Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making

Salimzadeh, S.

DOI
10.4233/uuid:8eddbb29-47d2-4344-8b93-aef440c1628a

Publication date
2024

Document Version
Final published version

Citation (APA)
Salimzadeh, S. (2024). *Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:8eddbb29-47d2-4344-8b93-aef440c1628a

Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making

Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus prof. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on Tuesday, 3rd of December 2024 at 12:30 o'clock

by

Sara Salimzadeh

Master of Science in Computer Science,
University of Amsterdam, the Netherlands,
Vrije Universiteit Amsterdam, the Netherlands,
born in Shemiran, Iran.

This dissertation has been approved by the promotors.

promotor: prof. dr. A. van Deursen
copromotor: dr. U.K. Gadiraju

Composition of the doctoral committee:

Rector Magnificus,	Chairperson
Prof. dr. A. van Deursen,	Delft University of Technology, promotor
Dr. U.K. Gadiraju,	Delft University of Technology, copromotor
Independent members:	
Prof. dr. A. Bernstein	University of Zurich, Switserland
Prof. dr. ir. G.J.P.M. Houben	Delft University of Technology, the Netherlands
Prof. dr. S. Stumpf	University of Glasgow, United Kingdom
Prof. dr. K. Verbert	KU Leuven, Belgium
Prof. dr. M.E. Warnier	Delft University of Technology, the Netherlands

This research was supported by the AI for Fintech Research, a collaboration between ING and Delft University of Technology.

SIKS Dissertation Series No. 2024. The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: Human-AI Decision-Making, Empirical Studies, Trust and Reliance on AI Systems, Human-AI Team, User Behaviour, Crowdsourcing, Contextual Factors

Printed by: Proefschriftspecialist

Cover: Maede Shakeri - In the age of AI, AI systems are increasingly being employed to assist humans in complex decision-making tasks, leveraging the strengths of both human and machine intelligence.

ISBN 978-94-93391-64-2

Contents

1

INTRODUCTION

We can't solve problems by using the same kind of thinking we used when we created them.

Albert Einstein

Decisions are not just a part of our lives; they are the essence of it. They are pivotal in our lives, influencing personal and professional spheres. Each day presents us with many choices that can affect our overall welfare and achievements. From small choices like our meals to major decisions such as choosing a career or making significant investments, these decisions not only affect us but also have broader implications for our families, communities, and society as a whole. The significance of decision-making in our everyday lives has prompted extensive efforts to understand the fundamental issues and components that influence decision-making across different fields, including psychology, management, economics, sociology, business, and more. These studies share a common goal of enhancing decision making processes and optimizing outcomes in diverse scenarios. This is crucial as making sub-optimal decisions can be costly, resulting in missed opportunities, financial losses, wasted resources, and negative impacts on individuals, organizations, and society [295]. The urgency to improve decision-making is also highlighted by abundant information, human cognitive biases, time constraints, and contextual factors that can lead to sub-optimal outcomes.

Technological advancements and the continuous development of AI systems present a growing potential to use these tools to improve decision-making processes and outcomes. These systems can analyze vast amounts of data, identify patterns and trends, provide real-time insights, and offer recommendations based on the analysis, all of which can empower humans in making more informed choices [258, 457]. While AI systems have shown tremendous potential in decision-making, the ultimate goal is not complete automation due to accountability and ethical and legal considerations [247, 255, 329]. Instead, the human-centered perspective is to foster a collaborative relationship between humans and AI systems, where AI systems can complement human capabilities and assist decision-makers in collecting and analyzing relevant data, generating different options, performing scenario analyses, and assessing the possible results of various decisions [8, 224, 225, 370]. Many research studies have explored incorporating AI systems into decision-making processes in different fields, promoting effective collaboration [13, 101, 109, 447]. For instance, explainable AI systems have been developed to provide transparency in the decision-making process, allowing humans to trust and validate the recommendations provided by AI systems [269, 419]. Different interfaces have also been designed to facilitate interaction between humans and AI systems, ensuring that decision-makers can easily understand the outputs of AI systems and incorporate them into their decision-making [49, 296].

When humans and AI work together on making informed decisions, various factors, such as individual characteristics [121, 313], the decision-making context [57, 408], the capabilities of AI systems [315, 321], and their interactions [22], play a crucial role in shaping the outcomes of these collaborative efforts. The success of these efforts hinges on understanding these factors and integrating them into the design and deployment of AI systems. Continuously adapting these AI systems to diverse situations can lead to ongoing changes in individuals' behavior as decision contexts become more complex and dynamic. These evolving AI systems have an impact on how decision-makers perceive, approach, and strategize while also influencing fundamental factors that determine decision outcomes [138, 209]. Therefore, there is a continuous need to refine the design and implementation of AI systems in decision-making environments while evaluating their effects on human behavior and decision outcomes.

Despite the high capabilities of AI systems in many areas, they are seldom perfect. Incorporating them into decision-making processes frequently results in reduced overall team performance when compared to either AI systems or humans working independently [162, 205]. This discrepancy highlights the need to understand current limitations and challenges in integrating AI systems into decision-making contexts and identify strategies to mitigate them. Current research predominantly centers on individuals and AI systems, overlooking the broader context in which decision-making occurs [247]. Thus, it is important to fully understand the impact of AI systems in different contexts in order to transfer findings across various domains and provide tailored solutions for specific decision-making scenarios. This thesis seeks to fill these gaps by investigating the dynamics of human-AI decision-making in different contextual settings, as well as gaining empirical insights into how human behavior might be influenced under different circumstances throughout Chapters 2 to 5. Chapters 6 and 7 present two approaches for gathering relevant information from online sources and databases to enhance decision-making.

1.1. General Overview of Human-AI Decision-Making

Decision-making is an integral part of human life that occurs in various contexts, from personal choices to professional settings. It involves selecting the best course of action from multiple alternatives based on available information and desired outcomes. Individuals have unique skills, traits, cognitive capacities, and experiences that shape their decision-making behavior and distinguish them from other individuals, Figure 1.1,

A. For example, some decision-makers may be more risk-averse, prioritizing certainty over potential gains, while others may exhibit a greater appetite for risk-taking. Human cognitive limitations also play a significant role in decision-making, as individuals often simplify complex problems, rely on heuristics [154, 372], and exhibit biases, particularly when they encounter extensive information, insufficient data to support their decisions, or have limited time for analysis [219]. While heuristics may not be based on logic, they align with ecological principles. They are neither inherently good nor bad, rational nor irrational [152]. Instead, they are relative to the context in which they are applied. For instance, confirmation bias could lead individuals to seek information supporting their pre-existing beliefs and ignore contradictory evidence, resulting in sub-optimal decisions [312]. On the other hand, heuristics can be beneficial when used in appropriate contexts, such as making quick decisions in familiar and low-stakes situations. Consequently, heuristics have the potential to either assist or impede decision-making, depending on the specific situation and desired outcome.

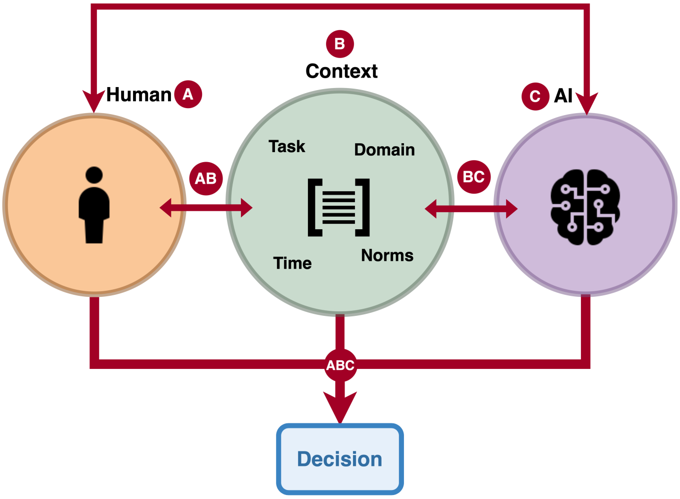


Figure 1.1: **General Overview of Human-AI Decision-Making**

A wide range of AI systems have been developed to assist people with various decision-making problems [156, 283, 320]. Several elements are essential for creating high-quality AI systems, including data accessibility, suitable model structures and algorithms, and efficient training methods, Figure 1.1, C. These AI systems can provide valuable insights by rapidly analyzing large volumes of data, recognizing patterns, and generating recommendations. The design and functionality of AI systems have also been determined by the context in which they are deployed. There are no AI systems that are fit for purpose in all situations, and thus, the choice of AI systems should be tailored to the specific requirements and characteristics of the contexts, Figure 1.1, BC. These contextual factors include the nature of the task, task attributes, the dynamic environment, time constraints, the domain of application, and the broader societal norms and regulations, Figure 1.1, B. For instance, in deterministic contexts, AI systems can be effective in automating certain tasks and making precise predictions based on available data. In these well-defined and predictable environments, AI systems can leverage their superior computational power and pattern recognition abilities to arrive at optimal decisions, while human involvement may be minimized. However, in stochastic and dynamic contexts characterized by high uncertainty, ambiguity, and rapidly changing conditions, AI systems may struggle to capture the full complexity of the situation and promptly adapt to changing circumstances.

Given the complementary strengths and limitations of humans and imperfect AI systems, integrating them into decision-making processes can leverage the unique capabilities of each to enhance overall performance, Figure 1.1, AC. In this collaborative decision-making paradigm, different attributes of AI systems, such as explainability, transparency, and the quality of outputs, can foster or hinder human-AI interaction, ultimately shaping the decision outcomes. While the explainability and transparency of AI decision-making processes are key factors for building trust and enabling human validation [12, 239, 315], the failures and errors of AI systems can also negatively impact the confidence and decision-making abilities of humans [221, 260]. Furthermore, the quality of AI-generated recommendations, including their accuracy, relevance, and alignment with human values and goals, can significantly influence the extent to which human decision-makers rely on and incorporate them. AI systems should also accommodate the diverse requirements and preferences of individuals with different characteristics in order to promote successful collaboration. For instance, many interventions, such as tutorials [78, 277], feedback mechanisms [277, 344], interactive interfaces [105, 291], and cognitive forcing functions [56], have been proposed to assist humans with different mental abilities, decision-making styles, and literacy in incorporating AI recommendations into their decision-making processes.

Individuals decision-making behavior and outcomes can also be influenced by the contextual factors surrounding the decision-making process, Figure 1.1, AB. Contextual factors such as domain knowledge, expertise, task-related features, and time constraints can determine the strategy individuals employ in making decisions. For example, experts might depend on their intuition to make choices when information is unconsciously processed [29, 107, 295, 379]. In contrast, non-experts may rely more on deliberate and analytical thinking processes, evaluating available information thoroughly before making a decision [379]. Prior studies indicate that while non-experts seek more information, experts rely more on their existing expertise, long-term memory, and knowledge [86, 231, 299]. It has also been recognized that individuals adhere to particular patterns when interacting with AI systems, which is referred to as the ASPECT model [209]: 1) attribute-based choices involve assigning values to different attributes [42, 196, 328]; 2) consequence-based decisions entail weighing potential outcomes before making a decision in uncertain circumstances [144, 220, 413]; 3) experience-based choices rely on past experiences and memories [41, 153, 436]; 4) socially-based decisions are based on others' opinions or exceptions [15, 85, 136]; 5) policy-based choices follow pre-determined rules or guidelines [284, 338, 343]; 6) trial and error-choices involve learning from mistakes over time [330, 339, 468]. These patterns are often used in combination, depending on the specific context and decision-making task at hand [217, 303]. Adapting AI systems to align with these patterns can significantly help decision-makers enhance their decision-making process and ultimately attain better outcomes [77, 390].

Establishing a good decision-making environment requires understanding the needs and preferences of individuals, as well as the contextual factors influencing the decision-making process and the attributes of the AI systems, Figure 1.1, **ABC**. Therefore, we should recognize and consider how the interaction between the decision-maker and the AI system, along with context-specific behavioral dynamics, can impact this process and, ultimately, decision-making outcomes. Despite extensive research in various aspects of the decision-making environment, there are still gaps in identifying contextual factors and their potential impacts on decision-making outcomes. Therefore, this thesis seeks to explore the role of contextual factors, particularly task-related elements, in the decision-making process and their influence on decision outcomes. These insights can contribute to the development of more effective AI systems and guide future research in human-AI decision-making field.

1.2. Scope of the Thesis

This thesis aims to uncover task-related contextual factors, Figure 1.1 **B**, and their influence on decision-making outcomes within the scope of human-AI decision-making. Tasks are generally defined as specific activities or problems that require decision-making processes to be carried out within a given context. Task-related elements include various task attributes such as the complexity of the decision task, the uncertain nature of the task, the availability and quality of information, the stake involved, and the time pressure or urgency associated with the task. By examining the interplay between task attributes **B**, decision outcomes **ABC**, individual and group behavior **A**, and their interactions with AI systems **AC**, this research contributes to a better understanding of how human-AI decision-making can be studied. This knowledge can guide the design of AI systems tailored to meet decision-makers specific requirements, thereby improving the decision-making process and its results. It is important to note the broader contextual factors, attributes of AI systems, or characteristics of decision-makers themselves are out of the scope of this thesis and will not be directly addressed.

This thesis focuses on non-expert decision-makers across different domains. These individuals often encounter challenges in making well-informed decisions due to their limited knowledge in a specific field. Biases, restricted access to information, and inadequate decision-support tools can further hinder their decision-making process. Even with abundant information, individuals may struggle to use it effectively, resulting in an exhausting decision-making process. This is known as the paradox of choice [365], which can detrimentally impact the quality of their decisions. For instance, investors with limited financial knowledge may struggle to make informed decisions in the stock market. They may rely on online forums for information, which can be biased and unreliable. Their emotions and recommendations from others can also sway them without considering their own goals and preferences. Inadequate analytical abilities and a lack of access to advanced decision-support tools can further contribute to sub-optimal decision-making outcomes. Moreover, while individuals can attain expertise in one area, they remain non-experts in others, emphasizing the necessity for decision-support systems that accommodate individuals with different proficiency levels.

The variety of decision-making contexts and how they have been operationalized in prior empirical research have led to a fragmented understanding of the effectiveness of AI systems across different domains. Therefore, the thesis first proposes a theoretical framework for systematically evaluating and comparing decision tasks, considering the complexity levels of the decision-making contexts. Using this framework, the thesis then reviews and analyzes existing literature to identify their strengths, limitations, and potential areas for improvement. It further develops these insights by conducting multiple empirical studies. These studies seek to evaluate how different task-related contextual factors influence individual behavior and performance outcomes. In researching the individuals' behaviors, we aim to comprehend the reasoning behind their choices and decisions by gathering data on their interactions with decision-support systems. This includes information access, time allocation for tasks, and decision-making patterns. Analyzing these factors along with findings further underscores the urgent need for tailored decision-support tools to improve decision-making outcomes for non-expert individuals in different domains. The thesis also proposes a modular framework for designing rigorous empirical research within real-world decision-making scenarios, promoting the generalizability and reproducibility of outcomes.

We then proceed to proposing methods for improving information access and individual's behaviors when interacting with information sources. Accessing relevant information is a crucial preparatory step before decision-making, especially for non-experts who often deal with a surplus of information, restricted access to relevant data, or time constraints. The presence of these factors may result in cognitive biases that can negatively impact the outcomes of decisions. Therefore, understanding how individuals seek, process, and utilize information is essential for creating practical tools to manage these biases and ultimately improve decision-making outcomes. To this end, this thesis proposes approaches to facilitate information access, such as providing relevant and customized recommendations and designing interfaces that facilitate efficient information retrieval. By conducting experiments and collecting data on individuals' behaviors, these proposed approaches are empirically evaluated and compared to traditional methods. The findings shed light on how the design and utilization of tailored tools can be further optimized to facilitate information access across various contexts.

1.3. RESEARCH QUESTIONS

There is a growing interest in using AI systems in various decision-making scenarios [128, 241, 283] to combine their complementary abilities with human decision-makers [381, 420]. Harnessing the full potential of AI systems requires consideration of various factors, including human-centric elements [121, 314] such as prior experience and cognitive biases, as well as attributes of AI systems [315, 321] like transparency and explainability. Furthermore, the context of decision-making also plays a crucial role in the effectiveness of AI systems [57, 408]. Previous research has predominantly centered on the initial two aspects, with comparatively fewer studies into the influence of task context on decision-making. This can greatly impact the transferability of study results across different scenarios. Moreover, it's unclear how decision-makers behavior and performance can be affected when using AI systems in various situations, such as complex tasks, incomplete information, or time-sensitive scenarios. In Part I of this study, we aim to fill this gap by evaluating how task context influences decision-making outcomes and behaviors in human-AI decision-making realm. Therefore, we seek to address the following research question in Part I of our research, specifically in chapters 2, 3, and 4:

> **RQ1** *How does task context impact user decision-making behaviour and outcomes when interacting with AI systems?*

Many decisions in real-life situations are made collaboratively by groups, such as determining a defendant's guilt or innocence in court cases and making major business strategies. Group decision-making is a complex process that involves multiple individuals with varying perspectives, biases, and decision-making styles. While our findings in Part I highlight the significant impact of task context on individual decision-making, it is unclear how the presence of AI systems affects group decision-making processes in different contexts where group dynamics can vary [80, 465]. Prior research has indicated that the value of group decision-making is especially evident in complex and uncertain tasks where diverse perspectives and expertise can reveal concealed information and counteract biases present in individual decisions [5, 203]. Nonetheless, group dynamics can improve or impede the quality of decision outcomes [32, 302] depending on their communication patterns, leadership structure, and conflict management strategies [233, 383], necessitating thorough exploration and understanding of their effects. Therefore, this section seeks to investigate the interplay between AI systems, group dynamics, and decision-making outcomes in various contextual settings. We delve into the following research question in part II of our study, comprising chapter 5:

> **RQ2** *How do task context and group dynamics influence user decision-making behaviour and outcomes when interacting with AI systems?*

Informed decision-making relies on gaining relevant and accurate information [95, 236, 237, 299]. Accessing such information is the essential preparatory step, particularly in complex scenarios involving AI systems. Information may exist in unstructured form, be reachable through search engines, or be structured and accessible from databases. However, navigating vast amounts of information to find relevant and reliable sources is often challenging. Moreover, individuals may struggle with interpreting and integrating diverse information sources to make well-informed decisions. Thus, the development of user-centric intelligent interfaces can potentially improve the outcomes of decision-making processes by improving information access [170, 282, 286, 384]. These interfaces can also utilize a range of technologies, such as machine learning, natural language processing, and data visualization, to assist users in locating pertinent information. Examples of such interfaces include search engines, recommender systems, natural language interfaces, and decision-support tools.Among these approaches, web search is still considered one of the primary sources of information. In addition, databases serve as the traditional repository of structured data that experts can rely on to support decision-making. Part III of the thesis explores how web search can be tailored to improve easier access to relevant information and examines ways to make databases accessible to a wider range of users without technical expertise. Thus, in this part of our research, chapters 6 and 7, we formulate the following research question:

> **RQ3** *How can we improve information access for users through novel interactions with web search and databases?*

1.4. RESEARCH METHODOLOGY

This thesis employs various research methodologies, such as empirical research [150, 349] and surveys [304], to address the three research questions (**RQ1** - **RQ3**). The research questions and hypotheses are grounded in established theories and existing literature on decision-making. In Chapters 4 through 7, we collect both qualitative [342] and quantitative [367] data to analyze how people make decisions and assess the effects of different strategies to improve information access. Through randomized assignments [91] of conditions to participants, this methodology maintains internal validity by addressing potential confounding variables that could affect decision-making outcomes.

Our qualitative research approaches encompass a range of questionnaires to fully comprehend individuals' perceptions and attitudes toward decision-support tools, their experiences with interactive interfaces, their decision-making strategies, and the success of the proposed interventions. In contrast, quantitative research approaches involve experimental designs and statistical analysis to gather numerical and behavioral log data to examine patterns and connections among different factors. These techniques enable measuring individuals' learning outcomes in accessing information and investigating how task-related contextual elements influence the performance of human-AI teams and individual reliance on AI systems.

This thesis applies technical HCI research methods [198] to improve decision outcomes and information access, as investigated in the first and third research questions (**RQ1**, **RQ3**). In Chapters 2 to 4, we particularly focus on designing and evaluating frameworks to enable comparative exploration and analysis of decision-making processes in the human-AI decision-making realm. In Chapters 6 and 7, we facilitate information access by developing interventions that allow the general public to quickly access relevant and reliable information. A series of tests, such as usability testing and human and machine performance tests, are conducted to evaluate the effectiveness of these proposed interventions and frameworks.

This research also leverages online communities as a research platform [394] to empirically evaluate decision-making processes or individual behavior when using decision-support tools or seeking for information in real-world contexts, addressing the three research questions (**RQ1** - **RQ3**). In Chapter 5, this thesis aims to enhance the ecological validity of the findings by conducting field studies in natural settings [288, 371]. This involves building on the application-based evaluation [115] of decision-support tools, focusing on authentic end-users and tasks as representatives of the decision-making context. This research method allows for the empirical examination of how individuals make decisions, seek information, and interact with decision-support tools in their daily lives, ultimately leading to practical recommendations for designing and implementing such tools in real-world decision-making scenarios.

In Chapters 4 to 7, we recruit participants [122] from online communities by publishing our studies on the Prolific crowd-sourcing platform, allowing us to access a diverse pool of potential candidates. Prior to their participation, we carefully screen individuals based on established criteria such as their previous study participation and success rates. Throughout the process, we closely monitor participant engagement using attention-check questions and log data to uphold data integrity. Additionally, we calculate the minimum sample size necessary to ensure our findings' validity and reliability. Furthermore, we offer compensation to ensure high engagement levels and accurate data collection through incentives for participants' time. Before full implementation, pilot studies are conducted with a small sample of crowd-workers and experts to refine research protocols and tools while validating our data collection methods. Note that all empirical studies are approved by the TU Delft ethics committee to ensure participant safety and data privacy [52].

This thesis includes observational and experimental log studies [120] to explore naturalistic individual behavior in addressing all research questions (**RQ1** - **RQ3**). Observational studies allow for understanding the pattern of behavior with existing tools and systems, while experimental log studies provide an opportunity to compare behaviors in different experimental conditions. This methodology enables a deep understanding of decision-making processes by capturing real-time behaviors and analyzing patterns and relationships among variables. The log data from Chapters 4 to 7 contains information about the type and timing of events occurring during user interactions with interfaces, offering valuable insights into how decision-making processes develop in real-world situations.

In spirit of open science principles and reproducibility, we make our data, frameworks, and analysis code public so that other researchers can validate and expand upon our findings. By employing rigorous research methodologies and diverse data collection methods, we aim to improve our understanding of decision-making processes and support the development of practical decision-support tools in various contexts.

1.5. Original Contributions

This thesis mainly contributes to the methodological and theoretical understanding of decision-making processes in various task contexts. To increase the external validity and generalizability of the findings, human-AI decision-making studies should include task-relate contextual factors, such as complexity and uncertainty, as these factors can significantly influence decision-making outcomes. While it is not feasible to account for every possible task-related contextual factor, this research aims to highlight the significance of considering and incorporating these factors when studying decision-making processes. Considering task-related contextual factors can lead to more realistic and applicable findings and contribute to the development of AI systems that better reflect and account for real-world complexities. Additionally, this thesis aims to enhance reproducibility and transparency in decision-making research by making the data, frameworks, and analysis code publicly available. Proposing an empirical framework to better investigate the impact of different parameters on decision-making outcomes, this research seeks to encourage transparent and rigorous research practices and improve the reliability of findings in the field of human-AI decision-making. The remainder of this section further explores each part of the thesis and its connection to the main contribution. It involves examining the task context, the impact of group dynamics, and the role of technology in aiding information access.

Part I: Understanding the Role of Task Context

In Chapter 2, we first assessed the existing literature on task-related contextual factors in human-AI decision-making. Having examined a few studies for reference, we acknowledged the necessity for further research on this topic. This gap is further exacerbated by the lack of standardized measures to evaluate task context in decision-making studies and compare findings across different studies. To this end, we proposed a theoretical framework for evaluating different decision tasks by their levels of complexity. Task complexity is a key factor in differentiating decision-making situations and determining the appropriate decision support. Using this framework, we compared current decision tasks, examined ongoing research efforts, and highlighted potential areas for future study.

Chapter 3 introduces 'DecisionTime', a configurable framework that integrates various variables and elements to establish a rigorous and controlled experimental setting. This modular framework allows researchers and practitioners to systematically manipulate task context parameters, enabling them to investigate the impact of these factors on decision-making behaviors and outcomes. The framework can also be customized to accommodate diverse decision-making scenarios, expanding the potential for wider applications and deeper insights. Chapters 4 and 5 utilized this framework to provide empirical evidence on how task complexity and uncertainty influence individual and group decision-making processes.

