which could be interesting when a similar experiment is executed.

A drawback found in the limitations is the distribution of the RSR being very sided to either 0, 1 or occasionally 0.5. For future research, it might be helpful to include more tasks per batch when researching appropriate reliance. The additional tasks per batch are needed to include more tasks in which the AI its advice is faulty. That is necessary, as only in the cases that the AI advice is faulty participants can have positive self-reliance.

To truly get the best use out of the tutorial intervention, more care should be put into creating a human-centred approach. In this thesis, it is ambiguous whether the improvement in calibration was a result of the tutorial intervention. This leaves us wondering what effect the tutorial actually had. Previous research suggests tutorial interventions are a good way to increase human ability by increasing performance [15]. In this research, we were not able to significantly increase the performance. As previously mentioned, human-AI team performance has previously been successfully improved, it would therefore be interesting to see what would happen with the (appropriate) reliance on AI in those cases. Future research can explore different ways of increasing human accuracy, and or understanding of the AI system, and research the influence that has on the (appropriate) reliance on AI.

Furthermore, the XAI in this study seemed to not be very effective overall, not in perceived helpfulness nor increasing accuracy, nor the anticipated increased effect on the hypotheses. The explanations were given in the form of text spans. It could be possible that this type of explanation is hard to interpret. Other ways of explainable AI and its effects on (appropriate) reliance, specifically for the group suffering from the DKE, might still be interesting. However, it would be suggested to explore different methods of XAI, ones that are more interactive and give a better understanding of the AI.

Another suggestion we would make is to include more participants in the study. As some results stay rather ambiguous, one way to increase the likelihood of proving the hypotheses would be by adding more participants. However, adding more participants would make the study rather large, which would make it expensive to carry out. One way to again decrease participant amount is by only researching a set of conditions in solitude rather than all four at the same time. For example, only collecting data about tutorial versus no tutorial and XAI versus no XAI, separately. When interesting results are found, it would nonetheless be interesting to research the four conditions in one experiment again to see the interaction. At this stage it might be better to research the tutorial and XAI interventions separately, to increase the participant count per condition.

Lastly, it would be very interesting to see similar studies being executed in different domains. As discussed previously, it is hard to estimate how well the results would transfer to another domain. The only real way to prove this is by doing more experiments in a different domain. It would be particularly interesting to see how domain experts would react as their trust relationship to AI could be very different to laymen. Specifically, domain experts should exhibit less of the DKE (as they have more knowledge) but still might show DKE due to not being sufficiently aware of how to interact with AI advice.

7

CONCLUSIONS

In this thesis, a quantitative study on the effects of the Dunning-Kruger Effect (DKE) on human-AI decision making is presented. The aim is to uncover how the DKE shapes reliance on AI systems, in research question 1: **How does the Dunning-Kruger Effect shape reliance on AI systems?**. It is found that participants who overestimate themselves have a lower reliance and lower appropriate reliance on AI systems. Based on that finding, we can tell that the DKE can result in lower (appropriate) reliance on AI systems. We try to find a way to mitigate the DKE in human-AI decision making tasks in research question 2: **How can the Dunning-Kruger Effect be mitigated in human-AI decision making tasks?**. We found that a better calibration of self-estimation is achieved, which suggests that the DKE is mitigated. However, the (appropriate) reliance is not improved after the tutorial intervention. So while the goal of decreasing the DKE is accomplished, decreasing the negative effects due to the DKE on the human-AI relationship is not achieved. These results give a better understanding of how the DKE affects human-AI decision making, which means that it might be beneficial to give attention to participants that overestimate themselves more as they are more likely to show a lower (appropriate) reliance on AI. Previous research showed that the DKE is present in human-AI decision making for a prisoner's dilemma like game [11], and influences the trust participants have on the AI system. The research in this thesis shows that the DKE influences (appropriate) reliance in another domain (e.g. logical reasoning questions), which gives an indication that it might influence even more domains within human-AI decision making.

While the tutorial intervention reduced the miscalibration, it is ambiguous whether this is due to a learning effect or due to the tutorial intervention. This leaves room for future research, to build upon and look more closely at how this tutorial intervention can be improved by focusing on improving the (appropriate) reliance on AI systems. Another aspect to consider is the application of explainable AI (XAI) in this thesis. The current implementation did not result in the expected increased effects of the hypotheses. Furthermore, it has been found that XAI did not improve (appropriate) reliance, so besides not increasing the effects of the hypothesis it also did not prove to be helpful by itself. This gives reason to take a further look at different methods of applying XAI and how those influence the human-AI trust relationship. The takeaway of this research is that the DKE does influence (appropriate) reliance in human-AI decision making. Mitigating the effects that are a result of this DKE (e.g. lower (appropriate) reliance on AI) is not as easy to accomplish as initially expected, and might need a more thorough intervention method.

We want to thank all participants for their participation. The data from all participants has been anonymised to our best capabilities and should not allow tracing back to specific participants.

A

INTERFACE

Screenshots of the complete interface can be found in this appendix. Examples of the interface have been given throughout the thesis in Figures 3.1, 3.2, 3.3 and 3.4. The main colour used in the interface is blue, blue gives a feeling of trust and neutrality [63, 64], this seemed the most applicable to what the study should represent. We specifically refrained from green and red as those could lead participants to believe something was wrong or correct based on colour coding. The continue button was green, as it is okay for that button to be perceived as "good", neutrality is not needed for this part. Furthermore, the continue button had to be clear from the rest and thus has a different colour. The tasks needed to fit tightly on one screen such that the context, questions and answers could be read easily without scrolling from one to another, as all would be needed for answering questions. After selecting an answer it became inverted colours such that the participant was aware of what they selected.

The interface was built in python using Flask¹. Flask allowed us to build an interface with python, HTML, CSS and javascript. To follow people their trajectory through the experiment, a login manager was used ², which allowed us to track participants based on their prolific id (that was also shared by prolific). By tracking in this way we were always sure that the right information was stored with the right user at all times, also if they re-entered the experiment, for instance after a break. Furthermore, WTForms ³ was used to receive the input given by the users and store it in a database in a correct manner.

¹<https://flask.palletsprojects.com/en/2.2.x/>

²<https://flask-login.readthedocs.io/en/latest/>

³<https://wtforms.readthedocs.io/en/3.0.x/>



Informed Consent

Welcome to this study! In this study you will be asked to answer 16 multiple choice questions related to logical reasoning. On answering the question, the same question will be shown again, but with an answer predicted and advised by an AI system. The accuracy of the AI is 63%. With this prediction you can reconsider your answer, and answer the question again. For each question you will be provided with some context. You should answer the questions according to the given context. You can only select one answer, choose the one that (most) accurately answers the question. After selecting the answer you can click the button "Confirm and Continue" to go to the next question.

At the end of the study you will also be asked to fill out two short questionnaires.

You will first start with part 1 consisting of 6 questions, then a part 2 consisting of 4 questions and last part 3 consisting of 6 questions. By clicking "Confirm and Continue" on this page you confirm that you are participating in this study and know that your answers are recorded for research. Note that no information pertaining to your platform identity will ever be shared publicly.

Confirm and continue

Figure A.1: Introduction page



Questionnaire

Please answer the following questions honestly and to the best of your ability.

1 = completely disagree, 2 = largely disagree, 3 = slightly disagree,

4 = slightly agree, 5 = largely agree and 6 = strongly agree

1. I like to occupy myself in greater detail with technical systems

 1 2 3 4 5 6

completely largely slightly slightly largely completely
disagree disagree disagree agree agree agree

2. I like testing the functions of new technical systems

 1 2 3 4 5 6

completely largely slightly slightly largely completely
disagree disagree disagree agree agree agree

3. I predominantly deal with technical systems because I have to.

 1 2 3 4 5 6

completely largely slightly slightly largely completely
disagree disagree disagree agree agree agree