Chapter 4 delves into the impact of task context on decision-making outcomes in real-world scenarios. We have identified task complexity and uncertainty as fundamental contextual elements that can significantly influence decision-making outcomes. Varying the levels of complexity and uncertainty in simulated decision tasks, we conducted a series of experiments to examine how individuals adapt their decision strategies based on the task context. We incorporated the DecisionTime framework to create a controlled experimental environment that closely reflects real-world decision-making scenarios. In this study, we measured various decision-making metrics, including decision accuracy, decision time, participant trust and reliance on decision support, and interaction patterns. Participants could apply a variety of decision-making patterns derived from the ASPECT model. This involves considering a wide range of characteristics for attribute-based patterns and evaluating potential consequences for each option to make optimal decisions while adhering to specific guidelines. Our results showed that the contextual factors significantly influence participants' behavior. They tend to rely more on AI systems as tasks become more complex and uncertain while consistently maintaining their trust in these systems.

The original contributions of Part I are the following:

- We provide an overview of the literature from the lens of task complexity on human-AI decision-making (Chapter 2), highlighting the gaps in research and identifying the need for further investigation in this area.
- We operationalize the concept of task complexity and propose a framework for assessing task complexity in decision-making scenarios (Chapter 2), which will serve as the foundation for future studies.
- We demonstrate that the lens of complexity can provide an axis to compare decision-making tasks across different contexts and settings (Chapter 2).
- We propose a customizable framework that allows researchers to create reproducible studies closely aligned with real-world decision-making scenarios (Chapter 3). This modular framework includes adaptable components tailored to fit specific research questions and contextual factors, enabling systematic manipulation and comparison of different configurations.
- We propose a framework to operationalize task uncertainty (Chapter 4), providing a comprehensive understanding of their effects on decision outcomes.
- We examine the impact of task uncertainty and complexity on human-AI decision-making, including factors such as trust, reliance on AI systems, and performance measures (Chapter 4). This study represents the first comprehensive effort to explore the influence of task context on decision-making outcomes, with particular emphasis on complexity and uncertainty.
- We provide empirical evidence to endorse the significance of considering task complexity and uncertainty in decision-making situations (Chapter 4), which can aid in developing specific interventions and approaches to improve decision making within complex and uncertain environments.

Part I is based on the following publications:

Chapter 2 is based on a peer-reviewed paper at **UMAP '23**:

- **Sara Salimzadeh**, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), June 26– 29, 2023, Limassol, Cyprus. ACM, New York, NY, USA, 13 pages. `https://doi.org/10.1145/3565472.3592959`

Chapter 3 is based on one peer-reviewed paper at **UMAP '24**:

- **Sara Salimzadeh** and Ujwal Gadiraju. 2024. "DecisionTime": A Configurable Framework for Reproducible Human-AI Decision-Making Studies. In Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24), July 01–04, 2024, Cagliari, Italy. ACM, New York, NY, USA, 4 pages. `https://doi.org/10.1145/3631700.3664885`

Chapter 4 is based on one peer-reviewed paper at **CHI ’24**:

- **Sara Salimzadeh,** Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision-Making. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3641905.

Part II: Adding Groups to the Mix: Human-AI group Decision-Making

After investigating the impact of contextual factors on individual decision-making outcomes in Part I, we shifted our focus to exploring the dynamics of group decision-making in Part II. In Chapter 5, we explored the impact of task complexity and uncertainty within group settings on their behavior and decision-making results. We conducted experiments in which participants were assigned to groups and given decision tasks similar to the previous part, with varying levels of complexity and uncertainty. We incorporated the DecisionTime framework to create a controlled experimental environment that closely reflects real-world decision-making scenarios. Through this study, we evaluated a variety of factors, focusing in particular on performance outcomes and group efficiency. In our task context, participants could apply various decision-making patterns from the ASPECT model to make informed decisions. Group members could consider different attributes, compare the consequences of choices, and adhere to predefined rules. Collaborating in a group setup promotes socially-based patterns where the perspectives and actions of others influence members. Our findings reveal that task complexity and uncertainty have a detrimental effect on group decision-making performance. Despite longer decision-making processes due to challenges posed by complexity, the study suggests that collaborative efforts enhance group efficiency in complex situation, leading to better outcomes in such environments.

The original contributions of Part II are the following:

- We investigate how group decision-making is affected by the uncertainty and complexity of the task (Chapter 5). This research marks the first step in thoroughly exploring the context of a task in a group setting, focusing specifically on complexity and uncertainty.
- We provide empirical evidence that integrating decision support systems within a group setting can significantly improve decision-making outcomes in complex environments(Chapter 5).

Part II is based on the following publications:

Chapter 5 is based on a peer-reviewed paper at **UMAP ’24**:

- **Sara Salimzadeh** and Ujwal Gadiraju. 2024. When in Doubt! Understanding the Role of Task Characteristics on Peer Decision-Making with AI Assistance. In Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 01–04, 2024, Cagliari, Italy. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3627043.3659567

Part III: Improving Information Access: The cases of Web Search and Databases

Individuals need to gather the necessary information before making informed decisions. Depending on the specific task, they frequently rely on external resources to inform their decision-making process. This information can be collected from a variety of sources, such as search engines, which provide unstructured information that requires active searching and evaluation, or databases, which offer structured and easily accessible information. In this section of the thesis, our focus is solely on improving information access independent of the decision-making process itself. Our specific goal is to enhance the interfaces of search engines and databases to facilitate better information accessibility in the upcoming chapters.

In Chapter 6, we introduced entity cards, or information cards, as a new component to facilitate information access in search engines. These cards provide summaries of entities along with key facts, images, and links to additional resources retrieved from search results and knowledge graphs. They are designed to present users with a structured and organized view of the relevant information, facilitating the exploration of complex and unfamiliar topics or domains. To assess the effectiveness of this approach, we conducted an empirical study to measure user learning outcomes and searching behavior using traditional search interfaces and interfaces enhanced with entity cards of varying quality. Our findings suggest that entity cards significantly impact users' information-seeking behavior, helping them to formulate more focused queries and explore related entities. Additionally, participants' learning outcomes are influenced by both the complexity of the search task and the quality of the entity cards, suggesting the potential to improve outcomes in decision-making contexts.

In Chapter 7, we redirect our attention from search engines to natural language interfaces for databases. These interfaces serve as a bridge between users and databases, allowing users to ask questions and access information conversationally. These interfaces have demonstrated the potential to enhance the accessibility of database systems for a wider range of users. Nonetheless, the performance of interfaces significantly declines for complex and ambiguous questions, limiting their usability in real-world scenarios. To address this issue, we propose a question decomposition technique that breaks down complex questions into relatively simpler sub-questions that the interfaces can accurately process. Having employed this technique on a benchmark dataset, we evaluated its efficiency when inputted into the interface compared to the initial complex questions. Our findings suggest that decomposition holds promise as an approach to improve the performance of these interfaces and enhance information access process. The crowd-powered decomposition could also be a scalable solution that leverages human intelligence to provide training data to build an automatic question decomposition system on top of existing natural language interfaces. This can further improve the efficiency and accuracy of natural language interfaces for databases without modifying the underlying model architectures.

The original contributions of Part III are the following:

- We investigate the impact of entity cards on information access and search behavior (Chapter 6), providing insights on their potential to influence search behavior. At the time of our research, entity cards had not been enhanced in search engines, making our findings unique in assessing their effectiveness.
- We propose a new approach for enhancing the performance of natural language interfaces without changing the underlying technology by leveraging question decomposition (Chapter 7).
- We show that the crowd-powered approach is a promising solution for generating training data for automatic query decomposition systems and improving the efficiency and accuracy of natural language interfaces for databases (Chapter 7).
- We publish a decomposed version of the standard benchmark (Chapter 7), which is a valuable resource for future research on natural language interfaces and create automatic question decomposition systems.

Part III is based on the following publications:

Chapter 6 is based on a peer-reviewed paper at **ICTIR '21**:

- **Sara Salimzadeh**, David Maxwell, Claudia Hauff. 2021. On the Impact of Entity Cards on Learning-Oriented Search Tasks. In Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3471158.3472255. **Honorable Mention for Best Student Paper.**

Chapter 7 is based on a peer-reviewed paper at **HT '22**:

- **Sara Salimzadeh,** Ujwal Gadiraju, Claudia Hauff, and Arie van Deursen. 2022. Exploring the Feasibility of Crowd-Powered Decomposition of Complex User Questions in Text-to-SQL Tasks. In Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '22), June 28-July 1, 2022, Barcelona, Spain. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3511095.3531282.

I

Understanding the Role of Task Context

2

A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making

Recent advances in the performance of machine learning algorithms have led to the adoption of AI models in decision making contexts across various domains such as healthcare, finance, and education. Different research communities have attempted to optimize and evaluate human-AI team performance through empirical studies by increasing transparency of AI systems, or providing explanations to aid human understanding of such systems. However, the variety in decision making tasks considered and their operationalization in prior empirical work, has led to an opacity around how findings from one task or domain carry forward to another. The lack of a standardized means of considering task attributes prevents straightforward comparisons across decision tasks, thereby limiting the generalizability of findings. We argue that the lens of 'task complexity' can be used to tackle this problem of under-specification and facilitate comparison across empirical research in this area. To retrospectively explore how different HCI communities have considered the influence of task complexity in designing experiments in the realm of human-AI decision making, we survey literature and provide an overview of empirical studies on this topic. We found a serious dearth in the consideration of task complexity across various studies in this realm of research. Inspired by Robert Wood's seminal work on the construct, we operationalized task complexity with respect to three dimensions (`component`*,*

This chapter is based on a peer-reviewed paper: **Sara Salimzadeh**, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), June 26– 29, 2023, Limassol, Cyprus. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3565472.3592959

`coordinative`, and `dynamic`) and quantified the complexity of decision tasks in existing work accordingly. We then summarized current trends and proposed research directions for the future. Our study highlights the need to account for task complexity as an important design choice. This is a first step to help the scientific community in drawing meaningful comparisons across empirical studies in human-AI decision making and to provide opportunities to generalize findings across diverse domains and experimental settings.

2.1. Introduction

Recent advances in the performance of machine learning algorithms have led to a rise in human-AI decision making in a wide variety of domains. For example, recidivism prediction algorithms have been used to help judges determine whether defendants are likely to re-offend [92, 118, 274, 283], medical diseases are being diagnosed with AI systems [65, 248, 260], and loan risk prediction algorithms are employed to approve or reject loan applications [38, 43, 84, 162].

To take advantage of AI systems and achieve an ideal complementary team performance, human decision makers need to recognize the strengths and weaknesses of AI systems and effectively use AI advice to optimize their decision making. To this end, a wide variety of mechanisms have been proposed to facilitate effective human-AI collaboration such as increasing transparency of AI systems, and their interpretability. For instance, many studies provide explanations along with AI decisions to help humans interpret AI systems' decisions [43, 44, 162, 164, 278]. It is also common to present information about the AI systems to create a better perception of their functionality among users [64, 176, 243]. Prior work has also examined how human trust and reliance on AI systems is affected by different design choices through empirical studies [53, 239, 277, 314].

Apart from different features of AI systems and inherent human factors, the choice of decision tasks also affects the performance of human-AI teams [11, 27, 38]. Although several studies have examined the role of human factors in shaping interactions with AI systems, there is a limited understanding of task characteristics in the human-AI decision making context [5, 26, 334, 405]. Even though some studies incorporate tasks with different characteristics [243, 246, 258], task attributes haven't been identified systematically in the literature, so their impact on human-AI complementary performance has not been fully investigated. Consequently, there is no standard and coherent way to compare decision tasks, hindering research efforts and preventing generalizability across domains explored in empirical studies. For example, it is difficult to say how human trust shapes in the context of recidivism prediction task [424] compares to movie recommendation task [239]. Although this is not a straightforward endeavor, being able to make such comparisons will allow us to build a deeper understanding of when, why, and how humans rely on AI systems and how users can be best supported in their interactions. [115] have argued that to create and advance a 'rigorous science' in the realm of human-AI decision making, there is room for empirical work that considers functionally-grounded explanations with proxy users and tasks, human-grounded evaluation with real users and simple tasks, and application-grounded evaluation with real users and real tasks. In practice, however, it is difficult to understand the transferability of findings across these levels of empirical work. Moreover, our exploratory analysis of rationales reported for the tasks considered in recent empirical work on human-AI decision making revealed a lack of depth. For instance, in a recent study, a specific task was selected due to the abundance of datasets in its domain [38], and in another because it has been used in previous studies [344]. We believe that this contributes to – and is indicative of – the opacity around tasks and the transferability of concomitant findings.

To facilitate comparison across distinct human-AI decision making tasks, we propose the lens of **task complexity** in this thesis. Complexity of a task is influenced by task characteristics which increase information load, information diversity, or rate of information change. The complexity of tasks is an important dimension differentiating one task from the other, playing a significant role in determining the performance of a human-AI team [5, 171, 273]. It has been found to be an essential predictor of human performance and behaviour [5, 75, 273], affecting the success of team work [26]. Task complexity can also impact trust and reliance on AI systems. Intuitively, more complex tasks demand more effort from decision makers to complete and one can expect that human decision makers perform worse on highly complex tasks. On one hand, more complex tasks may imply a greater need for humans to rely on AI systems [99] as a result of increased information overload in such tasks [75]. On the other hand, human decision makers may struggle to identify errors created by AI systems on complex tasks, leading to over-reliance [26, 334]. Note that we consider the construct of task complexity independently from the users' standpoint or abilities, *i.e.*, independently from factors which influence the perceived task complexity [71].

In this chapter, we first shed light on the extent to which task complexity has been considered in the design of recent empirical studies across research communities that have explored human-AI decision making. Next, we propose a means to operationalize task complexity to facilitate comparisons across empirical works and provide us with an instrument to gauge potential transferability of findings along this axis. We thereby address the following research questions:

RQ1: How has recent research in human-AI decision making considered the influence of task complexity?

RQ2: How can task complexity facilitate a comparative lens for empirical work on human-AI decision making?

To answer the RQs, we provide an overview of the current state of human-AI decision making research through a retrospective study. We focus on studies in which decision tasks were adopted within the human-AI team setting to evaluate or improve their performance, either as the team or individual component. We limited our scope to articles published in HCI conferences and journals in the last four years, considering most relevant articles have been published in the last four years based on our preliminary analysis on Google Scholar hits. We found little evidence of task complexity being considered or controlled as a factor within the study design. Inspired by Robert Wood's seminal construct of task complexity [435], we coded different aspects of existing decision tasks based on three dimensions of complexity — `component`, `coordinative`, and `dynamic` complexity. Next, we annotated the empirical study setups in different articles in our corpus ($N = 127$) corresponding to each dimension of task complexity, highlighted current trends, and proposed research directions for the future.

Original Contributions. We analyzed recent empirical studies of human-AI decision making from an under-explored but important perspective of *task complexity*. To the best of our knowledge, this is the first systematic analysis of task complexity across empirical human-AI decision making studies. We operationalized task complexity in decision tasks, measured and annotated task complexity of decision tasks considered in recent literature across research communities. We found that tasks in the literature are distributed across all levels and dimensions of complexity. Based on our analysis, most tasks that have been considered in empirical studies have a low or medium level of component complexity. We found that highly-complex tasks generally represent real-world problems by incorporating higher risk levels and requiring domain expertise that demands a greater level of trust and reliance by humans. Despite existing limitations in operationalizing task complexity such as difficulty in accounting for features like task stakes we argue that task complexity can provide us with an axis along which we can engage in comparisons across decision tasks in empirical human-AI studies. Our work offers a starting point on which we hope that future work can build upon, extend our framework, and model various aspects and attributes of decision-making tasks in greater depth. Our findings can assist researchers in making meaningful comparisons across studies, provide opportunities to generalize findings across diverse domains, and inspire future work to tackle issues pertaining to transferability of findings in empirical human-AI decision making research.

2.2. Related Work

Human-AI Decision Making

Since AI systems have shown promising performance on various intelligent tasks like financial risk estimation [318] and medical diagnosis [24], a growing number of researchers and practitioners have begun to propose such AI systems in augmenting human decision making [245]. One main goal of such human-AI collaboration is to achieve complementary team performance [274]. For this purpose, human decision makers are expected to identify when they should rely on AI and when they should work on the tasks themselves, thereby exhibiting '*appropriate reliance*' on AI systems [253]. Only a few empirical studies have reported such appropriate reliance [245, 274]. However, there is substantial evidence that corroborates how challenging it is to foster appropriate reliance among users on AI systems [277, 445]. To promote appropriate reliance on AI systems, different interventions including explanations of AI advice [424], cognitive forcing functions [53], and user tutorials [78, 79] have been proposed in empirical studies of human-AI decision making with varying extent of success and befitting varying contexts. Existing studies have also found that human-AI decision making is affected by a number of factors. The information shown to users along with AI advice can greatly impact their trust and reliance. Explanations [274, 424], stated performance [26, 449, 463], risk perception [161], and uncertainty [401] have been studied extensively in this context. User factors like expertise [108], machine learning literacy [78], and task characteristics like task subjectivity and proximity [58], and task types [155, 202, 243, 258, 306] have also been broadly investigated. Despite the significance of tasks in the human-AI decision making field, a limited number of studies have focused on explicitly considering task complexity and understanding the impact of varying task complexity. [26] defined the number of features in each task instance as task dimensionality and conducted a user study controlling the number of human-visible features. They found that human-AI team performance decreases as task dimensionality increases. Similarly, [334] investigated how the number of features impacts the capability of participants to simulate AI predictions. Participants struggled detecting errors when faced with more task features due to information overload, which can be detrimental to the complementary human-AI team performance. In contrast, having considered two levels of complexity in their task design, [399] found that complexity does not impact participants' performance due to a learning effect. In terms of comparing human and AI systems, according to [268], AI systems outperform humans when they have access to extensive amount of information.

In this thesis, we specifically focus on the task complexity, which is under-explored in human-AI decision making studies. We first reviewed studies published in recent four years to evaluate the extent to which they take task complexity into account while designing user studies. Towards this goal, we adapted a framework to conceptualize different dimensions of task complexity and annotated the tasks accordingly. We then evaluated how tasks with various levels of complexity have been distributed in the past based on the proposed framework. Based on this review, we present our findings on how various dimensions of task complexity could influence human-AI decision making.

Task Complexity

Task complexity became a point of interest for over 50 years. In the late-1980s, some frameworks were proposed to define and analyze task complexity; they were adapted in many domains such as psychology, management, information systems, and etc. [71, 173, 273, 435]. Among all works, the theories introduced by [71, 435] gained popularity with more than 2000 citations and became the basis of other frameworks. According to [71], complexity of a task is influenced by task characteristics that increase information load, information diversity, or rate of information change. More importantly, task complexity is defined independently of any task doer's ability [71]. Aligned with this definition, [435] recognized three factors contributing to task complexity which are (i) the number of distinct pieces of information required to complete the task, (ii) the number of steps, and (iii) any changes in either piece of information or steps over time. They named these factors as component complexity, coordinative complexity, and dynamic complexity.

Through adapting the framework proposed by [435], we operationalize task complexity in empirical studies in a human-AI decision making context. Note that we study tasks given to humans, not AI systems in our thesis. We then explore how decision tasks are distributed in existing work and discuss the limitations of current experimental studies and the implication for researchers to consider task complexity as their design choice.

2.3. Method

2.3.1. Scoping Our Literature Review

We followed a semi-systematic literature review, widely adopted in prior studies [316, 340], including the four stages summarized in Figure 2.1 below.

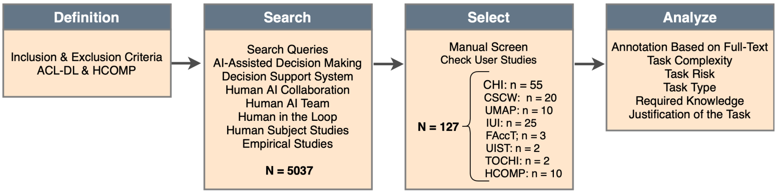


Figure 2.1: **A workflow diagram of the semi-systematic literature review process that we followed.**

Define Inclusion and Exclusion Criteria

The purpose of this study is to examine empirical human-subject studies pertaining to human-AI decision making, which evaluate or improve the performance of human-AI as a team or individual component. We applied the following inclusion and exclusion criteria to filter the articles.

- **Human-AI Decision Making**: Selected articles need to include at least one empirical human-subjects study in which humans are asked to accomplish a decision task with the aid of an AI system. We thus exclude non-empirical articles or articles focusing on tasks such as debugging, creativity, and sketching.
- **Qualitative Human-Subjects Studies**: Human-subjects studies must be evaluated quantitatively in the selected articles. Therefore, studies considering only interviews with humans to determine design decisions or asking about their preferences and understandings resulting in only filling questionnaires were excluded.
- **Proceedings**: Selected articles are published in HCI conferences or journals, including CHI, CSCW, IUI, UMAP, FAccT, TOCHI, HCOMP, and UIST within recent four years, as of January 2019 up to August 2022.
- **Format**: We included only full papers in our collection. Most HCI conferences and journal articles are published through ACM Digital Library, so we identify it as our source. For the articles in Proceedings of AAAI Conference on Human Computation and Crowdsourcing, which do not exist in ACM Digital Library, we retrieve the articles from their proceedings.

Search

We conducted an exploratory search in the ACM Digital Library to determine search queries. We searched for articles that included user studies in which participants were tasked to complete decision tasks. We retrieved 50 articles from the proceeding mentioned in our inclusion criteria using “empirical studies” and “human-AI decision making” as keywords. We manually analyzed these articles and extracted seven common keywords. We then utilize the keywords as our final search query. The search query included the following terms, "AI-assisted decision making," "decision support systems," "human AI collaboration," "human AI team," "human in the loop," "human subject studies," and "empirical studies." An initial search yielded 5037 articles after limiting the proceedings specified in our criteria.

SELECT

We manually screened the articles according to our inclusion and exclusion criteria. We looked for articles containing empirical studies by searching through the full texts of all articles with the keywords "user study" and "empirical study". We then examined the detail and type of study to decide whether we could add this article to the final collection. For instance, we removed articles containing only interviews or surveys as user studies. After excluding out-of-scope studies, we reached a collection of 127 articles.

ANALYZE

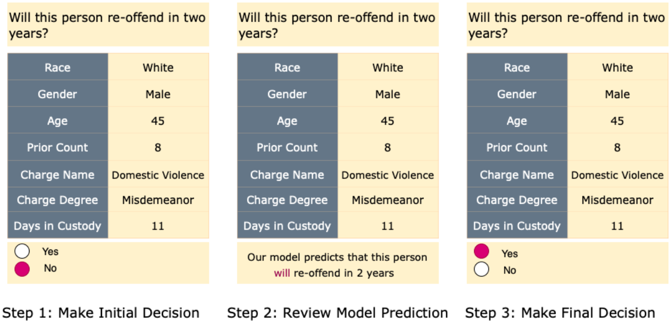
In order to evaluate how each article considered the influence of task complexity, we first reviewed the full text of the articles. We then started annotating the decision tasks by extracting relevant information such as: what kind of decision tasks, the risk of the decision tasks, how much knowledge is required to perform tasks, explicit justification of choosing decision tasks, and whether tasks are proxy tasks or actual decision making task [58]. Furthermore, we coded the component, coordinative, and dynamic complexity of decision tasks based on how many information cues are required to accomplish the tasks and the number of steps required to complete the tasks according to our rubrics explained in Section 2.3.3. In case of any changes in a number of information cues or steps, we reported dynamic complexity. We created our rubrics for operationalizing task complexity in a decision making context while annotating the articles and observing new scenarios. As identification of information cues and steps could be subjective, we discussed the rubrics iteratively along the way to ensure the integrity of the process. Authors of this chapter iterated over 30 articles, before finalizing and converging on the rubrics. The **full list of articles and our annotations** are publicly accessible for the benefit of the research community and in the spirit of open science [1].

2.3.2. OPERATIONALIZING TASK COMPLEXITY

In this section, we, introduce the general definitions represented by [435] and clarify the concepts using an example in the context of human-AI decision making. In the next subsection 2.3.3, we leverage these terminologies to explain how our framework to models task complexity in the context of human-AI decision making. Note that all of the terminologies and definitions in this section are adopted from the [435] work.

All tasks contain three essential components: products, required acts, and information cues. We follow the example in Figure 2.2 to elaborate on each component and introduce terminologies. The example presents a two-stage decision making process for recidivism prediction task. According to the defendant's profile, a human decision maker has to decide whether the defendant re-offends the crime in two years. We identified 3 constructs to calculate task complexity:

[1] https://osf.io/9bg8c/?view_only=7c0fedff68514fca892b16afa385a0e8



Will this person re-offend in two years?