Figure A.2: ATI first out of three

4. How many multiple choice logical reasoning questions does this study have as indicated in the instructions..

1 2 3 16 80 240

5. When I have a new technical system in front of me, I try it out intensively.

1 2 3 4 5 6

completely largely slightly slightly largely completely
disagree disagree disagree agree agree agree

6. I enjoy spending time becoming acquainted with a new technical system.

1 2 3 4 5 6

completely largely slightly slightly largely completely
disagree disagree disagree agree agree agree

7. It is enough for me that a technical system works; I don't care how or why.

1 2 3 4 5 6

completely largely slightly slightly largely completely
disagree disagree disagree agree agree agree

Figure A.3: ATI second out of three, includes attention check question

8. I try to understand how a technical system exactly works.

1 2 3 4 5 6

completely largely slightly slightly largely completely
disagree disagree disagree agree agree agree

9. It is enough for me to know the basic functions of a technical system.

1 2 3 4 5 6

completely largely slightly slightly largely completely
disagree disagree disagree agree agree agree

10. I try to make full use of the capabilities of a technical system.

1 2 3 4 5 6

completely largely slightly slightly largely completely
disagree disagree disagree agree agree agree

Confirm and continue

Figure A.4: ATI third out of three



Tasks

Context

One way to compare chess-playing programs is to compare how they perform with fixed time limits per move. Given any two computers with which a chess-playing program is compatible, and given fixed time limits per move, such a program will have a better chance of winning on the faster computer. This is simply because the program will be able to examine more possible moves in the time allotted per move.

Task 1/16

Which one of the following is most strongly supported by the information above?

- A** If a chess-playing program is run on two different computers and is allotted more time to examine possible moves when running on the slow computer than when running on the fast computer, it will have an equal chance of winning on either computer.
- B** How fast a given computer is has no effect on which chess-playing computer programs can run on that computer.
- C** In general, the more moves a given chess-playing program is able to examine under given time constraints per move, the better the chances that program will win.
- D** If one chess-playing program can examine more possible moves than a different chess-playing program run on the same computer under the same time constraints per move, the former program will have a better chance of winning than the latter.

Confirm and continue

Figure A.7: Task



Tasks

Context

Essayist: Only happiness is intrinsically valuable; other things are valuable only insofar as they contribute to happiness. Some philosophers argue that the fact that we do not approve of a bad person's being happy shows that we value happiness only when it is deserved. This supposedly shows that we find something besides happiness to be intrinsically valuable. But the happiness people deserve is determined by the amount of happiness they bring to others. Therefore, _____.

Task 2/16

Which one of the following most logically completes the final sentence of the essayist's argument?

- A** The judgment that a person deserves to be happy is itself to be understood in terms of happiness.
- B** People do not actually value happiness as much as they think they do.
- C** The notion that people can be deserving of happiness is ultimately incoherent.
- D** The only way to be assured of happiness is to bring happiness to those who have done something to deserve it.

AI advice

A

Your choice

B

Confirm and continue

Figure A.8: Task with XAI advice



Tasks

Context

Lyme disease is caused by a bacterium transmitted to humans by deer ticks. Generally deer ticks pick up the bacterium while in the larval stage from feeding on infected white-footed mice. However, certain other species on which the larvae feed do not harbor the bacterium. Therefore, if the population of these other species were increased, the number of ticks acquiring the bacterium and hence the number of people contracting Lyme disease would likely decline

Task

Which of the following, select answer C to confirm you have read the question?

- | | |
|---|---|
| A Ticks do not suffer any adverse consequences from carrying the bacterium that causes Lyme disease in humans. | B A single host animal can be the source of bacteria for many tick larvae. |
| C A deer tick feeds only once while in the larval stage. | D There are no known cases of a human's contracting Lyme disease through contact with white-footed mice. |

Confirm and continue

Figure A.9: First attention check out of three



Tasks

Context

Weight loss programs that guarantee results mislead their customers. No program can ensure that someone who follows it will lose weight. These programs prey upon unhappy and insecure people who are often driven more by emotion than by reason. Moreover, many people who lose weight while on a program eventually regain the weight within a year. So while the programs' claims may be true for a short period, customers will be disappointed in the long run.",

Task

Confirm that you have read the question by selecting answer D

- | | |
|---|--|
| A The first is a generalization; the second is a consequence of that generalization. | B The first is an assertion that the author questions; the second is evidence in support of that assertion. |
| C The first is the author's position; the second is a consideration in support of that position. | D The first is evidence in support of the author's main point; the second is the author's main point. |

Confirm and continue

Figure A.10: Second attention check out of three



Tasks

Context

The early universe contained only the lightest elements, hydrogen and helium. Confirm that you have read the context by selecting answer B. A recently discovered gas cloud contained carbon several billion years ago, when the universe was no more than two billion years old.

Task

If the statements above are true, which of the following must, on the basis of them, also be true?

- A** The carbon in the gas cloud later formed part of some stars.
- B** No stars identified to date are as old as the gas cloud
- C** The gas cloud also contained hydrogen and helium.
- D** Some stars were formed before the universe was two billion years old.

Confirm and continue

Figure A.11: Third attention check out of three



Survey

From the previous 6 questions, how many questions do you estimate to have answered correctly? (after receiving AI advice)

- 0 1 2 3 4 5 6

Confirm and continue

Figure A.12: Survey question on self-assessment



Survey

From the previous 6 questions, how many questions on **average** do you estimate **your peers** to have answered correctly? (after receiving AI advice)

- 0 1 2 3 4 5 6

Confirm and continue

Figure A.13: Survey question on assessment of peers

i

Survey

How do you estimate your own performance (after AI advice) compared to your peers from 0-100%. 0% being the worst, 50% exactly average and 100% the best one

35

Confirm and continue

Figure A.14: Survey question on how participant would place themselves amongst peers based on performance

i

Tutorial

The next 4 questions are a tutorial, the questions work the same way as the questions done before with one addition. After answering the question, you will be shown whether your final answer (after getting AI advice) was correct, if the answer is incorrect you will be shown an explanation on why another answer is better than you answer.

Confirm and continue

Figure A.15: Tutorial instructions

i

Tasks

Context

A certain cultivated herb is one of a group of closely related plants that thrive in soil with high concentrations of metals that are toxic to most other plants. Agronomists studying the herb have discovered that it produces large amounts of histidine, an amino acid that, in test-tube solutions, renders these metals chemically inert. Possibly, therefore, the herb's high histidine production is what allows it to grow in metal-rich soils, a hypothesis that would gain support if __.

Correct answer
D

Task 7/16

Which of the following most logically completes the argument?

A The concentration of histidine in the growing herb declines as the plant approaches maturity. **B** Cultivation of the herb in soil with high concentrations of the metals will, over an extended period, make the soil suitable for plants to which the metals are toxic.

C Histidine is found in all parts of the plant—roots, stem, leaves, and flowers. **D** Others of the closely related group of plants are also found to produce histidine in large quantities.

AI advice
C

Your choice
C

The answer is D rather than C. Answer C claims that histidine is found in the whole plant. However, that would not give support to the hypothesis, as the histidine could also only be in the roots, while still having the effect. Answer D focuses on that the other plants that can live in metal-rich soil also have large quantities of histidine, which would suggest that the fact that they all can grow in metal heavy soil, could be a result of their common factor of producing a lot of histidines. And if the others did not produce histidine it would be directly clear that histidine is not the reason the herb can grow in metal-rich soil.

Confirm and continue

Figure A.16: Tutorial



Final Tasks

You will now enter the last set of tasks. Consisting of 6 more tasks.

Confirm and Continue

Figure A.17: Entering last batch of questions



Questionnaire

Please answer the following question honestly and to the best of your ability.

1 = Not at all helpful, 2 = Slightly helpful, 3 = Rather helpful,

4 = Helpful, 5 = Very Helpful

1. To what extent was the explanation (i.e., the highlighted words/phrases) helpful in making your final decision?



Not at all helpful slightly helpful rather helpful helpful very helpful

Confirm and continue

Figure A.18: Question on helpfulness

Confirm and continue

Figure A.19: Disabled continue button

B

RESULTS

This appendix contains more distributions from the results found. These distributions were used to determine whether the distributions were normally distributed or not. For all distributions, we found they were not normally distributed.