Race	White
Gender	Male
Age	45
Prior Count	8
Charge Name	Domestic Violence
Charge Degree	Misdemeanor
Days in Custody	11

○ Yes
● No

Step 1: Make Initial Decision

Will this person re-offend in two years?

Race	White
Gender	Male
Age	45
Prior Count	8
Charge Name	Domestic Violence
Charge Degree	Misdemeanor
Days in Custody	11

Our model predicts that this person will re-offend in 2 years

Step 2: Review Model Prediction

Will this person re-offend in two years?

Race	White
Gender	Male
Age	45
Prior Count	8
Charge Name	Domestic Violence
Charge Degree	Misdemeanor
Days in Custody	11

● Yes
○ No

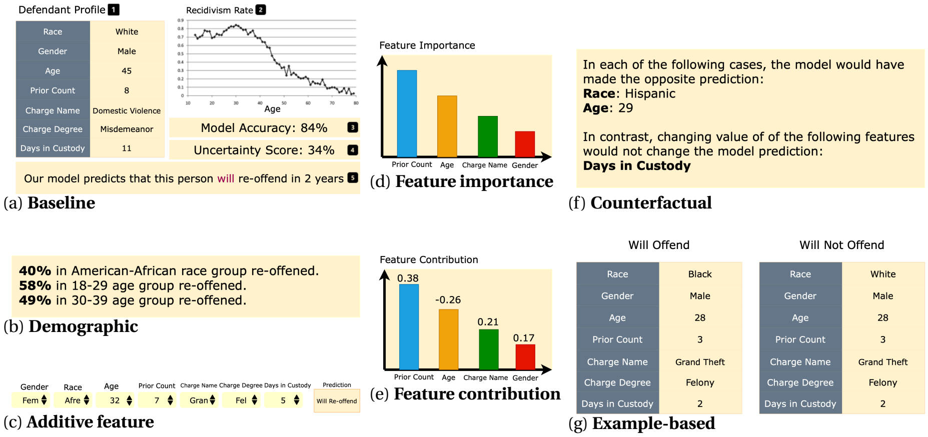
Step 3: Make Final Decision

Figure 2.2: **A decision task study. The features of the defendant profile are recognized as information cues, each step is an act, and the final decision is considered the product.**

- **Product**: Products are entities created by behaviours that can be identified separately from behaviours that produce them. They are identified as a set of assembled attributes such as an object or event and contain some defining attributes like quantity, quality, and cost. The final decision of a human is the Product in Figure 2.2.
- **Acts**: Acts serve as a specific activity or process carried out with some identifiable purpose. Acts are defined as the component of the task which is independent of an individual who performs them. In figure 2.2, making an initial decision, reviewing model prediction, and making a final decision are classified as acts.
- **Information Cues**: Information cues are pieces of information upon which an individual can make judgments during the performance of the tasks. Each variable in the defendant's profile, such as race, gender, etc., is considered an information cue. The model prediction is also a distinct information cue in Figure 2.2.

Acts and information cues are referred to as task inputs that determine the complexity of tasks. In other words, **task complexity** describes the relationship between task inputs and will be a significant determinant of individual performance. Task complexity is defined with three dimensions:

- **Component Complexity**: It refers to the total number of distinct information cues that need to be processed to perform the task. In our example, race, gender, age, prior count, charge name, charge degree, days in custody, and model predictions form component complexity as 8.
- **Coordinative Complexity**: It is defined by a number of sequences of acts that are required in the task performance. The number of steps to accomplish the task is three in figure 2.2.
- **Dynamic Complexity**: Changes in either value of information cues or number of acts lead to dynamic complexity. We count the number of information cues with variable quantities or additional steps required for accomplishing the task as dynamic complexity. In our example, both component and coordinative complexities are static during the process of decision making, indicating that the dynamic complexity is 0.



Defendant Profile 1

Race	White
Gender	Male
Age	45
Prior Count	8
Charge Name	Domestic Violence
Charge Degree	Misdemeanor
Days in Custody	11

Model Accuracy: 84% 3

Uncertainty Score: 34% 4

Our model predicts that this person will re-offend in 2 years 5

(a) **Baseline**

40% in American-African race group re-offened.
58% in 18-29 age group re-offened.
49% in 30-39 age group re-offened.

(b) **Demographic**

Gender	Race	Age	Prior Count	Charge Name	Charge Degree	Days in Custody	Prediction
Fem ▴▾	Afre ▴▾	32 ▴▾	7 ▴▾	Gran ▴▾	Fel ▴▾	5 ▴▾	Will Re-offend

(c) **Additive feature**

(d) **Feature importance**

(e) **Feature contribution**

In each of the following cases, the model would have made the opposite prediction:
Race: Hispanic
Age: 29

In contrast, changing value of of the following features would not change the model prediction:
Days in Custody

(f) **Counterfactual**

	Will Offend		Will Not Offend
Race	Black	Race	White
Gender	Male	Gender	Male
Age	28	Age	28
Prior Count	3	Prior Count	3
Charge Name	Grand Theft	Charge Name	Grand Theft
Charge Degree	Felony	Charge Degree	Felony
Days in Custody	2	Days in Custody	2

(g) **Example-based**

Figure 2.3: **The complexity of different experimental conditions of a decision Task. Participants are asked to make a prediction on whether this defendant would re-offend within two years on 30 trials. The study contains seven experimental conditions providing different types of explanations. The component complexity of each conditions is: a)12, b) 13, c) 12, d) 13, e) 14, f) 14, and g) 14 . This task has coordinative and dynamic complexities of 30 and 0, respectively. Except for condition 3, the dynamic complexity is 120. More details on how to calculate each dimension will be found on the companion page.**

Note that a task can be a combination of multiple sub-tasks. So, these definitions can be assessed at both task and sub-task levels. As a result, the overall complexity of the task in each dimension is the aggregation of the complexity across all sub-tasks. According to [435], the overall task complexity is expressed as the linear combination of component, coordinative, and dynamic complexities:

$$TC_{overall} = \alpha \cdot TC_{component} + \beta \cdot TC_{coordinative} + \gamma \cdot TC_{dynamic}.$$

However, it is not evident how the three dimensions of relate to one another [435]. Therefore, we consider each dimension separately in our study.

2.3.3. MEASURING TASK COMPLEXITY

We operationalize the theoretical model of task complexity proposed in seminal work by [435] in the realm of human-AI decision making. We model the information cues and required acts defined in previous Section 2.3.2 to decision tasks in our article collection and calculate three dimensions of complexity (i.e., component, coordinate, and dynamic) for each task. Such a framework can assist us in comparing complexity across various tasks and domains.

As a first step, we need to identify the information cues and required acts in the human-AI decision making context so we can determine each dimension of complexity defined in the previous section. To identify information cues, we created a set of rubrics. Additionally, the mandatory steps that are required to complete the task are the required acts. We categorize our rubrics into two groups: general rules applied to all cases and specific rules depending on each dimension of task complexity. In order to illustrate how these rubrics can be applied in practice, Figure 2.3 shows seven conditions of decision tasks along with their complexity.

General Rules

1 We excluded all task-independent components when calculating the task complexity: such as pilot studies, questionnaires for user factor assessment, tutorials before the actual decision task. As all of these factors do not contribute to the task complexity, they are discarded in our measurement.

2 When dynamic complexity is not zero, (due to changes in component or coordinative complexities), we report the minimum static complexity for component complexity/coordinative complexity. The dynamic dimension is indicated as the differential between maximum and minimum component/coordinative complexity.

3 Different experimental conditions of a decision task can vary in each dimension of complexity. We only consider the condition in which the authors investigate the effectiveness of their proposed approach or evaluate their primary hypothesis. Such a condition is typically the condition with the maximum complexity, among others.

4 We consider explanation methods as information cues. Although they are supposed to assist humans to interpret AI decisions, they augment task complexity as humans should digest them along with the AI decisions. However, such methods affect the complexity differently; one can directly increase component complexity,and the other may dynamically change leading to dynamic complexity.

5 Information cues can be presented in various ways, such as plots, paragraphs, tables, and images, each requiring a different amount of steps to interpret. Using a table as an information cue might be easier to digest than using a sophisticated plot. Since we do not have any references to determine the number of steps each require, we assume all of them have a similar coordinative complexity.

6 Tasks with different stakes (risks) may intuitively have different complexity levels. However, there is no way to map risk levels to either information cues or required actions. So, we consider it as the limitation of this framework as it can not capture them.

7 For each task, a set of features is required to make an informed decision. Missing any of these features will cause the task to be complex and error-prone. However, with this framework, we can not account for this type of complexity; since we cannot figure out this set of features for each specific task.

Component Complexity

1 We count the total number of distinct human-visible features, considering each as an information cue. The number of information cues indicates component complexity. Note that redundant information cues are not counted according to the definition.

2 In addition to features, the correlation between each combination of them is also considered a distinct information cue. If we have n features and their correlations, then the number of distinct information cues equals $2^n - 1$. For e.g., for $n = 2$, the

component complexity would equal three as we have three distinct information cues: feature_1, feature_2, and correlation between them.

3 Each of the following factors is examined as one information cue: 1) model prediction, 2) model uncertainty score, 3) model performance, and 4) overview of the model or algorithm distribution.

4 Each explanation method is counted as one or more information cues. 1) feature importance highlighting key features is considered as one information cue. 2) feature contribution showing top key features and their coefficients are counted as two information cues. 3) counterfactual explanation focusing on what changes in feature values result in an opposite AI prediction are recognized as two distinct information cues - as they provide both what features and which new values, 4) demographic-based information is one information cue, 5) example-based explanation such as nearest neighbour methods is considered as one information cue if only examples with similar predictions are presented; in case of providing examples with different predictions, based on the number of various predictions, they can be counted 2 to n distinct information cues.

5 The feedback regarding the performance of the humans or AI is also considered as one information cue.

COORDINATIVE COMPLEXITY

1 We count the total number of steps to accomplish the decision task as coordinative complexity.

2 Each task instance is recognized as one separate step.

DYNAMIC COMPLEXITY

1 When the quantity of any feature changes during the process of decision making (any changes in component complexity), the dynamic complexity should be greater than zero. Otherwise, the dynamic complexity is zero.

2 There is dynamic complexity if any feature affects the sequence of performing the task (changes in coordinative complexity).

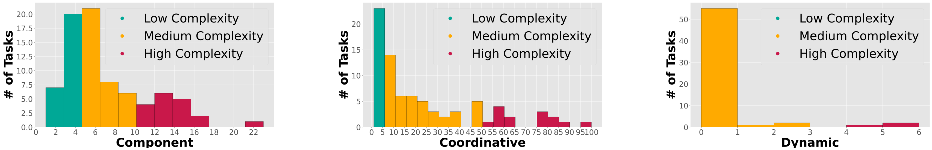
3 The additive feature attribution explanation method contributes to dynamic complexity. Providing additive feature attribution method, a human decision maker can modify the values of any features (Rule 1) and observe their correlation among other features and their impact on AI predictions. Based on Rule 2 in component complexity, the maximum component complexity with n features, given their correlation, is calculated as $2^n - 1$. As we report the differential of maximum $(2^n - 1)$ and minimum (n) component complexity as dynamic complexity, then, the dynamic complexity would be $2^n - 1 - n$.

4 If we let humans choose whether and when to see the AI recommendation, then the steps required to accomplish the task (coordinative complexity) would be dynamic depending on whether the decision makers request AI recommendation or not.

2.4. Results

2.4.1. Task Complexity in HCI Literature

RQ1 asks to what extent recent HCI literature has considered the impact of task complexity in the design of decision making tasks. Among all the relevant articles we collected, a limited number of studies have considered task complexity in designing their decision tasks [26, 334, 399]. This finding corroborates that there is no standard framework to quantify the complexity of decision tasks. We analyzed existing tasks according to the framework we proposed in Section 2.3.3, operationalizing the measurement of task complexity. We first shed light on the descriptive statistics; distribution across the component, coordinative, and dynamic complexities, the extremities observed in our data, and the outliers.



(a) **Component** (b) **Coordinative** (c) **Dynamic**

Figure 2.4: **Distribution of component, coordinative, and dynamic complexity in the decision tasks corresponding to our corpus.**

Component Complexity

Component complexity was found to be within the range of 1 to 23, shown in Figure 2.4a. The task with the component complexity score of 1 is related to a mind wandering detection task in which crowd workers were asked to identify whether people's attention in the presented video clip drifted away [49]. On the other end of the spectrum, a task related to music recommendation was found to have a component complexity of 23 [296]. In this study, a wide range of features associated with user's preferences, attributes of artists, and explanations for suggested songs were incorporated. The average component complexity of tasks in our data was found to be 6.9 (±4.3).

Research in neuroscience led by [298] revealed that the average human information processing capacity ranges between 5 and 9, which is the number of objects an average human can hold in their short-term memory. This indicates the range of component complexity for human decision makers. Based on this, we consider three bins of component complexity. First, tasks with the number of information cues (indicating component complexity) below 5 as those corresponding to low complexity. Next, tasks with 5-9 information cues are considered to have a medium component complexity, while those with more than 9 being highly complex. We note that the decision tasks considered in recent literature have a medium level of component complexity on average (6.9). In total, 33.7% of tasks have a low level, 40% have a medium level, and 26.2% have a high level of component complexity. Furthermore, 12% of tasks were found to be outliers with a high level of complexity between 24 to 132. The decision task with a component complexity of 132 relates to predicting the risk of not paying back a loan and a convict's chance of recidivism [38] within the same task. The study included 18 variables in which the three-dimensional relationship between some features was presented, including two decision tasks from different domains and many variables (shown as scatter plots) which increased the task complexity. All outliers leveraged specific datasets, included many features, and employed sophisticated plots.

Coordinative Complexity

We found that the coordinative complexity of tasks considered in our data lies between 1 and 100. There are four tasks with the lowest coordinative complexity, where participants were asked to react to the hypothetical scenario in which their Facebook account is suspended by an algorithmic content moderation system [407], movie recommendation [457], and medical diagnosis [33]. The task with the highest coordinative complexity was found to be bail decision making for 50 cases. We found that the coordinative complexity was 25.1±25.9 on average, meaning that participants must follow 25.1 steps to accomplish the task. Based on the Figure 2.4b, 75.6% of the complexity of tasks distributed between a range of 1 to 40.

We divided the level of coordinative complexity as low, medium, and high based on the quartiles; with the bottom quartile corresponding to low, top quartile corresponding to high, and the other two constituting the medium level. Tasks in the bin of low complexity corresponded to a coordinative complexity below 5; those with a medium level of complexity corresponded to between 5 to 50; highly complex tasks corresponded to a coordinative complexity of over 50. In total, 25.9% of tasks were found to correspond to a low level of coordinative complexity, 56.8% have a medium level of complexity, and 17.2% have a high level of complexity.

We also observed articles with coordinative complexity of 130 to 420 as outliers. In the task with a score of 420, participants were presented with 210 questions regarding quality control in a drinking glass-making factory scenario [450]. There is evidence to suggest that having more task instances, with a greater level of coordinative complexity, can cause mental fatigue. This is the result of prolonged periods of demanding cognitive activity [216] and has been shown to negatively affect performance [300, 427]. Therefore, it is important to set the number of task instances in empirical studies at a reasonable level to avoid the fatigue effect.

Dynamic Complexity

As explained, dynamic complexity depends on any changes in components or coordinative complexities. Our analysis revealed that dynamic complexity was distributed between 0 to 6 (cf. Figure 2.4c). We classified the level of dynamic complexity according to the bottom and top quartiles. As the result, tasks with a complexity of 0 corresponds to low dynamic complexity; task with dynamic complexity of 1 to 3 correspond to medium complexity; and tasks with dynamic complexity of 4 or more assigns to high dynamic complexity. We also observed that 95% of decision tasks have dynamic complexity between 0 to 2. This finding indicates that dynamic complexity is not common among decision tasks considered in empirical human-AI studies. The source of dynamic complexity was found to be the non-stationary nature of coordinative complexity. Some studies let the participants choose whether and when to see AI recommendations. This approach forced participants to be more cognitively involved in the decision making process by first probing the task inputs. This resulted in a variable number of steps depending on the participants, leading to dynamic complexity. We found outliers with a dynamic complexity ranging from 12 to 1890. The source of dynamicity for the study with the score of 1890 was changes in component complexity. This study includes a video activity searching tool to build specific queries and sort the number of videos about policies being followed by kitchen staff [314]. Dynamic complexity was a result of the fact that queries and responses were not constant.

ACTUAL AND PROXY TASKS

We also examined whether researchers conducted proxy tasks or actual tasks [58]. Participants in actual tasks are asked to make an informed decision with AI assistance, evaluating the performance of humans and AI as a team. In contrast, participants in proxy tasks have to simulate the model decision or decision boundaries. [58] showed how evaluations with proxy tasks do not predict the evaluation with actual tasks which can limit the generalizability of findings. In total, we observed that 86% of studies were conducted with actual tasks while the remaining 14% were proxy tasks in the set of articles in our corpus.

HIGH-STAKE AND LOW-STAKE

We also analyzed the risk of tasks as this is identified as one of four dimensions that vary in decision tasks by [245]. Among all, 67.7% of studies did not specify how risky their task was. Although this aspect could be inferred from the context, this suggests a potential lack of explicit consideration of stakes. Of the remaining, 18.3% are classified as high-stake, 8.6% low-stake, and 5.4% set up their studies in both conditions, either by changing the decision task [12] or artificially modifying the scenario. For instance, in a study by [169], participants were asked to check user requests for approval to run different software on company computers. In the high-stake domain, they were targeted by a malicious hacker, while in a low-stake setup, participants were told that they would be invited to a party if they performed well. As another approach for converting low-stake tasks to high-stake, participants were rewarded money/points in case of correct decisions and lost more amount of money/points serving as the punishment for incorrect decisions.

Note that we manually annotated tasks in which the stakes were not indicated. In total, 39.7% of tasks were found to be high-stakes, 54.8% were low-stake, followed by 5.4%, which contained both low-stake and high-stake scenarios. As there is limited understanding about the correlation between task stake and task complexity, future work could explore how these factors relate to each other and influence human-AI decision making. Example annotations of task complexity is shown in Tables 2.1a and 2.1b. The **full list of articles and our annotations** are publicly accessible for the benefit of the research community and in the spirit of open science.[2]

[2] https://osf.io/9bg8c/?view_only=7c0fedff68514fca892b16afa385a0e8

Decision Task	Complexity
Stroke Rehabilitation Assessment	(25,60,0) [255, 257]
Medical Image Retrieval	(14,6,1023) [65]
Medical Diagnosis	(7,4,2) [320],(6,1,0) [33], (6,240,0) [139]
Nutrition Prediction	(5,26,156) [56],(3,24,0) [58], (4,48,24) [147]
Recidivism Prediction	(10,64,32) [424],(11,200,50) [164], (8,12,6) [110],(132,10,0) [38], (11,50,0) [283]
Monitoring and Administration	(15,130,390) [325]
Job Application Approval	(5,24,0) [326]

(a) **Task complexity in high-stake domains**

Decision Task	Complexity
House Price Prediction	(9,15,0) [4],(11,24,12) [334], (12,40,10) [78]
Deceptive Review Prediction	(6,20,0) [243], (4,20,0) [244]
Sketch Recognition	(4,6,0) [64],(6,84,42) [460]
Movie Recommendation	(2,4,0) [239],(28,1,0) [457], (4,15,0) [266], (10,4,0) [34]
Place Recommendation	(12,2,0) [411], (51,12,0) [187]
Food Recommendation	(2,51,0) [158], (9,3,0) [305]
Image Classification	(3,12,0) [6],(5,216,0) [445], (4,90,0) [191], (3,40,0) [313]
Sentiment Analysis	(4,3,0) [377],(6,50,0) [27]

(b) **Task complexity in low-stake domains**

Table 2.1: **Example annotations of task complexity in low-stake and high-stake domains. Task complexity is shown as a tuple (component, coordinative, dynamic).**

2.4.2. TASK COMPLEXITY AS A COMPARATIVE LENS

RQ2 focuses on how task complexity can facilitate a comparative lens for empirical work on human-AI decision making. To address this research question, we examined tasks in each level of complexity, from low to high, and across all complexity dimensions. We found that there were some consistencies within each level of complexity and across dimension such as task stake, task expertise, and task type. However, there were also some differences across the levels in each dimension, which we discuss below.

DIFFERENT TASKS SAME COMPLEXITIES

Our analysis has indicated that there are different decision tasks with the same complexity. For most score levels of component complexity, there are at least two studies with the same score. Among all cases, score 3 and 5 is dominant, with 11 and 12 decision tasks, respectively. Considering low-complex tasks, they are comparable in terms of their stake, domain expertise, and task types. More than 93% are low-stake tasks that can be accomplished without domain knowledge. A majority of these tasks involve recommendations or binary decisions. Such binary decision tasks are primarily artificial, with no explicit real-world applications. As the tasks are straightforward, they imply lower demand for humans to rely on [99].

Looking towards tasks with higher levels of complexity, we observe diversity among tasks in terms of their stake and expertise, which is not comparable. For instance, we found two scenarios of child clinical decision making [227] and nutrition prediction [56] tasks with the same component complexity; the first high-stakes task requires extensive domain knowledge, while the second task can be performed without any background and has a low risk. In addition to binary decision and recommendation tasks, tasks in the bin of medium complexity were found to include multi-class, regression, and retrieval tasks. Compared to low-complex tasks, the number of tasks resembling real-world problems was found to be higher in the bin of tasks with medium level of complexity. Our examination, established that around half of the tasks were still artificial [323, 460] or do not necessarily require human intervention. For instance, sentiment analysis [377] and text classification [345] tasks can be fully automatic; thereby, human intervention may not be needed.

Lastly, on the other side of the spectrum, we found that highly complex tasks tend to be high-stake tasks requiring domain knowledge to complete. The existing low-stake tasks in this bin are dedicated to recommendation systems tasks. Including a wide range of features to capture human preferences makes such recommendation tasks complex. It is important to point out that high-complex tasks are found to be explicit examples of real-life problems. In our study, we found that as tasks got more complex, they resembled real-world use cases more, demanded more domain knowledge, and had a bigger stake. To simulate real-world problems and human interaction with AI systems, it is pragmatic to adapt actual tasks in which humans may want to rely on AI support.

We also observed decision tasks with similar scores of coordinative complexity, representing the number of steps required to finish the tasks. Tasks with low levels of complexity consist of low-stake tasks without the expertise needed. On the other end, high-complex tasks were found to have higher risks and require expertise. Interestingly, these tasks also have a medium or high score of component complexity at the same time. That could be due to the fact that researchers may increase task instances for such tasks to examine human behaviors over time. Consequently, human decision-makers familiarize themselves with AI systems, form their mental models, and calibrate their trust. Nevertheless, having a higher level of component complexity and stake for these tasks, the cognitive load of performing tasks could grow simultaneously. Such cognitive load could lead to mental fatigue in participants earlier [216]. It seems that researchers might neglect the fatigue effect in their studies. Comparing these user studies with actual scenarios, it's also rare for a human decision maker to do 100 tasks concurrently in real-life cases. Instead of doing a hundred cases in one session, it's recommended to examine human behavior over time in different sessions to mitigate the fatigue effect and model the real world better.

Based on our analysis, low-complexity tasks have almost no expertise required, are low-stake, and are easy to do, so findings across studies can be generalizable. With the increasing complexity of tasks, we observe a wide variety of task types, task stakes, more features, and a variety of explanation methods adapted, which makes it hard to carry findings from one study to another.

Same Tasks Different Complexities

In contrast to studies with the same complexity score, we found some similar decision tasks with varied complexity scores. Recidivism prediction, loan approval, movie recommendation, and image classification tasks dominate our corpus. These tasks are presented in each level of component complexity, low to high. The underlying dataset is similar throughout the studies for the recidivism prediction [38, 424] and loan approval tasks [156, 162]. They incorporate different explanation methods, enriched with additional visualizations along with the AI decision to modify the component complexity of those tasks. For the movie recommendation [239, 457], the reason for having a spectrum of complexity is integrating different user preferences to improve the quality of AI recommendations. Lastly, we can see distinct types of images in image classification tasks [6, 46], from clinical, nutrition-related, animal, and animal images. Since the context of images differs, the type and number of component complexity vary among them. Our survey showed that researchers could control component complexity by enhancing explanations, visualizations, and user preferences when making recommendations or changing domains.

2.5. Discussion and Implications

2.5.1. Implications of Our Work

Potential Reasons Why Task Complexity Has been Overlooked in Study Design

Reflecting on tasks with a high complexity, we observed that interest in promoting the need to rely on AI or opportunities to propose explanation methods can typically inform such task design choices. Researchers have shown that explanations can effectively inform mental models of humans and improve their understanding, especially for laypeople [157, 172, 390]. Additionally, to fill the knowledge gap between domain experts and laypeople or improve AI literacy, empirical studies engage with more explanations [77, 260, 356, 460]. A higher level of complexity can also result from adding more user preferences to improve the quality of recommendations [187, 305, 457]. Another reason to increase task complexity could be a need to study trust formation and reliance on AI systems in such contexts [78, 277, 334, 449, 457]. Incorporating various features can ensure that human decision-makers access salient features required to make better decisions, especially in high-stake domains [46, 65, 164, 255]. Missing a salient feature could be more harmful than presenting additional information. There's also a relationship between task complexity and the nature of the task. Tasks representing real-world cases, especially those with higher risk, tend to have more features and require more expertise to accomplish [38, 260]. Additionally, cognitive forcing interventions are applied in studies to engage human decision-makers more thoughtfully with AI systems. By increasing task complexity, such approaches affect human cognitive processes by: (I) asking humans to make decisions before seeing model predictions [424, 463], (II) varying AI systems response time [56, 323], and (III) providing feedback to humans [27, 164, 445, 450]. In terms of the arbitrary choice of task instances observed in many articles, researchers may include more instances to explore the impact of human-AI interaction over time. Having more time to collaborate with AI, human decision-makers familiarize themselves with AI systems, form their mental models, and calibrate their trust.

In contrast to orchestrating high task complexity, task complexity is mitigated in some studies. Due to the cost and limited accessibility to hire real-end users of AI systems, crowd workers simulate the decision making process. As crowd workers' knowledge is limited, decision tasks are either simplified, artificially created, or substitutes with common tasks that crowd workers have experience in are considered [58, 239, 377, 450]. Tasks with low complexity can help human decision makers have a better understanding of AI systems [64, 375].

Common Limitations and Challenges in Empirical Human-AI Studies

Our observations indicate an arbitrary selection of task design parameters like task type, number of features, etc is common in existing empirical studies. Depending on the factors that are investigated in experiments, task types can play a significant role. For instance, the pattern of reliance on AI systems in a decision task with high-stakes in healthcare domain could vary in comparison to a low-stakes task in the commercial domain. Features of a decision task are largely set from the dataset that the AI systems were trained on. However, not all of the features are relevant to a given task and some of them can even mislead users. Task parameters are also typically determined due to external factors such as associated costs, or available time. In many studies, the number of task instances, as well as the length of the study, are set according to the available budget. Furthermore, limited access to domain experts (where expertise is required) results in studies with fewer participants. Regarding whether tasks are actual or proxy tasks [58], we observed that 25% of studies in our corpus employed proxy tasks to evaluate their hypotheses. Evaluations of human-AI decision making using proxy tasks do not necessarily transfer to actual real-world tasks [58]. This pitfall can affect the generalizability and reliability of findings. A human-subject study often hinges on simulating real-world tasks accurately. While many parameters have to be simplified in isolated studies, the simulation still needs to be valid. There are sometimes tasks that are artificially created [64, 460], or tasks that do not fit into human-AI decision making are adapted [377, 445]. Explanation methods increase human understanding and transparency, but can inadvertently increase task complexity, which can be in conflict with what they are meant to achieve. It can be also detrimental to human-AI complementary performance if a lot of complicated and diverse visuals of task features are presented [390]. Although decision tasks are sometimes simplified for lay people to complete, some knowledge and familiarity, such as AI literacy [78], numeracy, and statistics background [166], may still be required to accomplish tasks, which may not be feasible to expect from all crowd workers — expert recruitment on-demand remains a challenge.

2.5.2. CAVEATS AND LIMITATIONS OF THIS STUDY

We limited our scope to articles published in HCI venues published in the last four years. Although our corpus is representative of literature, our sample frame may have resulted in not considering related articles published in other venues. We do not claim to provide exhaustive insights into why task complexity has not been widely considered in empirical human-AI studies. As the first to model task complexity in a human-AI decision making context, our thesis advances the current conversation in this community. Further work is required to extend the operationalization of task complexity to incorporate other task characteristics, and differentiate diverse methods of information visualization (e.g. plots, text, images), or task stakes. We hope to inspire future work in proposing methods to help inform and facilitate meaningful comparisons across empirical studies on human-AI decision making.

2.6. CONCLUSIONS AND FUTURE WORK

In this chapter, we examined to what extent recent literature in human-AI decision making has considered task complexity in the design of empirical studies (**RQ1**) and how task complexity can facilitate comparisons across experimental settings (**RQ2**). To answer our research questions, we reviewed the published literature on human-AI decision making tasks in the last four years. We found little evidence of its consideration as a design parameter. We then operationalized task complexity based on Robert Wood's seminal work. We analyzed different dimensions of task complexity and measured them using a set of well-defined rubrics. Our analysis found that tasks in the literature range in complexity across all levels and dimensions. Most of the tasks considered in empirical studies have low or medium complexity. The most complex tasks, which largely resemble real-world problems, were found to have higher risk levels, requiring domain expertise. Despite the limitations in our operationalization of task complexity — we did not account for other task characteristics that may effect task complexity — we found that it can still provide us with an axis for comparing decision tasks in human-AI studies. Such comparisons are particularly meaningful in tasks with lower levels of complexity. Based on our analysis of empirical human-AI studies, we found that it is important to measure and report the different types of expertise or domain knowledge that the participants might have (numeracy, AI literacy, familiarity with statistics or information visualization), so that comparisons across studies can be made meaningfully. Future work in this realm can consider explicitly controlling the level of task complexity across the experimental conditions. In the imminent future, we aim to expand our operationalization of task complexity to account for other task features and build a tool that can automatically measure the complexity of tasks across Wood's three dimensions and inform researchers in their design of empirical human-AI studies.

3

“DecisionTime”: A Configurable Framework for Reproducible Human-AI Decision-Making Studies

Empirical studies have extensively investigated human decision-making processes in various domains where AI systems are incorporated. However, comparing and replicating these studies can be challenging due to different experimental configurations. Moreover, the existing contexts often have limited scope and may not fully capture the complexity of real-world decision-making scenarios that are riddled with varying levels of uncertainty. Our framework addresses these practical gaps by providing a configurable and reproducible environment for conducting human-AI decision-making studies in the route planning domain that captures many complexities of real-world scenarios. Researchers can customize parameters, conditions, and factors involved in decision-making tasks to help address research and empirical gaps through rigorous experiments. With various modules such as map generation, chat components, and different AI systems available within the “`DecisionTime`” framework, researchers can effortlessly design experiments exploring multiple aspects of human-AI interaction and decision-making.

3.1. Introduction

AI systems have been increasingly adapted in various domains to assist individuals in making decisions that directly impact their lives, such as allocating resources [84], recommending products [239], and even predicting medical conditions [256]. AI systems

This chapter is based on a peer-reviewed paper: **Sara Salimzadeh** and Ujwal Gadiraju. 2024. “DecisionTime”: A Configurable Framework for Reproducible Human-AI Decision-Making Studies. In Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct ’24), July 01–04, 2024, Cagliari, Italy. ACM, New York, NY, USA, 4 pages. `https://doi.org/10.1145/3631700.3664885` can potentially enhance the effectiveness of decision-making processes and the overall quality of outcomes as decision-support tools [54]. With AI systems having the potential to make significant real-world decisions that can significantly affect individual lives, it is essential to guarantee these systems' trustworthiness, fairness, and accountability [247]. Researchers have investigated various factors affecting the interaction between humans and AI in decision-making, aiming to understand their impact on human behavior [167]. This understanding can contribute to developing approaches to enhance the human-AI decision-making processes.

Numerous empirical studies have explored the factors that influence interactions between humans and AI systems in decision-making situations, such as AI systems' explainability [408], the accuracy of AI suggestions [179], the level of trust and reliance on AI systems [25, 181], and the decision context in which AI systems are employed [334]. Empirical studies involve human participants engaging in decision-making activities with assistance from AI systems. This allows researchers to observe and analyze their behavior, decision-making processes, and perceptions. Previous studies have also developed strategies, methods, and techniques to address existing challenges and utilize empirical investigations to evaluate the efficiency of these strategies. These methods may involve creating explainable AI systems [315], offering visualizations and interfaces that improve transparency and trustworthiness [105], and implementing evaluation metrics [359] to assess the impact of AI systems in decision-making processes.

While these empirical studies provide valuable insights into the dynamics of human-AI decision-making, there is a need for a more systematic and reproducible approach to designing experiments in this field [115]. This is crucial for laying a solid groundwork for advancing and assessing AI systems in decision-making. The context and domain of the decision-making tasks is a critical factor in designing empirical studies, as it determines the relevance and applicability of the findings to real-world scenarios [247]. It also influences the selection of appropriate evaluation metrics and experimental protocols [247]. For instance, the trustworthiness of AI systems would be evaluated differently in healthcare decision-making with high-stakes and potential harm compared to a low-stakes scenario like recommending movies. However, current decision tasks vary widely and lack standardization, making comparing results across studies challenging. T hey are often too simplified or artificially created, lacking ecological validity and real-world complexity [354].

To address these challenges, we propose "`DecisionTime`", a configurable framework for designing reproducible studies in a route-planning context. Researchers and practitioners can manipulate and measure various variables in a controlled manner or extend the framework to incorporate more context-specific AI-related factors as needed. With these configurations, they can systematically investigate the impact of different variables on human-AI decision-making, including AI performance, system explainability, human trust and reliance, as well as contextual factors such as task complexity, time pressure, stakes, and information availability. Furthermore, a diverse range of participants with different levels of expertise can be engaged as individuals or in groups to perform tasks in route-planning domain. `DecisionTime` allows for a more precise representation of decision-making scenarios in terms of real-world complexity and user diversity, thereby enhancing the generalizability of the findings. Designing such a framework can be adapted for various decision-making domains to ensure reproducibility and enable meaningful comparisons across different studies. To the best of our knowledge, this is the first framework that addresses the need for systematic and reproducible experimental studies in human-AI decision-making using a configurable route-planning context.

The contribution of this work can be summarized as follows:

> A **configurable** framework that enables researchers to design reproducible studies in route-planning contexts and potentially other decision-making domains. The framework will facilitate the design of experiments that closely align with real-world decision-making scenarios, increasing the ecological validity and generalizability of the findings.
>
> A **modular** framework with adaptable components that can be customized to fit specific research questions and contextual factors. Our framework provides extensive control over the variables, allowing for systematic manipulation and comparison of various configurations. In prior research, experiments mainly have been conducted using the same dataset to train AI systems, resulting in limited flexibility.

3.2. Architecture

3.2.1. Decision Domain

We selected route-planning as the specific decision domain for this framework due to several factors. First, route-planning is a common decision-making context in everyday life, making it relevant and relatable to many individuals. Participants of empirical studies can also perform route-planning tasks individually or collaboratively in groups, allowing for diverse participation and capturing of different decision-making dynamics. Second, route-planning involves considering multiple factors, such as time constraints and various possible route options, and evaluating cost, efficiency, and potential risks associated with routes. Various goals, such as minimizing travel time, avoiding toll roads, or optimizing fuel consumption, can be incorporated into the decision-making process, enabling a rich and diverse set of decision scenarios and tasks to be explored within this domain. Third, route-planning often involves the use of AI algorithms and technologies, making it a suitable domain for studying human-AI decision-making interactions and evaluating the impact of varying attributes in AI systems on decision outcomes.

`DecisionTime` framework comprises three central layers: a map and metadata generator, a web server, and a user interface. Researchers and practitioners can choose to utilize the layers separately or together to construct a tailored framework based on their specific research objectives and needs. **All code for this framework** is made publicly available to support future research in the community [1].

3.2.2. Map and Metadata Generator

This is the core layer of the framework that generates realistic maps and relevant metadata that represent the decision-making task. The map and metadata generator utilizes

[1] https://anonymous.4open.science/r/rp-364E/README.md

spatial data and tailored functions to create maps with various features such as roads, landmarks, and points of interest. This metadata is based on OpenStreetMap data obtained using OSMnx [45], a Python library for retrieving, modeling, and analyzing street networks. The traffic, weather, and other related variables can also be incorporated into the generated maps to provide a more realistic environment for decision-making. Maps can be generated into HTML files, allowing easy integration with the web server layer, shown in Figure 3.1. The image of the generated map can also be exported for visualization purposes. Metadata is stored in JSON format, allowing researchers to access and manipulate the data for analysis and experimentation.

3.2.3. Web Server

The web server layer acts as the backbone of the framework, providing the infrastructure to support data storage, processing, and communication between different components. It handles requests from user interface, manages map data and metadata storage and retrieval, and facilitates communication with the user interface layer through the REST APIs. The existing APIs are programmed for decision-making studies, while they can be extended to include new functionalities based on specific research needs.

3.2.4. Interface

The user interface layer is responsible for providing a user-friendly interface through which participants of empirical studies can interact with the framework. It includes map visualization, route-planning, and decision-making functionalities. The design of the user interfaces enables researchers to use them as they are or customize them according to their specific study requirements. Additionally, the user interface allows researchers to collect data on participants' decision-making processes and interactions with the existing components.

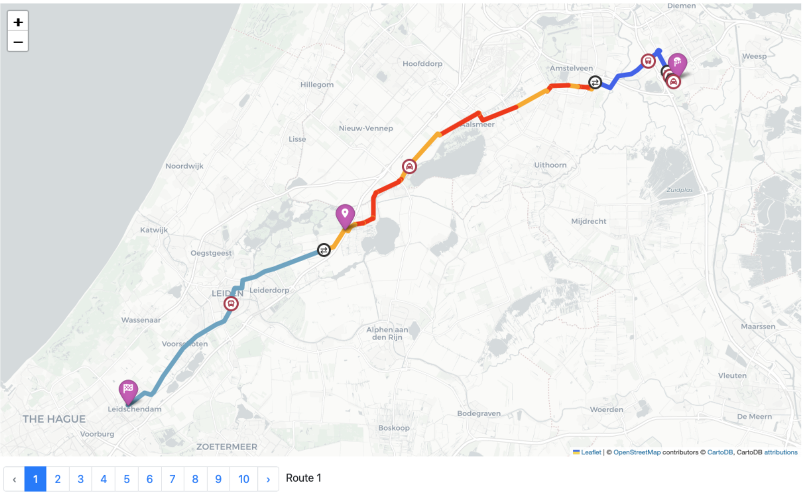


Figure 3.1: **An illustration of the map within the interface, allowing participants to engage with the system and make decisions.**

3.3. CUSTOMIZATION

Customization is a crucial aspect of the DecisionTime framework, allowing researchers and practitioners to tailor it to their needs and objectives. While every architectural layer provides default functionality and customization, the main contribution of this framework lies in the ability to customize the map and metadata generator, as well as the extensible user interface to widen the scope of decision-making studies.

3.3.1. BASIC MAP CONFIGURATION

The framework consists of several basic configurations for **map generation** that can be customized based on specific research requirements to ensure reproducibility and flexibility. These configurations include:

ROUTE-SPECIFIC ATTRIBUTES

Researchers can identify the particular characteristics of a route that aim to be considered in the decision-making processes. These attributes may include distance, time, traffic conditions, road quality, and potential hazards. The choice of particular attributes can depend on the research questions and design choices. The presentation of attributes can also be tailored, for example, by displaying them as numerical values or graphical representations on a map. The initial and final points of the route, as well as any stops in between, can be adjusted to align with the researchers' needs. Additionally, customizing visual elements such as colors and icons for presenting routes or locations is essential for creating an intuitive and user-friendly interface for participants of empirical studies.

MAP DISPLAY

DecisionTime framework provides different options for displaying the map interface for researchers. These options include different map styles (e.g., satellite view, street view) and zoom levels. Furthermore, researchers can include additional information on the map, such as points of interest, landmarks, and different types of roads (e.g., highways, local roads). The map display can also be interactive, allowing participants to zoom in and out and toggle between different layers of information. Maps can be formatted as HTML files to be displayed on web browsers or as images embedded in user interfaces or presentation materials.

3.3.2. MODULAR MAP AND METADATA GENERATOR

Researchers are able to study how various factors affect human-AI decision-making by adapting the framework's map and metadata generator module. These factors can be divided into contextual elements and attributes of AI systems, which can be systematically manipulated to create different experimental scenarios. Contextual factors include the complexity of the decision-making task, the level of uncertainty in the environment, the presence of time constraints, risk factors, the availability of information, and the cognitive effort required to process the data. Additionally, attributes of the AI systems can be manipulated, such as the reliability of the AI predictions, the level of transparency in the decision-making process, the degree of autonomy given to the AI system, updating the knowledge of AI systems over time, and the level of user control or interaction with the AI system. By systematically manipulating and operationalizing these factors on the generated maps and metadata, researchers can study how different combinations of contextual elements and AI attributes influence human-AI decision-making outcomes.

3.3.3. In-Built Scaffolding

The framework contains in-built scaffolding modules that enable researchers to assist participants before the decision-making phase. These supportive modules are designed to minimize potential biases or misunderstandings, ensuring participants make decisions based on accurate information and a comprehensive understanding of the task. This ultimately leads to more reliable and valid data for analysis. The scaffolding components consist of tutorials, detailed guides, and hand-in experiences designed to assist participants in understanding the contextual factors and getting acquainted with the interface and navigation features. Standardizing these modules while adapting them to specific task details ensures consistency across different studies while maintaining consistent support for all participants. Additionally, we incorporate quiz questions as part of our sample assessment process to gauge participant comprehension of material covered in the scaffolding modules and their readiness for engaging in the decision-making process. These modules can be augmented, removed, or tailored based on the specific needs and requirements of the study, allowing for customization and flexibility in the implementation process.

3.3.4. Chat Component Extension

The user interface of `DecisionTime` supports a chat component extension that facilitates collaborative decision-making and group discussions. This feature enables groups of various sizes to engage in real-time conversations, exchange perspectives, and collectively reach consensus. Additionally, participants can use the chat to ask questions, seek clarifications, and receive timely feedback from facilitators or researchers. It also provides an avenue for sharing feedback on AI predictions or suggestions, contributing to a more interactive and dynamic decision-making process.

3.3.5. Online AI System Extension

Our framework can integrate multiple AI systems to suggest and generate decision-making options for participants according to their preferences, task objectives, and contextual information in offline and online settings. In offline scenarios, the AI system provides pre-computed suggestions to participants during the decision-making process. This represents the default operation mode, during which maps and routes are pre-generated to facilitate efficient navigation and exploration. For real-time interactions in online setting, AI suggestions can be augmented into the user interface with live data feeds, allowing participants to receive up-to-date information and recommendations where maps and routes are dynamically generated based on real-time data.

3.4. Conclusion

In this work, we introduce our configurable framework, `DecisionTime`, for reproducible human-AI decision-making studies, which incorporates various modules to support effective decision-making processes. Our framework empowers researchers to design empirical studies tailored to their needs and research questions, highlighting the need for a systematic and rigorous approach to studying human-AI decision-making. It offers different layers of functionality and adaptability, from map generation to setting up the entire study flow. With scaffolding modules, a chat component extension, and an online AI system extension included, our framework provides a comprehensive tool-set for conducting human-AI decision-making studies. In the future, our goal is to improve our framework by broadening the scope of study scenarios and domain applications, enhancing data analysis capabilities, and incorporating additional templates for simple reproduction and adaptation for a larger community.

4

Dealing with Uncertainty: Understanding the Impact of Prognostic vs. Diagnostic Tasks on Trust and Reliance in Human-AI Decision-Making

While existing literature has explored and revealed several insights pertaining to the role of human factors (e.g., prior experience, domain knowledge) and attributes of AI systems (e.g., accuracy, trustworthiness), there is a limited understanding around how the important task characteristics of complexity and uncertainty shape human decision-making and human-AI team performance. In this work, we aim to address this research and empirical gap by systematically exploring how task complexity and uncertainty influence human-AI decision-making. Task complexity refers to the load of information associated with a task, while task uncertainty refers to the level of unpredictability associated with the outcome of a task. We conducted a between-subjects user study ($N = 258$*) in the context of a trip-planning task to investigate the impact of task complexity and uncertainty on human trust and reliance on AI systems. Our results revealed that task complexity and uncertainty have a significant impact on user reliance on AI systems. When presented with complex and uncertain tasks, users tended to rely more on AI systems while demonstrating lower levels of appropriate reliance compared to tasks that were less complex and uncertain. In contrast, we found that user trust in the AI systems was not influenced by task*

This chapter is based on a peer-reviewed paper: **Sara Salimzadeh,** Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision-Making. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3641905.

complexity and uncertainty. Our findings can help inform the future design of empirical studies exploring human-AI decision-making. Insights from this work can inform the design of AI systems and interventions that are better aligned with the challenges posed by complex and uncertain tasks. Finally, the lens of diagnostic versus prognostic tasks can inspire the operationalization of uncertainty in human-AI decision-making studies.

4.1. Introduction and Background

With the emergence of human-AI decision-making as a prominent paradigm across various domains, numerous investigations have been dedicated to understanding the factors that can impact trust and reliance on AI systems [277, 449, 463]. Such factors can be broadly classified into three primary categories: human-related factors [121, 313, 314], attributes of the AI systems [315, 321], and characteristics of the decision-making tasks [57, 202, 408]. Human factors such as prior experience [356, 390], cognitive biases [281, 326], and AI literacy [78], which can shape individuals' perceptions and interactions with AI systems. Attributes of the AI system include aspects such as predictions generated by AI [239, 252, 323], information about model predictions [27, 110, 311], as well as interventions that impact cognitive processes [54]. Furthermore, the level of trust and reliance on AI may differ across various domains and applications due to the attributes associated with decision tasks [155, 409].

The characteristics of tasks have been demonstrated to play a pivotal role in determining the level of reliance on AI systems, emphasizing the importance of methodically recognizing and comprehending these features in human-AI decision-making context. However, limited task characteristics have been systematically explored and their impact on human reliance on AI systems is not yet fully understood [247, 354]. Although a few studies have included multiple tasks with varying attributes [12, 46, 416], a systematic and empirical understanding of task features is notably absent from existing literature [247, 354]. Additionally, it remains unclear whether task attributes chosen in existing empirical studies have been appropriately considered, in a manner that is commensurate with the claims of the studies [155, 247, 258]. These limitations have the potential to undermine the credibility and generalizability of research findings, hindering our progress in developing effective strategies for human-AI decision-making [247, 354].

In this work, we propose empirically examining **task complexity** and **task uncertainty** as two essential objective task characteristics that are manipulable from the task's standpoint. *Task complexity* pertains to the characteristics of a task that contribute to an increased load of information [435], and it is distinct from task difficulty [324], which relates to an individual's perception of the task-based on their capabilities and previous experience [435]. It has been shown that task complexity is a crucial factor in determining both human performance and behaviour [5, 75, 273], as well as the success of human-AI teams [26]. Additionally, prior work has demonstrated that individuals tend to rely more heavily on AI systems when confronted with more complex tasks [100] due to the challenges associated with analyzing large volumes of information [75]. In line with work by [324, 408], we operationalize task complexity as an objective task-related characteristic that can be measured based on the number of constraints involved in the task. On the other hand, the level of *task uncertainty* refers to the extent of unpredictability inherent in a given task [17]. We operationalize uncertainty in our study using **diagnostic** and

prognostic tasks to capture different levels of uncertainty. *Diagnostic* tasks involve situations where participants are provided with detailed and comprehensive information about the task, (theoretically) enabling them to make accurate decisions. *Prognostic tasks*, on the other hand, involve situations where participants must make predictions about future events based on incomplete or limited information. By operationalizing uncertainty in this manner, we can effectively capture the diverse levels of uncertainty that arise from the inherent nature of a task and its connection to information availability. Intuitively, in prognostic tasks, users can benefit from using AI systems due to their ability to reduce uncertainties, particularly when choosing the optimal route for a future trip by considering anticipated weather and traffic conditions. Unlike planning immediate trips, this task entails a greater degree of uncertainty owing to future events' unpredictability.

Prior work has highlighted that appropriate trust and reliance play a critical role in achieving complementary human-AI team performance [207, 297, 450, 460]. Thus, it is essential to comprehend how task-related factors impact human trust and reliance on AI systems, as separate constructs [232, 297, 358], to foster successful collaboration between humans and AI. We thereby address the following research questions:

RQ1: How does task complexity influence user trust and reliance on an AI system?

RQ2: How does task uncertainty, characterized by *prognostic* versus *diagnostic* tasks, influence user trust and reliance on an AI system?

RQ3: How does task complexity interact with task uncertainty to shape user trust and reliance on an AI system?

To address these research questions, we selected the real-world scenario of *trip-planning* where both task complexity and uncertainty are prominent factors. In such scenarios, individuals are confronted with circumstances that necessitate a choice between relying on an imperfect AI system or exercising their own judgment. We conducted a 3 (*task complexity*) × 2 (*task uncertainty*) between-subjects study with 258 participants recruited from the Prolific crowdsourcing platform.