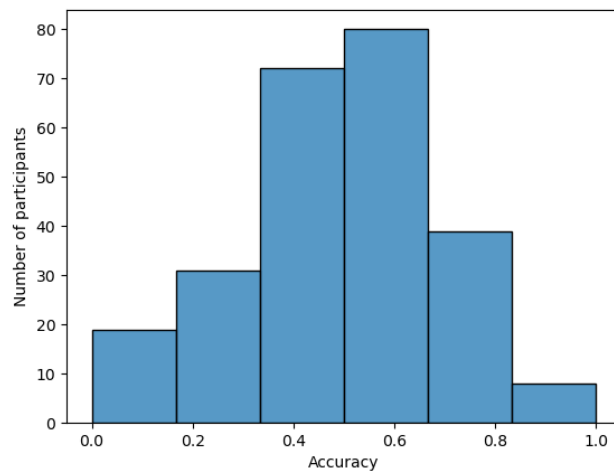


Figure B.1: Accuracy distribution of participants

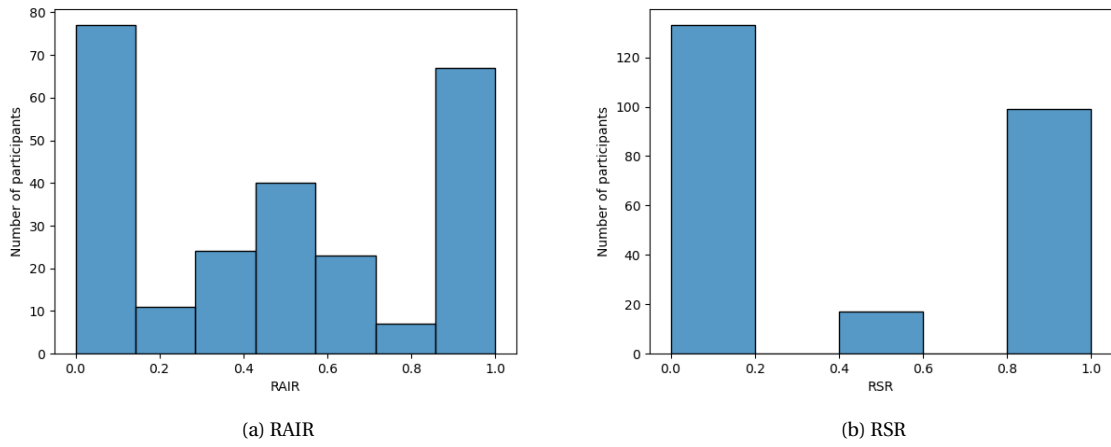


Figure B.2: RAIR and RSR distribution of participants

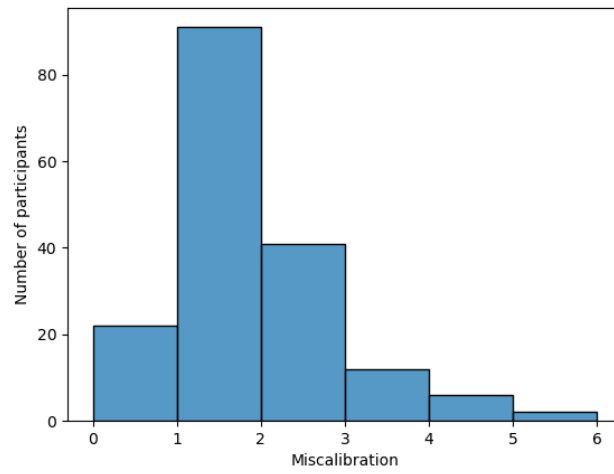


Figure B.3: Miscalibration distribution of participants

BIBLIOGRAPHY

- [1] O. Nov, Y. Aphinyanaphongs, Y. W. Lui, D. Mann, M. Porfiri, M. Riedl, J.-R. Rizzo, and B. Wiesenfeld, *The transformation of patient-clinician relationships with ai-based medical advice*, *Commun. ACM* **64**, 46–48 (2021).
- [2] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld, *Does the whole exceed its parts? the effect of ai explanations on complementary team performance*, (2020).
- [3] H. Liu, V. Lai, and C. Tan, *Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making*, *Proc. ACM Hum. Comput. Interact.* **5**, 1 (2021).
- [4] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, *Beyond accuracy: The role of mental models in human-ai team performance*, in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7 (2019) pp. 2–11.
- [5] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, *Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems*, *CoRR abs/2001.08298* (2020), 2001.08298 .
- [6] M. Schemmer, P. Hemmer, N. Kühl, C. Benz, and G. 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 (trAIIt)* (2022).
- [7] V. Lai, H. Liu, and C. Tan, *"why is 'chicago' deceptive?" towards building model-driven tutorials for humans*, in *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, edited by R. Bernhaupt, F. F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avelino, A. Goguy, P. Bjøn, S. Zhao, B. P. Samson, and R. Kocielnik (ACM, 2020) pp. 1–13.
- [8] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, and V. Gogate, *Anchoring bias affects mental model formation and user reliance in explainable ai systems*, in *26th International Conference on Intelligent User Interfaces, IUI '21* (Association for Computing Machinery, New York, NY, USA, 2021) p. 340–350.
- [9] J. Kruger and D. 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**, 1121 (1999).
- [10] U. Gadiraju, B. Fetahu, R. Kawase, P. Siehndel, and S. Dietze, *Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks*, *ACM Transactions on Computer-Human Interaction (TOCHI)* **24**, 1 (2017).
- [11] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer, *I can do better than your ai: Expertise and explanations*, in *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19* (Association for Computing Machinery, New York, NY, USA, 2019) p. 240–251.
- [12] S. Lee, M. Lee, and S. Lee, *What if artificial intelligence become completely ambient in our daily lives? exploring future human-ai interaction through high fidelity illustrations*, *International Journal of Human-Computer Interaction* **0**, 1 (2022), <https://doi.org/10.1080/10447318.2022.2080155> .
- [13] R. Fogliato, A. Chouldechova, and Z. Lipton, *The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies*, *Proceedings of the ACM on Human-Computer Interaction* **5**, 1 (2021).
- [14] B. Green and Y. Chen, *Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts*, *arXiv preprint arXiv:2012.05370* (2020).