We found that users' reliance on the AI system varied depending on the level of *complexity* and *uncertainty* in the task. Individuals facing tasks characterized by `medium` complexity and uncertainty *i.e.,* prognostic tended to rely excessively on the AI system. However, their ability to differentiate accurate AI advice from misleading advice was compromised, leading to a relatively low appropriate reliance, a higher over-reliance on AI, and subsequently lower overall task performance. However, we observed a point of transition where participants started to increase their appropriate reliance on the AI system. This led to enhanced overall performance in prognostic tasks with high complexity, revealing a significant interaction between complexity and uncertainty.

4.2. RELATED WORK

Human-AI Collaborative Decision-Making

In recent years, the use of AI technologies has evolved to encompass more collaborative approaches that involve both humans and AI systems working together [9, 64, 65, 257, 418]. While fully automated decision-making by AI systems may not always be appropriate, certain tasks still require human judgment. For example, in high-stake scenarios such as in the medical [139, 227, 241, 320], legal [12, 267, 283, 416], and financial [84, 129, 156, 160, 163] domains, individuals tend to exhibit a preference for human decision-makers over AI systems. This preference could be motivated by ethical and legal concerns [247, 255, 329], as well as a desire for individual agency and accountability [191, 244, 267, 375]. Additionally, it may also stem from the limited trust [61, 62] surrounding AI systems, coupled with concerns about potential biases or errors in algorithms [260, 399], particularly when human lives or ethical considerations are at stake due to possible failures of AI systems [247, 255, 329].

The primary objective of integrating human and AI is to unite their respective strengths, resulting in enhanced decision outcomes through complementary capabilities [54, 182]. To this end, previous research has focused on identifying the factors that influence human-AI decision-making. Recent studies have explored variables that contribute to the fairness [39, 110, 252, 417] and trustworthiness [119, 167, 266, 450] of AI systems, as well as the impact of assigning different decision-making roles to humans and AI on the reliance on such systems [188, 327, 403, 465]. Prior work has also been dedicated to developing and evaluating interfaces [49, 105, 291, 296] and visualizations [156, 166, 440, 445, 457] aimed at improving human-AI collaboration.

Trust and Reliance on AI Systems

It is important to distinguish between trust and reliance, as they have different implications for the context of human-AI decision-making. [253] proposed the following definition of trust, which we adopt for the scope of our work:

> *Trust is an attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability.*

Reliance, on the other hand, refers to the extent to which individuals rely on AI systems [253, 410]. When user decisions differ from AI advice, there are mainly three discernible patterns of reliance behavior [18, 359, 363], (i) appropriate reliance, switching to the AI advice when it is correct and overriding it when it is incorrect, (ii) over-reliance, excessively relying on AI advice even when it is incorrect, and (iii) under-reliance, not fully utilizing AI advice even when it is correct. While trust is an essential factor in determining the level of reliance on AI systems [200, 232, 253, 358], it is not always a guarantee. Prior studies have shown that individuals may not necessarily increase their reliance on AI systems even if they trust them [229, 232, 297]. Instead, they might rely more on their own judgments despite acknowledging the capabilities of the AI system. This highlights that the *trusting behavior* of users can differ from their *trusting beliefs*. The evaluation of the system's trustworthiness by individuals to establish perceived trustworthiness significantly influences (subjective) trust and trusting behaviour (i.e., objective reliance) [361]. Even if a system is trustworthy, it does not automatically ensure accurate perceived trustworthiness [25, 361]. To align the perceived trustworthiness of AI systems with their actual value, it is essential to consider aspects like the availability and relevance of system information and the detection and utilization of this information by human decision-makers [361]. Trust in AI systems, namely perceived trustworthiness, can be evaluated through different methodologies, including subjective self-reported measures [81, 213, 232, 430] and relatively objective trust-related behavioral measures [182, 410, 449, 463], such as agreement and compliance.

Through a wide range of studies, researchers have consistently found that reliance on AI systems is influenced by various factors including human-related aspects [130, 182, 267, 325, 399], attributes of the AI systems [156, 263, 345, 346], and characteristics of the decision-making tasks [26, 33, 57, 160, 408]. Human factors encompass a variety of individual characteristics, including previous experience [313, 356], cognitive biases [281, 326], and AI knowledge [78]. For instance, cognitive [121, 247, 314] or meta-cognitive biases [182] have the potential to influence how individuals comprehend and appraise the outcomes generated by AI systems which in turn can affect their reliance on AI. In addition, the attributes of AI systems can enhance decision-making outcomes [247], which include aspects such as predictions generated by AI [239, 252, 323], information about AI predictions or AI systems themselves [34, 244, 380, 442], and interventions that impact cognitive processes [238, 323, 334]. For instance, various explanation methods have been explored to enhance the interpretability and transparency of AI algorithms, allowing humans to better understand AI advice [4, 180, 239]. [25] discovered that reliance on AI systems is negatively affected when untrustworthy AI systems overstate their capabilities compared to trustworthy ones. This is primarily because users struggle to differentiate between the competence of trustworthy and untrustworthy AI systems, leading to deception and excessive reliance on the untrustworthy system. Moreover, the characteristics of the decision-making tasks can also significantly impact human reliance on AI systems [247, 354]. Hence, the level of reliance may differ across various domains and applications due to the attributes associated with decision-making tasks [155, 409]. For instance, in high-stake fields like healthcare or finance, individuals may exhibit distinct behaviours compared to low-stake areas such as entertainment [296, 442].

Recent research has revealed several challenges in fostering appropriate reliance on AI systems. Prior work has shown that depending on different factors [408, 464], users may blindly follow AI advice, leading to over-reliance [54], or underestimate the capabilities of AI, resulting in under-reliance [130, 416]. To overcome such challenges and improve performance-related outcomes, it is important to ensure that users can strike a balance between utilizing AI effectively while also considering the limitations of a given AI system. To this end, researchers and practitioners have explored the use of explanation methods [239, 315, 408], interventions such as tutorials [78, 277] and cognitive forcing functions [54] to foster appropriate reliance on AI systems with varying degrees of success.

Building on the body of literature, our study aims to enhance the comprehension of appropriate reliance on AI systems in human-AI decision-making by investigating how task complexity and uncertainty influence user trust and reliance. To this end, we conducted a between-subjects study in the context of trip-planning task. We measured the extent to which individuals rely on AI systems for decision-making in various conditions by leveraging a series of common metrics in the field.

Task Characteristics in Human-AI Decision-Making

Although much attention has been given to the effect of human and AI-related factors in shaping human reliance on AI, few studies have explored the influence of task characteristics. [258] found that individuals exhibited lower trust in AI systems in tasks that involve human skills, such as work evaluation, compared to tasks that require more analytical skills. Additionally, [408] has also examined the concept of task difficulty by considering the cognitive load required. Their findings indicate that as tasks become more difficult, there is a tendency among users to rely excessively on AI advice, leading to over-reliance. A few studies have also explored the effect of task features on human-AI team performance. [26] conducted a study where participants had to assess whether objects passing through a pipeline were defective or not. They manipulated the complexity by changing the number of the task features, such as color, shape, and size. They found that an excessive number of task features diminished the performance of human-AI teams significantly. Similarly, in a study by [334], participants were presented with varying numbers of features to predict apartment selling prices. The features included variables such as the number of rooms, area size, days on the market, distance to amenities, and building maintenance fees. They also found that participants struggle to distinguish AI errors in tasks with more features, leading to decreased performance. In contrast, [399] showed that the complexity of tasks did not significantly impact human-AI performance due to a learning effect. They conducted an experiment in which participants were tasked with finding a suitable house based on a set of constraints. The complexity of the tasks was manipulated, with some scenarios having three constraints (such as rent type, budget, and registration condition), while others had five constraints (including rental duration and proximity to amenities). [57] conducted a study examining the influence of proxy tasks, where participants were tasked to anticipate AI advice, compared to actual tasks where participants directly received AI advice. Their results indicate that participants' behavior in proxy tasks did not align with their behaviour in actual tasks, underscoring the importance of carefully designing experiments to draw valid conclusions. Additionally, high-stake [12, 160, 163, 320] tasks and low-stake [158, 160, 239] tasks have been studied individually in literature in relation to human reliance on AI systems.

Furthermore, there is a lack of comprehensive investigations into categorizing task attributes and their specific implications for human-AI decision-making [354]. [247] proposed a framework that categorizes task characteristics in terms of their domain, required expertise, risk, and subjectivity. According to [243], tasks can also be differentiated based on whether they are emulating human intelligence, like object recognition [63], or based on discovered patterns in data such as recidivism prediction [283]. Some prior works have also provided a taxonomy of task types existing in the literature [1, 307]. However, these taxonomies often focus on general task types rather than specifically addressing the impact of these characteristics on human-AI decision-making. [17] introduced diagnostic and prognostic tasks in which there is clear grand-truth in diagnostic tasks, while prognostic tasks involve making predictions about future outcomes. They emphasized that the level of inherent uncertainty in predicting future outcomes is a crucial factor that can impact human reliance on AI systems. Inspired by this work, we operationalize task uncertainty in our study using the distinction between diagnostic and prognostic tasks.

In this thesis, we aim to fill an empirical and research gap by examining the impact of task complexity and uncertainty, as important attributes in decision-making in real-world contexts. By providing application-grounded evaluation [116] with users relying on an AI system for assistance in practical tasks, our work is the first to explore task uncertainty and how task uncertainty interacts with task complexity in shaping human-AI decision-making.

4.3. HYPOTHESES AND TASK DESIGN

4.3.1. HYPOTHESES

The degree of task complexity is deemed one of the primary indicators for determining the success of Human-AI teams [5, 26, 75, 273]. Consequently, it can be anticipated that as tasks become more complex, their influence on human reliance on AI systems increases [75, 268, 334]. More complex tasks tend to require more cognitive effort [75], making individuals more likely to rely on AI systems for assistance. Moreover, as task complexity increases, the verifiability [140] and plausibility [206, 215] of AI advice tend to decrease. This can pose challenges for individuals in distinguishing misleading AI suggestions, leading to reduced levels of appropriate reliance on AI systems. Although there may not be a correlation between trust and reliance on AI systems [232, 253, 297, 362], prior work suggests a higher likelihood of individuals placing greater trust in AI systems for more complex tasks [190, 253].

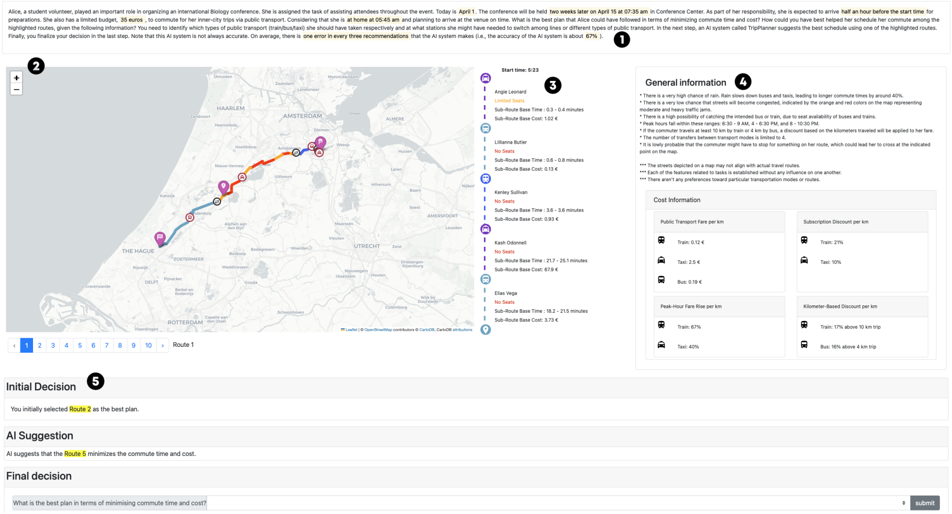
When faced with prognostic tasks, individuals are likely to perceive them as more complex and unpredictable, thus increasing their reliance on AI systems for assistance. With the presence of uncertainty in a task, individuals may lack sufficient capability to verify the correctness of AI advice and therefore rely more heavily on the AI systems [17], leading to reduced appropriate reliance on AI systems. Previous research has also demonstrated the influence of uncertainty on trust formation in AI systems [401]. Considering highly complex and prognostic tasks, we hypothesize that individuals exhibit higher levels of trust and reliance on AI systems while showing a decrease in appropriate reliance. This could be due to the high cost of engaging cognitively in complex decision-making processes, leading to a greater reliance on AI systems for guidance [408]. Therefore, we formulate our hypotheses as shown in Table 4.1.

Table 4.1: **Summary of Our Hypotheses.**

Hypothesis	Description
H1a	Users demonstrate a lower level of appropriate reliance on AI systems for complex tasks compared to relatively less complex tasks.
H1b	Users trust AI systems to a greater extent in complex tasks compared to relatively less complex tasks.
H2a	Users demonstrate a lower level of appropriate reliance on AI systems in tasks with high levels of uncertainty compared to tasks with low levels of uncertainty.
H2b	Users trust AI systems to a greater extent in tasks with a high degree of uncertainty (`prognostic`) compared to tasks with lower levels of uncertainty (`diagnostic`).
H3	Users demonstrate a relatively low level of appropriate reliance on AI systems in tasks with relatively high complexity and uncertainty.

4.3.2. Trip-Planning Task

We chose trip-planning to as the scenario for our study due to two primary reasons. Firstly, trip-planning is a common real-world problem that individuals frequently encounter and seek assistance from AI systems to make decisions. Secondly, this task allows us to meaningfully manipulate complexity levels (e.g., the number of constraints) and uncertainty levels in our experimental conditions, thereby enhancing the ecological validity of our findings. In our study, participants are presented with a practical scenario where external assistance is potentially useful to successfully accomplish the task. We utilized an imperfect AI system with a 66.7% accuracy rate for trip-planning and manipulated its features accordingly (cf. section 4.4.1). This setup with the necessary complexity creates the desired sense of vulnerability and uncertainty, making it a suitable situation for analyzing human trust and reliance on AI systems [207, 253]. Note that while trip planning is a frequently encountered real-world task, the inclusion of time and budget limitations makes it unique, affecting how individuals rely on AI assistance. We also employed the DecisionTime framework 3 to design the experiments described in this section.



Alice, a student volunteer, played an important role in organizing an international Biology conference. She is assigned the task of assisting attendees throughout the event. Today is April 1. The conference will be held two weeks later on April 15 at 07:35 am in Conference Center. As part of her responsibility, she is expected to arrive half an hour before the start time for preparations. She also has a limited budget, 35 euros, to commute for her inner-city trips via public transport. Considering that she is at home at 05:45 am and planning to arrive at the venue on time. What is the best plan that Alice could have followed in terms of minimizing commute time and cost? How could you have best helped her schedule her commute among the highlighted routes, given the following information? You need to identify which types of public transport (train/bus/taxi) she should have taken respectively and at what stations she might have needed to switch among lines or different types of public transport. In the next step, an AI system called TripPlanner suggests the best schedule using one of the highlighted routes. Finally, you finalize your decision in the last step. Note that this AI system is not always accurate. On average, there is one error in every three recommendations that the AI system makes (i.e., the accuracy of the AI system is about 67%).

Start time: 5:23

Angie Leonard
Limited Seats
Sub-Route Base Time : 0.3 - 0.4 minutes
Sub-Route Base Cost: 1.02 €

Lillianna Butler
No Seats
Sub-Route Base Time : 0.6 - 0.8 minutes
Sub-Route Base Cost: 0.13 €

Kenley Sullivan
No Seats
Sub-Route Base Time : 3.6 - 3.6 minutes
Sub-Route Base Cost: 0.93 €

Kash Odonnell
No Seats
Sub-Route Base Time : 21.7 - 25.1 minutes
Sub-Route Base Cost: 67.9 €

Elias Vega
No Seats
Sub-Route Base Time : 18.2 - 21.5 minutes
Sub-Route Base Cost: 3.73 €

General information

* There is a very high chance of rain. Rain slows down buses and taxis, leading to longer commute times by around 40%.
* There is a very low chance that streets will become congested, indicated by the orange and red colors on the map representing moderate and heavy traffic jams.
* There is a high possibility of catching the intended bus or train, due to seat availability of buses and trains.
* Peak hours fall within these ranges: 6:30 - 9 AM, 4 - 6:30 PM, and 8 - 10:30 PM.
* If the commuter travels at least 10 km by train or 4 km by bus, a discount based on the kilometers traveled will be applied to her fare.
* The number of transfers between transport modes is limited to 4.
* It is lowly probable that the commuter might have to stop for something on her route, which could lead her to cross at the indicated point on the map.

*** The streets depicted on a map may not align with actual travel routes.
*** Each of the features related to tasks is established without any influence on one another.
*** There aren't any preferences toward particular transportation modes or routes.

Cost Information

Public Transport Fare per km	Subscription Discount per km
Train: 0.12 €	Train: 21%
Taxi: 2.5 €	Taxi: 10%
Bus: 0.19 €	

Peak-Hour Fare Rise per km	Kilometer-Based Discount per km
Train: 67%	Train: 17% above 10 km trip
Taxi: 40%	Bus: 16% above 4 km trip

Initial Decision

You initially selected Route 2 as the best plan.

AI Suggestion

AI suggests that the Route 5 minimizes the commute time and cost.

Final decision

What is the best plan in terms of minimising commute time and cost?

Figure 4.1: **An overview of the trip-planning task interface that participants used including five components: (1) task scenario and description, (2) map, (3) route information, (4) general information, and (5) two-stage decision-making. Note that this screenshot is meant to convey a bird's-eye view of the interface. This interface is also dedicated to a highly complex scenario encompassing all constraints and the prognostic experimental condition with high uncertainty.**

Planning a trip involves determining the most suitable route for travel, taking into account factors such as time limitations and budget constraints. Participants are tasked to select the trip that minimizes both travel time and expenses. Each task typically consists of multiple components that support participants in making well-informed decisions, as depicted in a bird's-eye view of the task interface in Figure 4.1.

Quality Control: To ensure the accuracy and reliability of the collected data in our study, we employed multiple methods. We initially offered instructional materials on the interface and task-related features, followed by a training session for participants that included both theoretical instruction and hands-on practice. Secondly, we evaluated participants' comprehension by administering a quiz on task-related constraints. Individuals who scored below a certain threshold were excluded from the study to maintain the quality of data. Lastly, we incorporated four attention-check questions in the pre-questionnaire and post-questionnaire to screen out individuals who may not be fully engaged or attentive throughout the study. Detailed explanations of these methods are publicly available on our companion page[1].

Design Considerations and Setups: Task Complexity vs. Task Uncertainty

Wood's seminal work [435] proposed that task complexity consists of three constructs: component, coordinative, and dynamic complexities. *Component complexity* relates to the number of features in a task, while *coordinative complexity* pertains to executing sequences or steps within the task. *Dynamic complexity* arises from changing world states requiring further considerations at the point of decision-making. We utilized component complexity to define task complexity and also adjusted the uncertainty as incomplete information in our setup. In dynamically complex tasks, decision-making must adapt as the situation changes, with all information accessible at each point. However, uncertain tasks involve incomplete information at the point of decision-making, setting them apart from dynamically complex tasks. Therefore, it is valid to consider these factors as separate dimensions although task uncertainty can increase task complexity.

Task Complexity

To operationalize task complexity in our experimental conditions, we manipulated the number of constraints that are given to participants. This approach has been used in previous studies to control the level of complexity for a given task [26, 334, 399]. We categorized the tasks into three levels of complexity: `low`, `medium`, and `high`. In low-complexity tasks, participants are presented with **four** features to consider while in medium-complexity tasks, **eight** features are provided. High-complexity tasks entail **twelve** different features that must be taken into account. This design choice is guided by prior neuroscience research by [298], suggesting that human cognitive capacity for processing information is limited to around seven (± two) chunks of information at a time. Hence, we established five to nine task features as representative of a medium level of complexity based on this finding. Any number exceeding nine would classify as high complexity, while four or fewer would indicate low complexity [354].

Task Uncertainty

Diagnostic tasks entail circumstances where participants are given access to well-defined and comprehensive information about the current task, allowing them to make precise judgments [17]. **Prognostic** tasks, on the other hand, involve scenarios in which participants are presented with restricted or unclear data and need to generate predictions regarding future outcomes [17]. The necessity to anticipate uncertain results gives

[1] https://osf.io/kt8m4/?view_only=c6930ba990c8412cb3948c2cf2b0a39c

rise to increased uncertainty throughout the process of making decisions. To operationalize uncertainty in the contrasting experimental conditions pertaining to diagnostic and prognostic tasks, we employed various strategies.

For diagnostic tasks, participants are instructed to schedule a trip for the present moment within the narrative, while for prognostic tasks, participants are assigned to plan a trip that will take place two weeks later. Next, we customized the way task attributes are presented to align with the level of uncertainty. In situations involving diagnostic tasks, participants are given precise values for each constraint, eliminating any potential ambiguity. On the other hand, in prognostic tasks, a certain degree of uncertainty is introduced by offering participants ranges or estimates instead of exact values for each attribute. We also presented the probability of different outcomes for certain constraints. For example, we highlighted the high likelihood of encountering traffic congestion during the rush hour or the low chance of experiencing rain during the scheduled trip.

We created one task scenario for each task. In total, we generated 24 different scenarios, with four scenarios in each experimental condition that differed in terms of task complexity and uncertainty. The **full list of these task scenarios** and **all code for our implementation** is publicly accessible for the benefit of the research community and in the spirit of open science[1].

TASK FEATURES

We designed task features to impart and define constraints in the decision-making tasks such that they do not affect each other and can be independently manipulated and measured. We communicated this independence explicitly and implicitly by ensuring that each feature is presented separately and does not rely on or interact with other features. All task features were inspired by considerations typical in real-world trip-planning contexts. In our research, we can classify task characteristics from two different viewpoints: each feature has the potential to influence either the overall duration of travel, the associated expenses, or both factors. Furthermore, each feature can be categorized as being either time-dependent or time-independent. Time-dependent features, such as traffic conditions and weather patterns, are prone to temporal changes based on external factors and their presentation differs when considering diagnostic tasks versus prognostic tasks. In tasks that have low complexity, we designed an equal distribution of time-dependent and time-independent features. However, for tasks with medium or high complexity, we increase the number of time-dependent features to enhance the degree of uncertainty that need to be considered in decision-making processes. Detailed explanations of all features are publicly available on our companion page[1].

4.4. STUDY DESIGN

4.4.1. EXPERIMENTAL CONDITIONS

Our study was approved by our institutional ethics board. We designed a between-subject study with a 3×2 factorial design. The three levels for task complexity were categorized as `low`, `medium`, and `high`, while the two distinct levels for uncertainty were `diagnostic` and `prognostic` tasks. We refer to these conditions as LowDiag, LowProg, MedDiag, MedProg, HighDiag, and HighProg. Participants were randomly assigned to one of the six experimental conditions while ensuring a balanced distribution of participants across the different task complexity and uncertainty levels. In each condition, participants were presented with three different task instances to complete with the assistance of an AI system. The three task instances were determined based on each condition's assigned complexity and uncertainty levels. Detailed explanations regarding the complexity and uncertainty levels are provided in section 4.3.2.

We fine-tuned the AI system to suggest routes that satisfy the given criteria with an accuracy of 66.7% across all experimental conditions. This level of accuracy was chosen since it is helpful if the system is relied on but still involves some risks. Hence, it calls for appropriate reliance instead of blindly following the AI system's advice. This design choice is motivated by prior work emphasizing the role of uncertainty in dictating the need to facilitate appropriate reliance [253]. This implies that within each batch of three task instances that a participant completes, to control for potential ordering effects, we ensure that incorrect advice is offered by the AI system once at random.

4.4.2. Measures

We leveraged a set of objective metrics to quantify participants' reliance on the AI system (cf. Table 4.2) [207, 292, 297, 361, 450, 460]. These metrics include Agreement Fraction, Switch Fraction [182, 449, 463], and Accuracy with Disagreement [182], Relative Positive AI Reliance, and Relative Positive Self-Reliance [359]. These parameters are commonly adapted in literature to capture the level of reliance within the human-AI interaction context. In addition to these measures of reliance, we also evaluated participants' decision-making accuracy, demonstrating the human-AI team performance [27, 348]. By measuring trust and reliance variables alongside human-AI team performance, we can gain a deeper understanding of whether performance outcomes result from under-reliance, appropriate reliance, or over-reliance on AI systems.

The subjective trust in the AI system was assessed using the Trust in Automation questionnaire (TiA) [232], which is a commonly employed and validated tool for measuring trust [262, 364, 399]. The questionnaire comprises multiple items that evaluate various aspects such as participants' perceptions regarding Reliability/Competence (TiA-R/c), Understanding/Predictability (TiA-U/P), Familiarity (TiA-Familiarity), Intention of Developers (TiA-IoD), the Propensity to Trust (TiA-PtT), and the overall level of trust placed in the AI system, Trust in Automation (TiA-Trust).

We collected information about participants' perceived numeracy skills as well as their affinity for technology in the pre-task questionnaire. To measure numeracy skills, we employed the Subjective Numeracy Scale [132], which is a self-report measure of perceived ability to perform various mathematical tasks and preference for the use of numerical information. Additionally, we administered the Affinity for Technology Interaction Scale (ATI) [143] to determine participants' level of comfort and familiarity with technology [399].

4.4.3. Participants

We first estimated the required sample size using G*Power software, considering a medium effect size of 0.25, a power of 0.90, and a significance level of 0.05, leading to a recommended minimum sample size of 210 participants, *i.e.*, 35 participants in each

Table 4.2: **An overview of the different metrics that we considered in our user study.**

Metric Type	Metric Name	Value Type	Value Range
Performance	Accuracy	Continuous	[0,1]
Reliance	Switch Fraction	Continuous	[0,1]
	Agreement Fraction	Continuous	[0,1]
Appropriate Reliance [182, 359]	Accuracy-wid	Continuous	[0,1]
	RAIR	Continuous	[0,1]
	RSR	Continuous	[0,1]
Trust	TiA-ReliabilityCompetence	Likert	5-point, strong distrust to strong trust
	TiA-UnderstandingPredictability	Likert	5-point, strong distrust to strong trust
	TiA-Intention of Developers	Likert	5-point, strong distrust to strong trust
	TiA-Trust in Automation	Likert	5-point, strong distrust to strong trust
Covariates	Subjective Numeracy (SNS)	Likert	6-point: from low to high
	Affinity for Technology (ATI)	Likert	6-point: low to high
	TiA-Familiarity	Likert	5-point, strong distrust to strong trust
	TiA-Propensity to Trust (TiA-PtT)	Likert	5-point, strong distrust to strong trust

of our experimental condition. To obtain a sufficient sample for our study while accounting for potential exclusion, we enlisted the participation of 285 individuals using the Prolific crowdsourcing platform. To ensure the reliability of the data gathered, we applied inclusion criteria that were designed to select native English speakers with a minimum approval rate of 95% on the platform and at least 100 completed studies. A total of 27 participants who failed any attention-check questions or the quiz were excluded from participation in the study, resulting in a final sample size of 258 participants. On average, participants took approximately 25 minutes to complete the entire study. All participants were compensated at the fixed rate of 8 GBP per hour regardless of their performance in the study. Additionally, participants received bonus rewards amounting to 0.2 GBP for each accurate response they provided during the study period. Overall, participants earned an average of 8.44 GBP per hour, well over the wage considered to be '*good*' and recommended by the Prolific platform.

4.4.4. PROCEDURE

The entire workflow of the study is illustrated in Figure 4.2. When participants entered the study, they were first provided with informed consent, a brief overview of the study's goals, and instructions on how to complete the tasks (step 1). If they consented to participate, they were directed to the pre-task questionnaire in step 2, where they were presented with a series of questions related to their numeracy skills and affinity for technology. Participants were then randomly assigned to one of the six different experimental conditions. According to the assigned condition, participants were presented with an interface tutorial and task tutorial that provided step-by-step instructions on how to navigate and complete the task followed by a training session on a sample task. The participants were given sufficient time to familiarize themselves with the sample task and the interface. To ensure the understanding of the task, participants were required to answer a quiz related to the task features before proceeding to the main task. If participants did not pass the quiz, they were excluded from the study. Otherwise, they received immediate feedback on their quiz performance to ensure that participants proceeded to the main task with a complete understanding of the task and devoid of familiarity or comprehension-related biases. Participants were then asked to complete three trip-planning tasks. Each task instance consisted of a decision-making scenario, where participants had to analyze the information provided and make an AI-assisted decision. Lastly, participants were directed to fill out a post-task questionnaire to assess their perception of the task features and trust in the AI system.

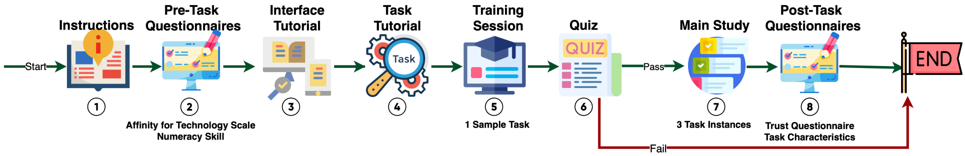


Figure 4.2: **Illustration of the procedure participants followed within our study.**

4.5. Results

4.5.1. Descriptive Statistics

The resulting sample of 258 participants had an average age of 38 years old ($SD = 11.8$) and consisted of 39% females and 61% males. To account for potential confounding variables, we gathered information about the participants' subjective numeracy skill (SNS), affinity for technology (ATI), TiA-Familiarity, and TiA-Propensity to Trust (TiA-PtT). Participants reported a moderate level of perceived numeracy ($M = 4.28, SD = 0.80$) on the 6-point scale. Similarly, participants were found to have a moderate affinity for technology interaction ($M = 4.04, SD = 0.56$) measured on a 6-point scale, low familiarity ($M = 2.87, SD = 1.17$), and a moderate propensity to trust AI ($M = 3.72, SD = 0.49$) measured on a 5-point scale.

4.5.2. Hypothesis Tests

H1a. Impact of task complexity on appropriate reliance

To explore the main effect of complexity on appropriate reliance, we conducted a Kruskal–Wallis test, Table 4.3. Subsequently, we conducted Dunn's post-hoc test to determine which levels of complexity resulted in significant differences in appropriate reliance. We reported *adjusted p-values*, calculated using Bonferroni correction to account for the increased likelihood of falsely declaring statistical significance when conducting multiple tests. If the *adjusted p-value* for an individual hypothesis is less than the significance level (0.05), then the null hypothesis is rejected, indicating a statistically significant result [441]. We first report the influence of complexity on reliance, followed by our examination of appropriate reliance.

The observed significant difference in switch fraction between high and low-

Table 4.3: **Kruskall-Wallis test for the main effect of task complexity on reliance. † indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column.**

Dependent Variable	*adjusted-p*	$M \pm SD$ (Low)	$M \pm SD$ (Medium)	$M \pm SD$ (High)	Post-hoc Results
Agreement Fraction	.8	0.62 ± 0.20	0.54 ± 0.25	0.55 ± 0.28	-
Switch Fraction	**.003**†	0.18 ± 0.32	0.26 ± 0.30	0.34 ± 0.36	Low < Medium < High
Accuracy	**<.001**†	0.79 ± 0.22	0.58 ± 0.27	0.61 ± 0.29	Low > Medium, High
Accuracy-wid	**.001**†	0.61 ± 0.40	0.40 ± 0.37	0.50 ± 0.35	Low > Medium, High
RAIR	**.001**†	0.22 ± 0.41	0.33 ± 0.41	0.43 ± 0.45	Low < Medium, High
RSR	**<.001**†	0.64 ± 0.48	0.34 ± 0.48	0.38 ± 0.49	Low > Medium, High

complexity tasks implies that task complexity does indeed exert an influence on **reliance**. In tasks with higher complexity levels, individuals tend to shift from relying on their own judgment to relying on the AI system. This can be attributed to a decrease in self-confidence regarding their decision-making abilities and, as a result, seeking guidance from the AI system.

Tasks of higher complexity tend to diminish the **appropriate reliance** on the AI system. Participants demonstrated significantly lower levels of *Accuracy-wid* in tasks with greater complexity compared to those with lower complexity. A similar trend is observed when examining *RSR*, wherein participants displayed significantly reduced levels of confidence in themselves during tasks with higher complexity than those with lower complexity. Consistent with these findings, participants exhibited a contrasting trend in displaying a significantly higher level of reliance on the AI system for tasks that were more complex compared to those of lower complexity, as indicated by higher *RAIR*. The rise in RAIR does not necessarily imply a higher appropriate reliance on the AI system. Rather, it suggests that individuals under-rely on the AI system in tasks with relatively lower complexity, and over-rely on the AI system in tasks with relatively higher complexity without being able to recognize when the advice may be inaccurate. This excessive reliance can ultimately have a negative impact on performance by reducing appropriate reliance levels.

Furthermore, we found that the **accuracy** of participants is significantly lower in tasks with higher levels of complexity than those with lower complexity. This finding provides additional evidence to our previous findings regarding the influence of task complexity on appropriate reliance. Overall, these results **partially support** our hypothesis **H1a**.

H1b. Impact of task complexity on trust

We aimed to examine the main effect of task complexity on trust in the AI system. Therefore, we conducted a two-way ANCOVA to consider the potential confounding effects of the covariates, namely subjective numeracy skill, affinity for technology, TiA-Familiarity, and TiA-Propensity to Trust. We did not find a significant effect of task complexity on human trust in the AI system, leading us to **reject** our hypothesis **H1b**. However, this finding supports that the subjective nature of trust in the AI system does not always follow the objective measure of reliance on the AI system [297, 362].

H2a. Impact of task uncertainty on appropriate reliance

We investigated the main effect of task uncertainty on reliance by conducting the Kruskal–Wallis test, reported in Table 4.4. We found that task uncertainty significantly affects participants' **reliance** on the AI system. Participants showed significantly higher levels of switch fraction when faced with prognostic tasks, indicating their tendency to rely more on the AI system due to lower self-confidence. Our findings further suggest that individuals can accurately assess the level of uncertainty in a task and adjust their reliance on the AI system accordingly.

Table 4.4: **Kruskall-Wallis test for the main effect of task uncertainty on reliance. † indicates the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column.**

Dependent Variable	*adjusted-p*	$M \pm SD$ (Diagnostic)	$M \pm SD$ (Prognostic)	Post-hoc Results
Agreement Fraction	**.01**†	0.60 ± 0.23	0.54 ± 0.26	Diagnostic > Prognostic
Switch Fraction	**.02**†	0.22 ± 0.32	0.31 ± 0.34	Diagnostic < Prognostic
Accuracy	**<.001**†	0.72 ± 0.30	0.60 ± 0.24	Diagnostic > Prognostic
Accuracy-wid	**.04**†	0.56 ± 0.43	0.45 ± 0.33	Diagnostic > Prognostic
RAIR	**.02**†	0.27 ± 0.42	0.38 ± 0.44	Diagnostic < Prognostic
RSR	.1	0.50 ± 0.50	0.40 ± 0.49	-

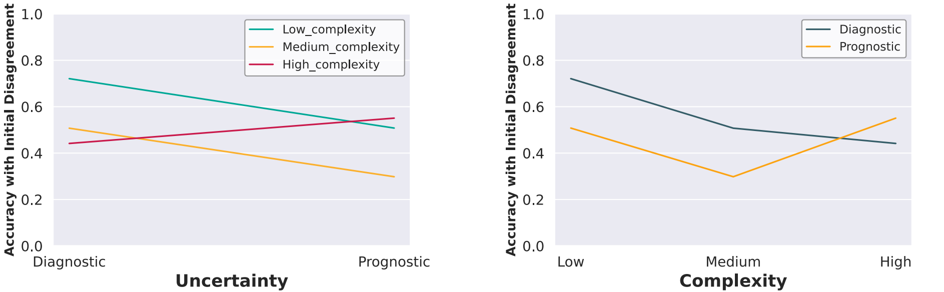
Furthermore, our findings revealed that the degree of uncertainty in a task significantly influenced participants' **appropriate reliance** on the AI system. We found that participants were more likely to appropriately rely on the AI system in diagnostic tasks, leading to higher accuracy rates, as indicated by higher *Accuracy-wid* compared to prognostic tasks. In line with this finding, we also observed that participants exhibited a slightly higher level of reliance on their own decision-making skills (*RSR*) when faced with diagnostic tasks. On the other hand, in prognostic tasks, participants showed significantly higher degree of reliance on the AI system as indicated by higher *RAIR*. This finding suggests that participants tend to rely heavily on the AI system in uncertain situations. However, this does not necessarily lead to appropriate reliance. It can be challenging for them to distinguish between accurate and inaccurate AI advice in prognostic tasks, resulting in lower appropriate reliance on the AI system and decreased accuracy levels. As a result, our findings **partially support** the hypothesis **H2a**.

H2b. Impact of task uncertainty on trust

The main effect of task uncertainty on trust in the AI system was also examined in this study through the ANCOVA test. The results indicated that there was no significant main effect of task uncertainty on any trust subscales. These findings indicate that participants' trust in the AI system remains relatively stable regardless of the level of uncertainty in the task. Thus, we **reject** our hypothesis **H2b**.

H3. Interaction effect of task complexity and uncertainty

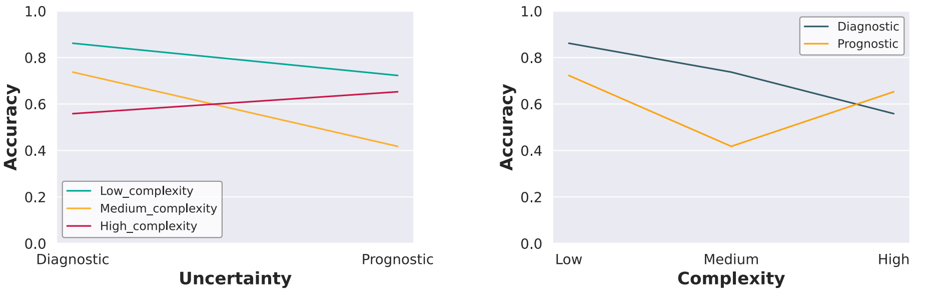
We conducted an ANOVA to investigate the interaction effect of task complexity and uncertainty on appropriate reliance and trust. We found a significant interaction effect between task complexity and uncertainty on *Accuracy-wid* as a measure of appropriate reliance. Figure 4.3a illustrates the interaction effect of task complexity and uncertainty on Accuracy-wid, focusing on different levels of complexity. We observed that the



(a) **Different task complexity levels across uncertainty levels** (b) **Different task uncertainty level across complexity levels**

Figure 4.3: **Interaction effects between task complexity and uncertainty on the *Accuracy-wid* metric reflecting appropriate reliance.**

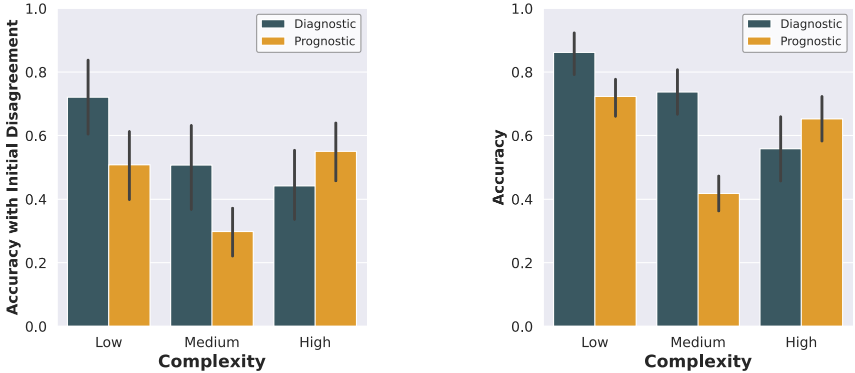
trend of Accuracy-wid is descending for tasks with low and medium complexity while increasing the level of uncertainty. However, for tasks with high complexity, the trend is the opposite, where Accuracy-wid increases with increasing uncertainty. Although we found earlier that participants have a lower *Accuracy-wid* for prognostic tasks, the interaction effect suggests that the impact of uncertainty on appropriate reliance depends on the level of task complexity. This finding suggests that participants tend to engage more cognitively in tasks they perceive as less complex, believing they can make accurate judgments. This trend is also observed in diagnostic tasks with high complexity. However, when faced with highly complex and prognostic tasks, participants are more likely to relinquish some cognitive control and rely heavily on the AI system. This could be attributed to their perception of the task's complexity exceeding their own capabilities. Participants may also view the AI advice as being more reliable and trustworthy, resulting in increased agreement and appropriate reliance. This finding is further supported by the significant interaction effect identified in *Accuracy*, Figure 4.4a, demonstrating that participants' ability to make accurate predictions increases when they are faced with prognostic tasks with high complexity, compared to prognostic tasks with medium and low complexity. Consequently, their level of accuracy aligns with that of the AI system due to their increased appropriate reliance. Figures 4.5a and 4.5b illustrate the Accuracy and Accuracy-wid for different levels of task complexity and uncertainty.



(a) **Different task complexity levels across uncertainty levels** (b) **Different task uncertainty levels across complexity levels**

Figure 4.4: **Interaction effect between complexity and uncertainty on *Accuracy* metric.**

We can observe the interaction effect of complexity and uncertainty for diagnostic



(a) **Mean of Accuracy-wid** (b) **Mean of Accuracy**

Figure 4.5: **Mean of Accuracy-wid and Accuracy across different levels of task complexity and uncertainty.**

and prognostic tasks in Figure 4.3b. For diagnostic tasks, the trend *Accuracy-wid* is descending as the complexity of the task increases. However, for prognostic tasks, different effects are observed. Participants tend to have lower *Accuracy-wid* as we increase the complexity from low to medium. In medium-complexity tasks, *Accuracy-wid* reaches its local minimum. So, as we further increase the complexity to high levels, *Accuracy-wid* starts to rise again, suggesting that participants rely more appropriately on the AI system, and their accuracy improves in highly complex prognostic tasks, aligning more closely with accuracy of the AI system (cf. Figure 4.4b). Furthermore, we can see that the appropriate reliance is always greater for diagnostic tasks compared to prognostic tasks, except for high complexity, where the values for prognostic tasks surpass those for diagnostic tasks, further supporting our findings. In summary, we found that the interaction effect between complexity and uncertainty in conditions with high complexity and uncertainty plays a significant role in human-AI decision-making. While the appropriate reliance drops as the complexity and uncertainty of a task increase, there is a turning point where participants start to rely more appropriately on the AI system, resulting in increased accuracy in prognostic tasks with high complexity. Thus, our findings **reject** hypothesis **H3**.

4.6. Discussion

4.6.1. Key Findings

Our study examined the impact of task complexity and uncertainty on human-AI decision-making. The results of our study demonstrated that increasing the level of **complexity** and **uncertainty** in decision-making tasks led to significant differences in **users' reliance** on the AI system. In more complex and uncertain tasks, we found that users were often in initial disagreement with the advice provided by the AI system. However, they demonstrated a heavy reliance on AI advice during the second stage of the decision-making process, leading to higher *Switch Fraction*. This can be attributed to the potential recognition that AI offers valuable insights for decision-making under complexity and uncertainty, coupled with a lack of confidence in their own judgment, corroborating what has been uncovered by other work in human-AI decision-making [75, 334]. Furthermore, the greater cognitive effort linked to complex tasks may also be a contributing factor. The cost of relying on the AI system would prove to be less compared to evaluating the reliability of the AI advice, thereby prompting individuals to lean towards following AI advice [408]. Additionally, users showed higher engagement and information-gathering behavior in prognostic scenarios, demonstrated by significantly more clicks on *route control buttons*, indicating greater inclination to explore different route options.

We also found that the **appropriate reliance** on the AI system varied significantly depending on **task complexity** and **uncertainty**. Users exhibited lower appropriate reliance on the AI system (lower *Accuracy-wid*), leading to lower accuracy in tasks with `medium` complexity or uncertainty compared to those with `low`. However, users demonstrated higher appropriate reliance on the AI system, resulting in improved accuracy in the experimental conditions with tasks with `high` complexity or uncertainty compared to those with `medium` complexity or uncertainty. Users perceived that tasks with higher complexity and uncertainty required greater effort and information processing, making them more willing to rely on the AI system. In such scenarios, their performance approaches AI accuracy, indicating the effectiveness of integrating AI in decision-making.

Our findings showed that individuals generally place significantly more **reliance** on the AI system when faced with tasks characterized by **high uncertainty**. However, in such prognostic tasks, their ability to **appropriately rely** on AI advice is lower compared to diagnostic tasks, subsequently affecting their overall performance. Tasks that involve inherent uncertainty are often those where humans tend to rely on AI systems for advice, such as loan approval [39, 84, 119], recidivism prediction [110, 165, 283], house price estimation [4, 34, 78], and student admission [34, 77]. Individuals may be more inclined to adhere to AI advice in these types of tasks. This could stem from the belief that AI systems possess advanced analytical abilities and have access to a greater amount of data [258]. On the other hand, when individuals are faced with tasks that have lower uncertainty, such as annotation and classification task [6, 260, 375], they tend to rely less on the AI advice and rely more on their expertise and judgment. Since the heavy reliance on AI systems in uncertain situations does not always lead to improved decision-making accuracy, several mechanisms have been proposed to optimize the combination of human and AI decisions to achieve the best outcomes and facilitate appropriate reliance on the AI system. These mechanisms include providing interpretable explanations for AI advice [65, 256, 404], using cognitive forcing functions[54, 165, 323], and incorporating feedback loops to enhance the interaction between humans and AI systems [27, 28, 450]. Despite implementing a two-stage decision-making process to encourage individuals to be cognitively involved in the procedure, as well as incorporating visual and textual explanations for increased transparency, our research emphasizes the necessity for additional exploration into strategies that can facilitate appropriate reliance on AI systems in contexts characterized by high levels of uncertainty.

The **complexity** of tasks plays a significant role in determining the degree of **reliance** on AI advice, consistent with the findings of [26, 334]. The more complex a task is, the more individuals may be inclined to rely on the AI system. We use the number of features or constraints as the measure of task complexity similar to previous studies [26, 334, 399]. Tasks with a larger number of constraints that need to be accounted for in decision-making are often more challenging for individuals to process, making them more likely to seek guidance from AI [232, 297, 358]. Our findings, which were based on objective measures, align with [408] study and suggest that users tend to rely more heavily on AI systems when faced with complex tasks that demand higher cognitive effort. This is further backed by [324] indicating that the complexity of a task can elevate its perceived difficulty, potentially resulting in greater reliance on AI systems. As shown by [354], the majority of tasks that have been studied in the context of decision-making are characterized by `low` and `medium` complexity. Prior studies that investigated tasks exceeding individual information processing capabilities (i.e., 9 constraints [298]) suggested employing visualization techniques to assist individuals in understanding the AI advice and the underlying decision-making process [156, 445, 457]. We used visual and textual techniques to support individuals in understanding the factors playing a role in shaping the given AI advice. However, in higher complexity scenarios, an individual still lacks cognitive engagement with the AI system and may be more likely to rely heavily on its advice. This is supported by the tendency of individuals to rapidly make their decision within approximately twenty seconds after receiving advice from AI, without carefully reassessing the provided information or exploring alternative route options. Although these visual and textual strategies have shown promise in improving decision-making outcomes in literature, they were not sufficient to mitigate over-reliance on AI advice in high complexity tasks.

According to the Trustworthiness Assessment Model (TrAM) [358], accurate perceived trustworthiness of AI systems is essential for establishing meaningful trust and reliance on AI systems. Factors such as relevance and availability of system information, as well as the ability of individuals to detect and utilize this information, play a crucial role in determining accurate perceived trustworthiness. In our study, we only presented relevant task features using visual and textual formats to participants. We utilized user behavior metrics and validation of participant perceptions through training and quizzes to ensure the detection of these features. However, we expected the complexity and uncertainty of tasks to impact the availability and utilization of system information, thus affecting perceived trustworthiness [321]. However, participant trust remained consistent regardless of task complexity or uncertainty, which was in contrast to what is suggested by the TrAM framework.

4.6.2. Implications of Our Work

Implications for Methodology and the HCI Community

The implications of methodology in HCI research pertain to the design and analysis of studies [37]. These implications specifically address data collection methods and the construction of new knowledge. Our work has important implications for the methods used to study human-AI decision-making, for increasing the external validity of empirical work and strengthening the understanding of the transferability of findings across different studies. It has been observed that task characteristics, such as complexity and uncertainty, are seldom examined or analyzed systematically in human-AI decision-making studies. While it may not be experimentally feasible to account for every facet of a task, our research emphasizes the significance of considering these factors when assessing human-AI collaboration. Future research should consider the incorporation of methodologies that take into account task-related features when evaluating human-AI decision-making. Our findings also contribute to the interpretation of human behaviour and reliance on AI systems through the lens of task complexity and uncertainty. Current studies often focus on generic decision-making scenarios or tasks with low to medium complexity, which may not fully reflect or represent the challenges and dynamics of the full range of real-world scenarios. This is particularly important in highly complex tasks coupled with high uncertainty, where humans tend to require, appreciate, and rely on advice from an AI system. Future research should consider the systematic identification and inclusion of task-specific characteristics in the design of studies in the realm of human-AI decision-making.