- [15] Z. Lu and M. Yin, *Human reliance on machine learning models when performance feedback is limited: Heuristics and risks*, in *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, edited by Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. M. Drucker (ACM, 2021) pp. 78:1–78:16.
- [16] L. Chong, G. Zhang, K. Goucher-Lambert, K. Kotovsky, and J. Cagan, *Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice*, *Computers in Human Behavior* **127**, 107018 (2022).
- [17] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, *Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making*, in *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, edited by M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggeri, L. Taylor, and G. Zanfir-Fortuna (ACM, 2020) pp. 295–305.
- [18] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, *Deep learning for identifying metastatic breast cancer*, (2016).
- [19] S. E. Dilsizian and E. 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 (2014).
- [20] A. E. Khandani, A. J. Kim, and A. W. Lo, *Consumer credit-risk models via machine-learning algorithms*, *Journal of Banking & Finance* **34**, 2767 (2010).
- [21] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, *Towards a science of human-ai decision making: a survey of empirical studies*, arXiv preprint arXiv:2112.11471 (2021).
- [22] J. D. Lee and K. A. See, *Trust in automation: Designing for appropriate reliance*, *Human factors* **46**, 50 (2004).
- [23] C. Chiang and M. Yin, *Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models*, in *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, edited by G. Jacucci, S. Kaski, C. Conati, S. Stumpf, T. Ruotsalo, and K. Gajos (ACM, 2022) pp. 148–161.
- [24] C.-W. Chiang and M. Yin, *You'd better stop! understanding human reliance on machine learning models under covariate shift*, in *13th ACM Web Science Conference 2021* (2021) pp. 120–129.
- [25] X. Wang and M. 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* (2021) pp. 318–328.
- [26] D. Spiller, *Assessment matters: Self-assessment and peer assessment*, *The University of Waikato* **13**, 2 (2012).
- [27] H. Andrade and A. Valtcheva, *Promoting learning and achievement through self-assessment*, *Theory into practice* **48**, 12 (2009).
- [28] A. Papanthymou and M. Darra, *The contribution of learner self-assessment for improvement of learning and teaching process: A review*, *Journal of Education and Learning* **8**, 48 (2019).
- [29] R. A. Jansen, A. N. Rafferty, and T. L. Griffiths, *A rational model of the dunning-kruger effect supports insensitivity to evidence in low performers*, *Nature Human Behaviour* **5**, 756 (2021).
- [30] N. Yates, S. Gough, and V. Brazil, *Self-assessment: With all its limitations, why are we still measuring and teaching it? lessons from a scoping review*, *Medical Teacher* **44**, 1296 (2022), PMID: 35786121, <https://doi.org/10.1080/0142159X.2022.2093704>.
- [31] S. A. Nisly, J. Sebaaly, A. G. Fillius, W. R. Haltom, and M. M. Dinkins, *Changes in pharmacy students' metacognition through self-evaluation during advanced pharmacy practice experiences*, *American Journal of Pharmaceutical Education* **84** (2020), 10.5688/ajpe7489, <https://www.ajpe.org/content/84/1/7489.full.pdf>.

- [32] R. E. Mayer, A. T. Stull, J. Campbell, K. Almeroth, B. Bimber, D. Chun, and A. Knight, *Overestimation bias in self-reported sat scores*, *Educational Psychology Review* **19**, 443 (2007).
- [33] D. Dunning, C. Heath, and J. M. Suls, *Flawed self-assessment: Implications for health, education, and the workplace*, *Psychological science in the public interest* **5**, 69 (2004).
- [34] A. B. Ocay, *Investigating the dunning-kruger effect among students within the contexts of a narrative-centered game-based learning environment*, in *Proceedings of the 2019 2nd International Conference on Education Technology Management*, ICETM 2019 (Association for Computing Machinery, New York, NY, USA, 2020) p. 8–13.
- [35] A. Selbst and J. Powles, *"meaningful information" and the right to explanation*, in *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, *Proceedings of Machine Learning Research*, Vol. 81, edited by S. A. Friedler and C. Wilson (PMLR, 2018) p. 48.
- [36] A. Jacovi, J. Bastings, S. Gehrmann, Y. Goldberg, and K. Filippova, *Diagnosing AI explanation methods with folk concepts of behavior*, *CoRR abs/2201.11239* (2022), [2201.11239](https://arxiv.org/abs/2201.11239).
- [37] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, *Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems*, *CoRR abs/2001.08298* (2020), [2001.08298](https://arxiv.org/abs/2001.08298).
- [38] C. S. Bradley, K. T. Dreifuerst, B. K. Johnson, and A. Loomis, *More than a meme: The dunning-kruger effect as an opportunity for positive change in nursing education*, *Clinical Simulation in Nursing* **66**, 58 (2022).
- [39] W. Yu, Z. Jiang, Y. Dong, and J. Feng, *Reclor: A reading comprehension dataset requiring logical reasoning*, arXiv preprint arXiv:2002.04326 (2020).
- [40] J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang, *Logiqa: A challenge dataset for machine reading comprehension with logical reasoning*, *CoRR abs/2007.08124* (2020), [2007.08124](https://arxiv.org/abs/2007.08124).
- [41] R. Fogliato, S. Chappidi, M. Lungren, P. Fisher, D. Wilson, M. Fitzke, M. Parkinson, E. Horvitz, K. Inkpen, and B. Nushi, *Who goes first? influences of human-ai workflow on decision making in clinical imaging*, in *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22 (Association for Computing Machinery, New York, NY, USA, 2022) p. 1362–1374.
- [42] B. Green and Y. Chen, *Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments*, in *Proceedings of the conference on fairness, accountability, and transparency* (2019) pp. 90–99.
- [43] E. Peer, J. Vosgerau, and A. Acquisti, *Reputation as a sufficient condition for data quality on amazon mechanical turk*, *Behavior research methods* **46**, 1023 (2014).
- [44] T. Draws, N. Tintarev, U. Gadiraju, A. Bozzon, and B. Timmermans, *This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics*, in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 295–305.
- [45] S. Mohseni, J. E. Block, and E. Ragan, *Quantitative evaluation of machine learning explanations: A human-grounded benchmark*, in *26th International Conference on Intelligent User Interfaces*, IUI '21 (Association for Computing Machinery, New York, NY, USA, 2021) p. 22–31.
- [46] Y. Liel and L. Zalmanson, *What if an ai told you that $2+2$ is 5? conformity to algorithmic recommendations*. in *ICIS* (2020).
- [47] J. Sawler, *Economics 101-ism and the dunning-kruger effect: Reducing overconfidence among introductory macroeconomics students*, *International Review of Economics Education* **36**, 100208 (2021).
- [48] V. Lai and C. 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* (2019) pp. 29–38.
- [49] A. Sheth, M. Gaur, K. Roy, and K. Faldu, *Knowledge-intensive language understanding for explainable ai*, *IEEE Internet Computing* **25**, 19 (2021).