To initiate a systematic evaluation of task characteristics, we propose the lens of diagnostic and prognostic tasks as a framework for modeling uncertainty in decision-making, which can be used as a basis for designing experiments and gathering data on human-AI interactions. This approach acknowledges the inherent uncertainty in determining or estimating different constraints that influence decision outcomes. Additionally, it offers a relatively more precise representation of decision-makers' challenges. Incorporating this lens into research methodology would involve designing studies that specifically control the uncertainty inherent in diagnostic and prognostic tasks and exploring their impact on human-AI decision-making processes and outcomes. We also encourage researchers to consider highly complex tasks in their experiments to capture the challenges and nuances of decision-making in real-world scenarios. This can be achieved by developing scenarios or simulations that closely resemble complex decision-making situations in different domains. Our task details and all code for the interface are made publicly available to support future research in the community[1].

Our study also highlights the need for further examination and development of techniques tailored specifically to support high-complexity and prognostic tasks in human-AI decision-making. Although many interventions have been developed for decision-making in various domains, there is still a need to focus on the unique challenges posed by high complexity and prognostic tasks. Such interventions could be targeted to offer users indicators that can help them accurately assess the reliability, plausibility, and verifiability of the AI advice. Consequently, these methods will promote appropriate reliance on the AI system in complex and uncertain decision-making scenarios. There is a heightened urgency in developing and creating these mechanisms to prevent potential deception arising from the complexity and uncertainty of tasks, which can make it challenging to detect untrustworthy AI systems [25]. By reducing the cost of verifiability and plausibility of such XAI techniques, decision-makers can gain a better understanding of the basis for AI advice based on their own expertise and judgment, potentially leading to improved performance and appropriate utilization of AI systems.

The decline in performance of human-AI teams when tackling tasks of medium complexity suggests that users may have faced challenges in accurately assessing their own abilities and the capabilities of AI systems, primarily by overestimating their own abilities [235]. This aligns with previous research findings, highlighting the need for interventions to assist users in evaluating their skills and appropriately adjusting their reliance on AI systems [78, 182, 243]. This may be particularly important in tasks with relatively moderate complexity which may lead to illusory self-assessments among some users, compared to tasks with evidently low or discernibly high complexity.

Implications for Theory

Theoretical implications focus on the understanding of task characteristics and their impact on human-AI decision-making. Based on our findings, it is evident that the complexity and uncertainty of tasks significantly influence how humans rely on AI systems. This study serves as the application-grounded evaluation [116] in the context of trip-planning, centering on the individuals the system intends to support in actual tasks. It empirically validates the commonly held belief that task complexity and uncertainty play a crucial role in determining human reliance on AI systems. While the primary objective of combining humans and AI is to achieve enhanced performance through collaboration, an over-reliance on AI can potentially impede the advantages offered by human judgment and decision-making abilities. Therefore, it is crucial for researchers to develop theoretical frameworks that can help identify and motivate the optimal balance between human and AI involvement in decision-making, taking into consideration task complexity and uncertainty.

Contrary to previous research suggesting that trust in AI systems increases with the complexity and uncertainty of tasks, our findings indicate that trust is orthogonal to these factors. These results suggest that trust is not the sole determinant of reliance on AI advice, and other factors such as task characteristics play a significant role. This also indicates the difference between human trustworthy beliefs and behavior toward AI systems, where trust may not always translate into increased reliance, highlighting the need to measure, calibrate, and understand factors beyond trust that influence human-AI decision-making.

4.6.3. Caveats and Limitations

According to the checklist of cognitive biases provided by [117], it is important to acknowledge that humans are prone to cognitive biases. In our task, we identify the familiarity bias and availability heuristic, which can cause individuals to exhibit an inclination towards decisions that align with their pre-existing beliefs or past experiences. Although we created artificial routes, individuals may still tend to prefer familiar or known options or prefer specific transport modes due to personal biases. Confirmation bias and overconfidence bias are other potential limitations, as individuals may be more likely to seek out and give more weight to information that confirms their preconceived notions or beliefs regarding AI capabilities and their decision-making abilities. We should also consider the self-interest bias, where individuals may prioritize their own monetary reward over objective decision-making criteria.

The findings discussed in this chapter are not universally applicable to all decision-making tasks. Different tasks may have varying characteristics and contexts that can influence human-AI decision-making. Although this is a valid approach to operationalize uncertainty, it is important to acknowledge that there could be other approaches to capturing task uncertainty that were not explored in this study (e.g., missing data or conflicting information). Future research should consider exploring different operationalizations of task complexity and uncertainty to further understand their impact on human reliance on AI systems. It is worth noting that we asked participants in our study to consider that the traffic features were unrelated to each other and carried equal weights in determining the best route. This may not always be the case in real-world contexts. We also considered traffic conditions in both diagnostic and prognostic scenarios, although, in the real world, traffic conditions can change over time and at the time of decision-making, making them predominantly prognostic.

4.7. CONCLUSION AND FUTURE WORK

In this study, we explored how task complexity (**RQ1**) and task uncertainty (**RQ2**) and their interaction (**RQ3**) inform user trust and appropriate reliance on AI systems. To this end, we conducted a user study with 258 participants across six experimental conditions varying in three levels of task complexity (`low`, `medium`, and `high`) and two levels of task uncertainty (`diagnostic` and `prognostic`). We selected trip-planning as the decision-making task and evaluated participants' trust, reliance, and decision-making behaviors when interacting with an AI system. The study showed that task complexity and uncertainty significantly impact human reliance on AI systems. Participants tended to rely more on AI in tasks with higher complexity and uncertainty, with no significant differences in human trust across different levels of complexity and uncertainty.

Future studies should further explore the relationship between task complexity and uncertainty to better understand their interconnections in human-AI decision-making. Further research is needed across a range of domains and task types to fully understand the impact of task complexity and uncertainty. We encourage researchers to investigate the impact of other task characteristics, such as time pressure and information overload, on human-AI decision-making. Future work should also focus on understanding how to effectively present AI-generated predictions and explanations to enhance human understanding and decision-making, particularly in complex and uncertain situations. Given the increasing complexity and uncertainty of tasks, it becomes crucial to develop strategies that can help users evaluate the reliability and verifiability of AI advice in these scenarios.

II

Adding Groups to the Mix: Human-AI Group Decision-Making

5

When in Doubt! Understanding the Role of Task Characteristics on Peer Decision-Making with AI Assistance

With the integration of AI systems into our daily lives, human-AI collaboration has become increasingly prevalent. Prior work in this realm has primarily explored the effectiveness and performance of individual human and AI systems in collaborative tasks. While much of decision-making occurs within human peers and groups in the real world, there is a limited understanding of how they collaborate with AI systems. One of the key predictors of human-AI collaboration is the characteristics of the task at hand. Understanding the influence of task characteristics on human-AI collaboration is crucial for enhancing team performance and developing effective strategies for collaboration. Addressing a research and empirical gap, we seek to explore how the features of a task impact decision-making within human-AI group settings. In a 2×2 *between-subjects study (*$N = 256$*) we examine the effects of task complexity and uncertainty on group performance and behaviour. The participants were grouped into pairs and assigned to one of four experimental conditions characterized by varying degrees of complexity and uncertainty. We found that high task complexity and high task uncertainty can negatively impact the performance of human-AI groups, leading to decreased group accuracy and increased disagreement with the AI system. We found that higher task complexity led to higher efficiency in decision-making,*

This chapter is based on a peer-reviewed paper:**Sara Salimzadeh** and Ujwal Gadiraju. 2024. When in Doubt! Understanding the Role of Task Characteristics on Peer Decision-Making with AI Assistance. In Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 01–04, 2024, Cagliari, Italy. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3627043.3659567

while a higher task uncertainty had a negative impact on efficiency. Our findings highlight the importance of considering task characteristics when designing human-AI collaborative systems, as well as the future design of empirical studies exploring human-AI collaboration.

5.1. Introduction

The increasing capabilities of AI systems to perform tasks with high accuracy have led to increasing interest in incorporating these systems into human decision-making processes across various fields, such as finance [84, 128, 156, 160], healthcare [139, 227, 241, 320], and the legal domain [12, 267, 283, 425]. The main goal of such collaboration is to leverage the complementary strengths of humans and the AI systems to improve overall performance [27, 224, 381, 420]. Human-AI collaboration is also crucial for mitigating potential issues that may arise from relying solely on AI systems [244, 247, 255, 275, 375]. Empirical research in the HCI community has investigated factors that affect human-AI collaborative decision-making. This includes exploring the impact of human expertise [260, 313, 356], the level of human trust and reliance on the AI systems [55, 130, 408, 425], and the context of decision-making tasks [11, 155, 258].

Numerous studies have focused on group recommendation systems and AI support for individual decision-making. However, there is still a gap in the research concerning how AI can assist in group decision-making processes, specifically regarding task characteristics and their impact on the decision-making process [80, 208, 287, 465]. For instance, in healthcare, AI systems can assist multidisciplinary teams of doctors in diagnosing and planning treatment for patients, while in group trip-planning scenarios, individuals may rely on AI advice to make itinerary decisions. The dynamics of group decision-making can be complex, with various social and cognitive factors influencing the process and outcomes which need to be carefully considered when designing human-AI collaborative systems [19, 141, 276, 302, 465]. It is important to understand how human-AI collaboration can be fostered effectively in group decision-making settings, where multiple individuals interact with one or several AI systems to make joint decisions. Understanding these aspects can also offer insights into designing AI systems and interventions to promote effective collaboration among group members with AI systems, enhancing the overall outcomes.

In this thesis, we aim to explore the potential of human-AI collaboration in group decision-making by investigating **the role of task characteristics** on group dynamics and outcomes. Task features are the predictor factors that could impact group decision-making performance [5, 402, 403]. While existing research has examined the role of task characteristics within individual human-AI decision-making realm [26, 155, 258, 408], there is a limited understanding of how these factors influence human-AI group decision-making processes. In our work, we specifically examine the influence of **task complexity** and **task uncertainty** on the performance and interaction between human peers and AI systems. These elements have been recognized as crucial factors in determining the effectiveness of group decision-making processes [211, 403, 412]. Prior studies have also shown that people tend to need a group to collectively make a decision when faced with complex and uncertain decision-making scenarios [203]. Task complexity is defined by the amount of information that needs to be processed due to task features, such as the number of variables, interdependencies, and decision constraints [435]. On the other hand, task uncertainty pertains to the degree of unpredictability linked with the outcome of a task [17]. To the best of our knowledge, this is the first study that explicitly investigates the role of task characteristics in human-AI collaboration within a group decision-making context. We thereby address the following research questions in our study:

RQ1: How does task complexity influence user behaviour and performance in AI-assisted collaborative decision-making?

RQ2: How does task uncertainty influence user behaviour and performance in AI-assisted collaborative decision-making?

To address these research questions, we selected the real-world context of group trip-planning as our study domain. This complex decision-making scenario is characterized by numerous variables and elements of uncertainty, requiring peers to identify the most efficient route from a set of options by relying on an imperfect AI system or exercising their group judgments. We conducted a 2 × 2 between-subjects study with 256 participants randomly assigned to one of the four experimental conditions, manipulating task complexity (*high* vs. *low*) and task uncertainty (*low* vs. *high*). The complexity levels were determined by considering the different number of constraints in the task, while task uncertainty was altered by giving participants precise value of task constraints as opposed to probabilities or likelihood estimates. For instance, in low uncertainty conditions, participants would know the specific values of traffic conditions and weather forecasts, whereas in high uncertainty conditions, participants would only be provided with probabilities or a potential range of values corresponding to these variables. Note that when at least two individuals collaborate with an AI system, it can be characterized as a group decision-making process, as at least three distinct entities are involved: the two human participants and the AI assistant. In this chapter, we refer to the two human participants as the pair or peer, and the entire arrangement is considered a group decision-making scenario.

We found that task complexity and task uncertainty significantly influence user behaviour and performance when collaborating with an AI system in a group setting. Performance of groups in the high complexity or uncertainty conditions was significantly lower compared to the low complexity or uncertainty conditions. Moreover, incorporating of AI advice for final decisions resulted in increased performance compared to the initial decisions across all conditions, specifically in high complexity tasks. This performance gain is not attributed to the higher agreement with the AI advice but rather to the ability of participants to integrate the AI advice with their own judgment and resulting in more informed decisions. Interestingly, participants demonstrated a higher level of efficiency in tasks with high complexity, while task uncertainty was detrimental to group efficiency as it led to longer discussion times after receiving AI advice.

Original contributions: Our study contributes to the understanding of how task complexity and uncertainty impact group trip-planning when collaborating with an AI system. The context of group planning serves as an application-grounded evaluation [115], focusing on the individuals that such systems are designed to assist in making real-life decisions. To the best of our knowledge, this is the first study to explore the combined effects of task complexity and uncertainty on user behaviour and performance in a group setting with AI collaboration. Our study also provides empirical evidence that integrating AI advice with human groups can enhance performance and efficiency, particularly in high complexity tasks. Our work highlights the importance of considering task complexity and uncertainty when designing AI systems for group collaboration, and has important implications for the UMAP community.

5.2. Related Work

Human-AI Decision-Making

With the increasing performance of AI systems, there has been growing interest in understanding how humans can effectively collaborate with AI systems in decision-making tasks [58, 127]. The main goal of such collaboration is to leverage the complementary strengths of humans and AI systems to improve overall performance, exceeding what either humans or AI systems could achieve alone [27, 224, 381, 420]. However, reaching such complementary performance is not always achievable [247, 274] due to various factors including human cognitive biases [121, 181, 281, 314, 326], the degree of trust and reliance on AI systems [130, 267, 293, 325], and human understanding of the boundaries within which AI systems can make errors [26, 140, 215, 391, 463]. The context of decision-making tasks has been found to have a substantial influence on the extent to which AI systems are trusted, consequently affecting overall performance [11, 155, 258, 347]. Each domain has its own unique characteristics and requirements, which might not be transferable to other domains, highlighting the need for domain-specific studies on human-AI decision-making [247, 354]. Decision-making tasks can have varying levels of risk and stake across different domains, which can influence user behaviour, especially in high-stake situations where vulnerability and potential consequences are significant [12, 162, 207, 255]. The significance of creating suitable tasks in studies to arrive at valid findings has been emphasized by researchers [58, 247, 354]. For example, in proxy tasks that require users to predict AI advice, user behaviour, and performance might vary from tasks where users make decisions directly based on AI advice [58]. The level of complexity in tasks can also affect the performance of human-AI teams [26, 75, 334], with individuals tending to rely more on the AI systems for complex tasks that demand specialized knowledge or extensive analysis [258]. Task complexity may be gauged by factors such as the number of constraints involved [26, 334, 400] or the depth of mathematical calculations and analysis needed [155]. [408] also manipulated task difficulty levels by adjusting the cognitive effort needed to complete the task, thereby expanding the decision space. In addition to the complexity of the tasks, the uncertainty associated with task constraints can also impact human behaviour and reliance on the AI systems [17, 401]. In this study, we investigate how group performance is affected by task complexity and uncertainty in a collaborative decision-making scenario involving two individuals and an AI system.

Group Decision-Making

Group decision-making involves a collective process of reaching a decision within a group consisting of two or more individuals with their own perceptions and personalities, all accessing the same information to address a shared problem [60]. Research in group decision-making has shown that the group dynamics and interaction within a group can significantly influence the decision outcomes [189, 233, 383]. Although many studies suggest that group decision-making can result in better outcomes than individual decision-making [23, 32, 308], there are also factors that can hinder effective group decision-making [276, 302], such as group-think [31, 432], social loafing [185], and conformity biases [134, 402]. To achieve effective group decision-making, it is essential to understand the factors that influence the performance and outcomes of groups. Group performance refers to the collective ability of a group to achieve its goals and objectives [106], arising from their interactions, coordination, and cooperation rather than simply being the sum of individual capabilities [20]. Prior studies have investigated the factors that affect the performance of group decision-making [30, 48, 59, 102, 103, 197]. Composition of the group has been identified to be an important determinant of the group performance [192, 193, 373, 439]. A few studies have also focused on the role of task characteristics in influencing group performance and behaviour. [5] found that group efficiency surpasses the highest-scoring and most efficient members of nominal groups, similarly sized collection of individuals working independently, when dealing with complex tasks as opposed to relatively simple ones. They also observed that both individuals and groups have lower performance in complex tasks compared to simple tasks. [402] also explored the impact of task uncertainty and group size on the group performance, finding that higher levels of uncertainty with larger group sizes can negatively impact decision-making outcomes. In our research, we create a group of two individuals with an AI system to investigate how task complexity and uncertainty affect the collective intelligence and decision-making performance of the group.

Group Decision-Making with AI Assistance

Extensive research has been conducted on individual decision-making with AI systems and group recommender systems [208, 287]. However, there remains a gap in the literature regarding how AI can adequately facilitate group decision-making processes, particularly when considering distinct characteristics of tasks and their impact on the decision-making process [67, 94, 287, 398]. [20] highlighted the factors that contribute to successful human-AI group decision-making, such as the cognitive processes, algorithms, and psychological constructs that can provide a framework to model and understand the dynamics of human-AI team decision-making. [19] also found that individual expertise and cognitive biases play a crucial role in shaping social influence and decision-making dynamics within a human-AI group. [465] explored the equal power of AI in a group decision-making process, where AI systems have an equal say in the final decision. [228] also found that the collective intelligence, a factor measures group ability to perform together on a range of task, is one of the key predictor variables of group performance, specifically in complex tasks. [80] compared group and individuals along six aspects, including decision accuracy and confidence, reliance on the AI system, understanding AI system, fairness, and accountability in the recidivism risk assessment task. They found that groups over-rely more on AI systems compared to individuals but their performance may not necessarily be superior. In this study, we aim to delve deeper into the impact of task complexity and uncertainty on the interactions and performance of groups of two individuals with an AI system.

5.3. Hypotheses and Task Design

5.3.1. Hypotheses

The complexity of tasks have been identified as a key factor in affecting the performance of groups in general, either as humans-only groups [5, 402] or individual human-AI groups [26, 75, 334]. Based on prior studies, we hypothesize that as the complexity of tasks increases, the performance of groups of humans with an AI system would be negatively affected. As tasks become more complex, the likelihood of interpersonal conflict among group members may rise [382], resulting in sub-optimal outcomes and reduced performance [5]. Cognitive biases like social loafing [226] and group-think [210] are more likely to be noticeable in complex tasks, which can further hinder group performance. On the other hand, integrating input from each individual and the AI system may require more time, resulting in prolonged decision-making processes and a potentially reduced overall efficiency in decision-making.

(H1a.) Groups exhibit a lower performance in complex tasks compared to relatively less complex tasks.

(H1b.) Groups spend more time for decision-making in complex tasks compared to relatively less complex tasks.

We hypothesize that the presence of uncertainty in tasks could further exacerbate the negative impact on group performance. Uncertainty can introduce further challenges for human peers in reaching a consensus and making timely effective decisions. It may also lead to increased reliance on AI systems [401], potentially impeding group coordination.

(H2a.) Groups exhibit a lower performance in tasks with high uncertainty compared to tasks with relatively lower uncertainty.

(H2b.) Groups spend more time in tasks with high uncertainty compared to tasks with relatively lower uncertainty.

5.3.2. Task Scenario

In our study, we devised a trip-planning scenario in which participants were tasked with identifying the most efficient route that minimizes both time and budget from a selection of ten possible routes. Participants worked in pairs and were presented with practical situations in which they could receive guidance from an AI system on the optimal route or make decisions based solely on their own judgments as a team. We also employed the DecisionTime framework 3 to design the experiments described in this section.

The context of group trip-planning serves as an application-grounded evaluation [115], focusing on the individuals that such systems are designed to assist in making real-life decisions. We chose trip-planning as our task scenario to test our hypotheses for several reasons: participants are familiar with the concept of the task, allowing us to simulate a realistic scenario. Nevertheless, including time and budget constraints makes this task unique in affecting participants' behaviour and decision-making process. Furthermore, a trip-planning scenario allows the meaningful manipulation of task complexity and uncertainty, thus enhancing the ecological validity of our findings. We incorporated an imperfect AI system to evoke the intended feeling of uncertainty and vulnerability, prompting deliberate collaboration and exchange of information among team members to validate the accuracy of the AI advice rather than relying solely on it.

Task Complexity

Inspired from prior work [26, 334, 400], we manipulate the complexity of the tasks by varying the number of constraints that participants have to consider when planning their trip. We curated two levels of task complexity: **low complexity** and **high complexity**. In the low complexity condition, participants only have to consider four constraints when planning their trip (e.g., length of the route, transportation, travel time, and transportation fare). In the high complexity condition, participants have to consider eight constraints when planning their trip (e.g., length of the route, transportation, travel time, transportation fare, weather conditions, traffic jam, seating capacity of the transportation, and ticket subscription). We determined the number of constraints per condition according to an individual's information processing ability [298], suggesting that individuals can efficiently process between five and nine variables simultaneously.

Task Uncertainty

Task uncertainty pertains to the level of unpredictability associated with the given task. This can be influenced by various factors such as the amount and reliability of information available, the likelihood of unexpected events occurring, and the level of variability in task conditions. We operationalized task uncertainty by manipulating the amount and reliability of information available in two levels: **low uncertainty** and **high uncertainty**. In the low uncertainty condition, participants are provided with accurate and reliable information about all constraints involved in planning their trip. In the high uncertainty condition, participants are given a range of possible values or probabilities for certain constraints, reflecting the inherent unpredictability and variability in real-world conditions. Such constraints include weather conditions, traffic jam, availability of transportation options, and their seating capacity, as these constraints can change over time and are subject to various unforeseen events.

Design Consideration: Task Complexity vs. Task Uncertainty

Wood's seminal work suggests that task complexity could be devised into three dimensions: *component*, *coordinative*, and *dynamic* complexity [435]. Component complexity refers to the number of distinct constraints that need to be considered in a task, while coordinative complexity relates to the number of steps required to complete the task and the interdependencies between those steps. Dynamic complexity, on the other hand, arises when the world states change requiring to potentially adjust decisions based on the changing conditions. In our work, we operationalized task complexity as component complexity taking into account the number of constraints involved in planning a trip [26, 334, 400]. Task uncertainty pertains to the level of unpredictability associated with the task, reflecting the missing information and likelihood of unexpected events. By definition, dynamic complexity and uncertainty are two distinct constructs; in dynamically complex tasks, all the information are accessible and can be considered for decision-making at each point, while in uncertain tasks, decision-making is inherently challenged by the lack of complete and reliable information. Although these two constructs could interact and influence each other, it is valid to consider that task complexity and task uncertainty as separate constructs that can independently influence decision-making processes.

5.4. Study Design

5.4.1. Experimental Conditions

Our study was approved by our institutional ethics board. We designed a between-subject study with 2 × 2 factorial design. The experimental conditions included two independent variables: task complexity (high vs. low) and task uncertainty (high vs. low). We randomly assigned participants to one of four experimental conditions and balanced the number of participants in each condition across the different task complexity and uncertainty levels. Participants in each scenario were given three distinct task instances, according to the complexity and level of uncertainty associated with the specific condition. In groups of two individuals with an AI system, participants were required to make decisions based on the task constraints, while also considering the AI advice.

AI System

Our AI system was developed to consider factors like distance, traffic, weather conditions, and time and budget limitations to offer the most efficient route from the possible ten options. We fine-tuned the AI system at a 66.7% accuracy rate, such that the AI system provided incorrect advice in one out of the three tasks. This design choice aimed to encourage participants to critically evaluate the AI system's advice while still benefiting from its guidance.

Four-stage Decision-Making

To engage each member in the decision-making process, we implemented a four-stage procedure [203]. The first stage includes recording individual initial decisions in isolation to prevent bias or influence from others [177, 333]. In the second stage, initial decisions are shared in a chat box and openly discussed. The AI system then suggests the best route based on the available information in the third stage. The peers then engage in a collaborative discussion during the fourth stage to reach a consensus on the final decision.

Ice-breaking

We included an ice-breaking activity at the beginning of the collaboration to provide an opportunity for peers to build common ground and enhance their communication throughout the trip-planning task [7, 16, 212, 428]. To this end, we suggested a number of questions that encouraged participants to share personal experiences, interests, and goals, inspired by [218]. Peers are also allowed to share any concerns or questions they may have, further fostering open communication and collaboration within the group.

5.4.2. MEASURES

Measures: Our measures aim to assess the effectiveness of collaboration within the group, individual decision-making skills, and the impact of the AI system on the decision-making process across all conditions. These measures included objective performance criteria such as the accuracy of the final decision made by the group (*Group Accuracy_Final*), the accuracy of individual initial decisions (*Individual Accuracy_Initial*), and the peer agreement with AI advice (AI Agreement). Additionally, we recorded the time taken to reach an individual initial decision (*Average Individual Decision Time_Initial*), the time taken to reach a consensus for the final decision (*Average Group AI Consideration Time_Final*), and the total task completion time (*Average Decision Time*). To get insight into participants' engagement and exploration of solution space, we also tracked participants' interaction with the map and exploration of alternative routes. Furthermore, we calculated group efficiency (*Group Efficiency_Final*) and individual efficiency (*Individual Efficiency_Initial*), defined as the performance divided by the duration. To assess the potential confounding effect of participants' mathematical skills on their task performance, we incorporated a pre-task questionnaire to evaluate their perceived numerical skills [132]. Additionally, we administered the Affinity for Technology Scale questionnaire [143] to gauge participants' level of comfort and proficiency in using technology [400].