- [50] M. B. Zafar, P. Schmidt, M. Donini, C. Archambeau, F. Biessmann, S. R. Das, and K. Kenthapadi, *More than words: Towards better quality interpretations of text classifiers*, arXiv preprint arXiv:2112.12444 (2021).
- [51] H. Yan, L. Gui, and Y. He, *Hierarchical interpretation of neural text classification*, arXiv preprint arXiv:2202.09792 (2022).
- [52] F. Xu, J. Liu, Q. Lin, Y. Pan, and L. Zhang, *Logiformer: A two-branch graph transformer network for interpretable logical reasoning*, in *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, edited by E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai (ACM, 2022) pp. 1055–1065.
- [53] M. Yin, J. Wortman Vaughan, and H. 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* (2019) pp. 1–12.
- [54] M. Körber, *Theoretical considerations and development of a questionnaire to measure trust in automation*, in *Congress of the International Ergonomics Association* (Springer, 2018) pp. 13–30.
- [55] S. Tolmeijer, U. Gadiraju, R. Ghantasala, A. Gupta, and A. 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, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, edited by J. Masthoff, E. Herder, N. Tintarev, and M. Tkalcic (ACM, 2021) pp. 77–87.
- [56] T. Franke, C. Attig, and D. 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**, 456 (2019).
- [57] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, *Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses*, *Behavior research methods* **41**, 1149 (2009).
- [58] G. Pallier, *Gender differences in the self-assessment of accuracy on cognitive tasks*, *Sex roles* **48**, 265 (2003).
- [59] A. Rachmatullah and M. Ha, *Examining high-school students' overconfidence bias in biology exam: a focus on the effects of country and gender*, *International Journal of Science Education* **41**, 652 (2019), <https://doi.org/10.1080/09500693.2019.1578002>.
- [60] T. Draws, A. Rieger, O. Inel, U. Gadiraju, and N. Tintarev, *A checklist to combat cognitive biases in crowdsourcing*, in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9 (2021) pp. 48–59.
- [61] T. Miller, *Contrastive explanation: A structural-model approach*, *The Knowledge Engineering Review* **36** (2021).
- [62] C. D. Mellinger, *Metacognition and self-assessment in specialized translation education: task awareness and metacognitive bundling*, *Perspectives* **27**, 604 (2019), <https://doi.org/10.1080/0907676X.2019.1566390>.
- [63] P. Broeder, H. Snijder, et al., *Colour in online advertising: Going for trust, which blue is a must*, *Marketing-from Information to Decision Journal* **2**, 5 (2019).
- [64] M. Amsteus, S. Al-Shaabani, E. Wallin, and S. Sjöqvist, *Colors in marketing: A study of color associations and context (in) dependence*, *International Journal of Business and Social Science* **6** (2015).