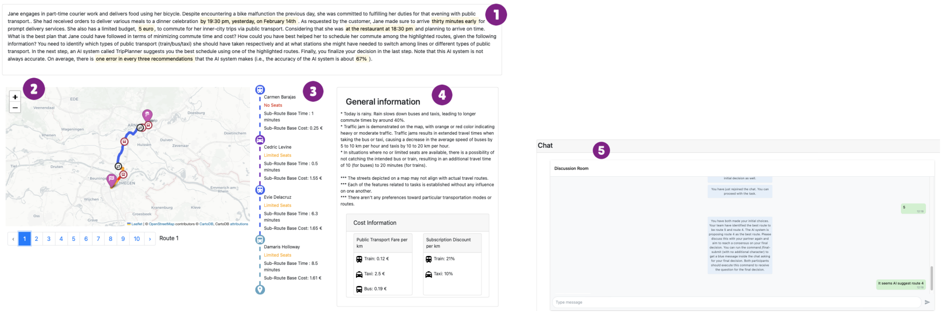
5.4.3. PARTICIPANTS

We calculated the minimum sample size required for this research by conducting a power analysis using G*Power software, resulting in 256 participants, i.e. 64 participants (32 pairs) per experimental condition with a medium effect size of 0.25, the significance level of 0.05, and the power of 0.80. We totally recruited 303 participants from the Prolific crowdsourcing platform, where 47 participants rejected due to our quality control criteria, resulting in a final sample size of 256 participants. On average, participants spent 45 minutes completing the entire study. All participants were compensated at the fixed rate of 9 GBP per hour regardless of their performance, as deemed *good* according to the standards set by the Prolific platform.

Quality Control: To ensure the quality of our study, we implemented various measures. First, we conducted a pilot test with a small sample of participants to identify any potential issues or ambiguities in our instructions and procedures. Second, we recruited native English participants who had at least 100 previous successful task completions in the Prolific crowd sourcing platform, with at least 95% approval rate ensuring a certain level of experience and reliability. Third, we provided detailed training to participants to ensure they had a clear understanding of the interface, the task, and how to collaborate within the group to submit their decisions. We also evaluated their understanding through a brief quiz at the end of the tutorial and practice session. Participants who did not pass the quiz were excluded from the study. Finally, we closely monitored the participants' progress throughout the study and were available for any clarifications or assistance they may have required.

5.4.4. Procedure

When participants entered the study, they were provided with a consent form that explained the procedure of the study. If they agreed to participate, they were randomly assigned to one of the four experimental conditions. In the first step, participants were given a pre-task questionnaire to help us capture their numeracy skills. On completing this questionnaire, participants were introduced to the interface and familiarized with its functionalities through a tutorial. They were then instructed on the details of the trip-planning task, including the specific goals and constraints involved. In the next stage, participants were presented with a sample scenario and were asked to make decisions based on the provided information to practice and apply their understanding of the task. To ensure participants' understanding and consistency across experimental conditions, participants were given a brief quiz at the end of the practice session. The quiz aimed to ensure participants' comprehension of the task and the impact of each constraint on the decision-making process. Participants who passed the quiz proceeded to the lobby, while those who did not were excluded from the study[1]. In the lobby, participants had to wait for a partner to be matched with. During this waiting period, participants were provided with optional fillers– a game and breathing exercises– to help them relax and stay engaged. Such activities were designed to keep participants engaged and reduce the perceived waiting time [2, 3]. Participants in the lobby had to wait to be matched with a partner. Upon the arrival of another participant under the same experimental conditions, they were grouped and given guidance on how to collaborate within the chat environment for the main tasks. Figure 5.1 displays a bird's-eye view of the interface without the chat component.



Jane engages in part-time courier work and delivers food using her bicycle. Despite encountering a bike malfunction the previous day, she was committed to fulfilling her duties for that evening with public transport.. She had received orders to deliver various meals to a dinner celebration by 19:30 pm, yesterday, on February 14th . As requested by the customer, Jane made sure to arrive thirty minutes early for prompt delivery services. She also has a limited budget, 5 euro, to commute for her inner-city trips via public transport. Considering that she was at the restaurant at 18:30 pm and planning to arrive on time. What is the best plan that Jane could have followed in terms of minimizing commute time and cost? How could you have best helped her to schedule her commute among the highlighted routes, given the following information? You need to identify which types of public transport (train/bus/taxi) she should have taken respectively and at what stations she might have needed to switch among lines or different types of public transport. In the next step, an AI system called TripPlanner suggests you the best schedule using one of the highlighted routes. Finally, you finalize your decision in the last step. Note that this AI system is not always accurate. On average, there is one error in every three recommendations that the AI system makes (i.e., the accuracy of the AI system is about 67%).

Carmen Barajas
No Seats
Sub-Route Base Time : 1 minutes
Sub-Route Base Cost: 0.25 €

Cedric Levine
Limited Seats
Sub-Route Base Time : 0.5 minutes
Sub-Route Base Cost: 1.55 €

Evie Delacruz
Limited Seats
Sub-Route Base Time : 6.3 minutes
Sub-Route Base Cost: 1.65 €

Damaris Holloway
Limited Seats
Sub-Route Base Time : 8.5 minutes
Sub-Route Base Cost: 1.61 €

General information

* Today is rainy. Rain slows down buses and taxis, leading to longer commute times by around 40%.
* Traffic jam is demonstrated on the map, with orange or red color indicating heavy or moderate traffic. Traffic jams results in extended travel times when taking the bus or taxi, causing a decrease in the average speed of buses by 5 to 10 km per hour and taxis by 10 to 20 km per hour.
* In situations where no or limited seats are available, there is a possibility of not catching the intended bus or train, resulting in an additional travel time of 10 (for buses) to 20 minutes (for trains).

*** The streets depicted on a map may not align with actual travel routes.
*** Each of the features related to tasks is established without any influence on one another.
*** There aren't any preferences toward particular transportation modes or routes.

Cost Information

Public Transport Fare per km	Subscription Discount per km
Train: 0.12 €	Train: 21%
Taxi: 2.5 €	Taxi: 10%
Bus: 0.19 €	

Chat

Discussion Room

initial decision as well.

You have just rejoined the chat. You can proceed with the task.

5
12:18

You have both made your initial choices. Your team have identified the best route to be route 5 and route 4. The AI system is proposing route 4 as the best route. Please discuss this with your partner again and aim to reach a consensus on your final decision. You can run the command /final-submit (with no additional character) to get a blue message inside the chat asking for your final decision. Both participants should execute this command to receive the question for the final decision.

It seems AI suggest route 4
12:18

Figure 5.1: **An overview of the trip-planning task interface that participants used including five components: (1) the task scenario and description, (2) map, (3) route information, (4) general information, and (5) chat box, located at the bottom of the interface. Note that this screenshot is meant to convey a bird's-eye view of the interface. This interface corresponds to a *highly complex task* scenario encompassing all constraints.**

During the collaboration phase, peers were provided with the same interface and could communicate with each other in a chat environment. They shared their opinions, received AI advice simultaneously, and submitted their decisions as text messages within the chat interface. At the beginning of their interaction, participants could use three minutes to communicate with their partners on the ice-breaker questions sug-

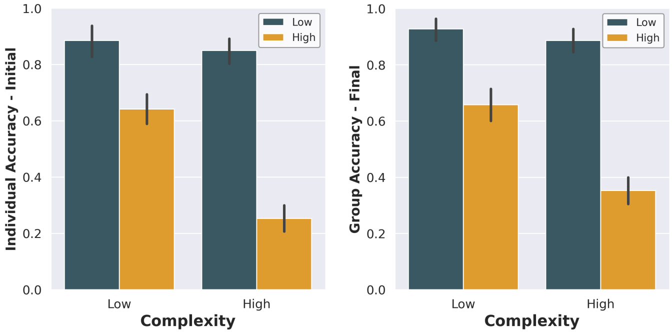
[1] All participants were compensated fairly irrespective of whether or not they passed the quiz.

gested within the chat environment. This aimed to initiate conversation and facilitate a comfortable common ground between the participants that could aid collaboration[2]. Participants were tasked to follow the four-stage decision-making process to submit their final decision for each trip-planning task. To ensure that participants adhere to the decision-making procedure, prompts were provided at each stage of their collaboration. While there were no time limits for each stage, these prompts served as a road-map for systematic progression. After completion of three consecutive tasks, participants were redirected to the post-task questionnaire where they were asked to provide feedback on their experience. All code for our implementation of the interface along with task scenarios and in-depth details of our user study are made publicly available to support future research in the community and in the spirit of Open Science[3].

5.5. RESULTS

5.5.1. DESCRIPTIVE STATISTICS

Our study involved a sample of 256 participants, with 59.7% being male and 40.3 % female. The average age of the participants was 34 years, ranging from 18 to 61 years. Participants' numeracy skills were also found to be moderately high ($M = 4.40, SD = 0.77$), with no significant differences or confounding effects observed across the four experimental conditions. Similarly, participants reported their affinity with technology to be moderately high ($M = 4.09, SD = 0.57$), indicating that they were familiar and comfortable using technological tools for communication and collaboration.



(a) Mean of *Individual Accuracy_Initial*

(b) Mean of *Group Accuracy_Final*

Figure 5.2: **Mean of *Individual Accuracy_Initial*and *Group Accuracy_Final*across different levels of task complexity and uncertainty. The accuracy improved after the peers received AI advice.**

5.5.2. HYPOTHESES TESTS

H1A. IMPACT OF TASK COMPLEXITY ON GROUP PERFORMANCE

To investigate the main effect of task complexity on group performance, we conducted a Kruskal-Wallis test, comparing the accuracy of groups across complexity levels (cf. Table

[2] Note that analyzing the role of relationship strengths in the context of peer decision-making with AI assistance is beyond the scope of this work.

[3] https://osf.io/kvt7p/?view_only=0d90e14a2eeb4ea8889000b409720987

5.1). We also assessed the accuracy of initial decisions submitted per individual in a group to evaluate if the AI advice provided after their individual choices had an effect on group performance. Similarly, we evaluated the level of agreement between the peers decisions and AI advice to gauge the impact of the AI system on decision-making.

The observed significant difference between low complexity tasks and high complexity tasks indicates that task complexity does impact group performance (*Group Accuracy_Final*). In the low complexity tasks, groups exhibited higher accuracy compared to high complexity tasks, suggesting that the level of task complexity has a negative impact on group performance. Furthermore, we observed that the performance of each individual in making initial decisions (*Individual Accuracy_Initial*), also significantly differed across complexity levels. Similarly, the initial performance are lower in high complexity tasks compared to low complexity tasks, indicating that the complexity of the tasks also affects individual performance. We found that the collective performance significantly improved (*Accuracy Gain*) when participants were given AI advice following their initial decision-making, in situations where the task complexity was high compared to low complexity tasks, demonstrated in Figures 5.2a and 5.2b. However, there was no significant difference between AI agreement across complexity levels. This suggests that AI advice and group discussions can positively impact group performance, particularly in high complexity tasks, regardless of the level of agreement between the peers decisions and AI advice. Overall, these results **support** our hypothesis **H1a**.

Table 5.1: **Kruskall-Wallis test for the main effect of task complexity on performance.** † **indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column.**

Dependent Variable	*adjusted-p*	*M* ± *SD* (Low)	*M* ± *SD* (High)	Post-hoc Results
Individual Accuracy_Initial	**<.001**†	0.76 ± 0.26	0.55 ± 0.35	Low > High
Group Accuracy_Final	**<.001**†	0.79 ± 0.24	0.62 ± 0.32	Low > High
Accuracy Gain	**0.017**†	0.03 ± 0.15	0.07 ± 0.17	Low < High
AI Agreement	0.06	0.56 ± 0.21	0.51 ± 0.25	-

H1b. Impact of task complexity on group behaviour

To examine the influence of task complexity on group behaviour, we conducted a Kruskal-Wallis test comparing the average time taken by groups to make decisions across complexity levels, reported in Table 5.2. We also considered the time taken by individual participants to make their initial decisions and the time taken for peers to reach a consensus after receiving AI advice. We also assessed the effectiveness of group decision-making by evaluating how quickly the peers reached correct decision during each minute of their decision-making process. Additionally, we investigated the individual efficiency in making initial decisions to compare their performance before and after receiving AI advice.

We found no significant difference in the average time taken by groups (*Average Decision Time*) to make decisions across complexity levels. The time taken by individual participants to make their initial decisions (*Average Individual Decision Time_Initial*) and the time taken for peers to reach a consensus after consideration of AI advice (*Average Group AI Consideration Time_Final*) also did not significantly differ across complexity levels. This indicates that task complexity does not have a significant impact on the speed at which groups make decisions or reach consensus. However, task complexity did significantly affect the group efficiency (*Group Efficiency_Final*) in terms of the performance per unit of decision-making time. In high complexity tasks, groups tended to be more efficient in reaching a correct decision compared to low complexity tasks. However, the impact of task complexity on individual decision-making efficiency (*Individual Efficiency_Initial*) was not significant, suggesting that task complexity primarily affects group dynamics and enhances the collective efficiency. We also recorded the user interaction with the interface including time spent navigating different routes on the map, the number of clicks on different routes, and the extent they explored the available features and options. We observed that participants exhibited similar exploration and utilization of the map and route options regardless of the complexity of the task at hand. The presence of additional features and information did not appear to prompt participants to explore or make extensive use of them. Thus, we **reject** our hypothesis **H1b**.

Table 5.2: **Kruskall-Wallis test for the main effect of task complexity on behaviour. † indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column. Note that Efficiency is the measure of performance over one minute. The times are also reported in seconds.**

Dependent Variable	*adjusted-p*	*M* ± *SD* (Low)	*M* ± *SD* (High)	Post-hoc Results
Individual Efficiency_Initial	0.13	0.30 ± 0.10	0.28 ± 0.11	-
Group Efficiency_Final	**<.001**†	0.03 ± 0.02	0.04 ± 0.02	Low < High
Average Decision Time	0.91	483 ± 171	477 ± 179	-
Average Individual Decision Time_Initial	0.59	389 ± 149	376 ± 137	-
Average Group AI Consideration Time_Final	0.31	94 ± 52	100 ± 77	-

H2A. IMPACT OF TASK UNCERTAINTY ON GROUP PERFORMANCE

We investigated the main effect of task uncertainty on group performance by conducting a Kruskal-Wallis test on a number of performance metrics including group accuracy, individual accuracy, and peer agreement with the AI system, Table 5.3.

Tasks uncertainty is found to have a significantly negative impact on group performance (*Group Accuracy_Final*). Participants in high uncertainty tasks showed lower accuracy in their decisions compared to groups performing low uncertainty tasks. Additionally, the level of task uncertainty also significantly affected the initial decision accuracy (*Individual Accuracy_Initial*), with individuals facing high uncertainty demonstrated lower initial decision accuracy compared to those facing low uncertainty. This suggests that task uncertainty hinders the overall performance and decision-making ability of groups, leading to lower accuracy in decisions either with or without the AI advice. The remarkable gap between group performance in high and low uncertainty tasks highlights the significant role that task uncertainty plays in influencing group performance and decision-making accuracy. We also found that task uncertainty significantly increased disagreement with AI advice, with high uncertainty tasks experiencing lower levels of AI agreement compared to low uncertainty tasks. The impact of AI advice on improving accuracy (Accuracy Gain) from the initial decision is not significantly influenced by task uncertainty, shown in Figures 5.2a and 5.2b. However, this suggests that presenting the AI advice may still have assisted peers in enhancing their decision-making process, even in high uncertainty tasks where agreement with the AI advice was lower. As a result, our findings **support** hypothesis **H2a**.

Table 5.3: **Kruskall-Wallis test for the main effect of task uncertainty on performance. † indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column.**

Dependent Variable	*adjusted-p*	$M \pm SD$ (Low)	$M \pm SD$ (High)	Post-hoc Results
Individual Accuracy_Initial	**<.001**†	0.87 ± 0.21	0.45 ± 0.28	Low > High
Group Accuracy_Final	**<.001**†	0.91 ± 0.16	0.51 ± 0.26	Low > High
Accuracy Gain	0.12	0.04 ± 0.16	0.06 ± 0.16	-
AI Agreement	**<.001**†	0.65 ± 0.14	0.43 ± 0.24	Low > High

H2b. Impact of task uncertainty on group behaviour

We examined the main effect of task uncertainty on group behaviour by analyzing the decision making time of groups as wells as their efficiency using Kruskal-Wallis tests, reported in Table 5.4.

Our findings illustrate that task uncertainty did not significantly impact the decision-making time of groups (*Average Decision Time*) and the time it takes to submit the initial decision (*Average Individual Decision Time_Initial*). However, the time taken to reach from initial decision to final decision (*Average Group AI Consideration Time_Final*) was significantly longer for groups facing high uncertainty compared to those facing low uncertainty tasks. This indicates that task uncertainty prolongs the decision-making process when presenting the AI advice, potentially leading to more thorough analysis and consideration of the uncertain tasks. Task uncertainty also had a significant impact on efficiency (*Group Efficiency_Final*), with groups facing high uncertainty demonstrating lower efficiency in their decision-making process compared to those in low uncertainty experimental conditions. However, the individual efficiency (*Individual Efficiency_Initial*), measured by the time taken to submit the initial decision, did not show a significant difference between high and low uncertainty tasks. These findings suggest that task uncertainty not only slows down the decision-making process but also reduces overall efficiency in groups. We also observed that participants exhibited similar exploration and utilization of the map and route options regardless of the uncertainty of the task at hand. One explanation for this could be that participants did not get extra information from the interface or map that could help reduce the uncertainty in high uncertainty tasks. As a result, they invested additional time with their partner engaging in discussion and exploring possible solutions after receiving AI advice, relying on their collective knowledge and perspectives to navigate the uncertain task and make informed decisions. These results **support** hypothesis **H2b**.

Table 5.4: **Kruskall-Wallis test for the main effect of task uncertainty on behaviour. † indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column. Note that Efficiency is the measure of performance over one minute. The times are also reported in seconds.**

Dependent Variable	*adjusted-p*	$M \pm SD$ (Low)	$M \pm SD$ (High)	Post-hoc Results
Individual Efficiency_Initial	0.17	0.30 ± 0.11	0.28 ± 0.10	-
Group Efficiency_Final	**<.001**†	0.04 ± 0.02	0.02 ± 0.02	Low > High
Average Decision Time	0.35	464 ± 150	500 ± 196	-
Average Individual Decision Time_Initial	0.92	373 ± 119	392 ± 163	-
Average Group AI Consideration Time_Final	**.01**†	91 ± 64	108 ± 66	Low < High

5.6. Discussion and Implications

5.6.1. Key Findings and Implications of Our Work

Our study investigated the impact of task complexity and uncertainty on group performance and behavior. We found that task complexity negatively affects the accuracy of decision-making in groups, with higher complexity leading to lower accuracy. This finding is consistent with prior studies that have shown the challenges posed by complex tasks which can lead to increased decision-making errors [5, 26, 75, 334]. The higher performance gain obtained with AI advice in complex tasks while maintaining a similar level of agreement with the AI system suggests that leveraging the AI system can help mitigate the negative impact of task complexity on group accuracy. One potential explanation for this finding is that the AI advice provides a point of discussion and reference for peers, helping them navigate through the challenges and potentially improving their decision-making process. Although the AI advice was not appropriately followed by peers in tasks with high complexity, it was utilized to facilitate more thorough analysis and consideration of the complex task, leading to improved decision-making outcomes.

We also did not find any significant influence of task complexity on the time taken to reach the consensus or the total time spent on decision-making. However, task complexity positively impacted the efficiency of the groups, with groups facing high-complexity tasks demonstrating a higher efficiency compared to those facing low-complexity tasks. One explanation for this contrasting finding could be that peers were able to utilize their collective insights and knowledge to make more efficient decisions in complex tasks, whereas in low-complexity tasks, the decision-making process may be more straightforward and less strategic and thorough. Another possible explanation could be that the higher complexity tasks prompted peers to allocate more resources towards thorough discussions and information gathering, leading to more informed and carefully considered decisions. We did not observe any significant difference in the efficiency of individuals in their initial decisions across tasks of varying complexity, suggesting that task complexity does not inherently impact individual efficiency. Although we expected that individuals facing high-complexity tasks may experience decreased efficiency, our findings suggest that the presence of a group and the opportunity for collaborative thinking and initial discussion can offset such a potential decrease in individual efficiency caused by task complexity.

Our study also revealed that task uncertainty plays a crucial role in group performance. The levels of task uncertainty were found to negatively influence the collective accuracy of the groups. Task uncertainty also significantly impacts the agreement with the AI advice, with higher levels of uncertainty leading to lower agreement. Nevertheless, the use of AI advice led to improved decision-making outcomes across all levels of task uncertainty. This finding highlights the potential of the AI system as a valuable tool in decision-making processes, especially in situations where task uncertainty is high. In such situations, the AI system could provide additional insights that can help mitigate the negative effects of task uncertainty and support more accurate decision-making.

We found that task uncertainty significantly impacts the time taken to reach a consensus after receiving the AI advice. This could be because high task uncertainty creates more divergent viewpoints within the group, requiring more time for discussion and consideration of various perspectives along with AI advice. Nevertheless, the total decision time and time taken to make the initial individual decision remained relatively consistent across different levels of task uncertainty. With the lower group accuracy and longer time reaching the consensus which is associated with higher task uncertainty, we observed that the group efficiency drops in high levels of task uncertainty compared to tasks with relatively low uncertainty. We did not observe any significant difference in the efficiency of individuals in their initial decisions across tasks of varying uncertainty, suggesting that the presence of a group and collaborative thinking before making an individual decision could help mitigate the negative impact of task uncertainty on efficiency.

Comparing the performance of the human-AI group with the accuracy of the AI system (0.66) revealed that the group was able to achieve higher accuracy in tasks with relatively low complexity (0.79) or uncertainty (0.91). This suggests that in situations with low task complexity or uncertainty, collective intelligence and collaboration within the group can lead to more accurate outcomes compared to relying solely on the AI system, achieving complementary performance. In high-complexity situations, however, the AI system's accuracy slightly surpassed that of the group (0.62), suggesting that the AI system may have an advantage in handling complex tasks. The AI system's accuracy exceeds that of the group in tasks with high levels of uncertainty (0.51), demonstrating its potential as a reliable and precise tool for making decisions. Therefore, peers should consider leveraging AI advice to augment their decision-making processes, particularly in situations with high levels of task uncertainty or complexity.

Overall, this study emphasizes the importance of considering task complexity and uncertainty in a human-AI decision-making context and highlights the potential benefits of incorporating AI advice to enhance accuracy and efficiency in group decision-making. Although AI advice has been shown to improve accuracy potentially through facilitating improved discussions, there is still a need for the design and development of more nuanced approaches and support systems to help groups fully leverage AI advice, especially in highly complex and uncertain scenarios.

5.6.2. Caveats and Limitations

Individuals and groups are prone to a range of biases that can impact the accuracy and effectiveness of their decision-making processes. In our task, we identify the familiarity bias that could lead individuals or groups to rely too heavily on their own experiences and knowledge, neglecting the valuable insights provided by the AI system or the partner. We also recognize the potential for groupthink, wherein group members may conform to a dominant opinion or suppress dissenting views to maintain harmony or consensus within the group. While we formed small groups and delegated tasks accordingly, it is crucial to account for the potential differences in outcomes when working with larger groups in future work. Furthermore, group communication in face-to-face settings may differ from chat communication, potentially influencing the decision-making process. The findings from this study should be interpreted with caution as they may not generalize to all decision-making contexts. Different contexts may have different levels of task uncertainty or complexity, and the dynamics within groups may vary. The use of AI systems with different attributes may also impact the decision-making process. Although we operationalized task complexity and task uncertainty in this study, further research is needed to explore the impact of other factors such as group size, group diversity, and communication dynamics on the accuracy of decision-making outcomes. Analyzing the impact of task characteristics on peer trust and reliance on AI systems was beyond the scope of this work. However, it could provide valuable insights into how group performance was affected by these factors. Additionally, our study did not explore the content and number of messages exchanged within the group, which may have influenced decision-making processes. Future research should aim to investigate these factors to gain deeper insights into the role of AI systems in human-AI group decision-making.

5.7. Conclusion and Future Work

In this study, we investigated the impact of task complexity **(RQ1)** and task uncertainty **(RQ2)** on the performance and behaviour of human-AI group decision-making. Each group consisted of two participants with an AI system who collaborated on three decision-making tasks in the context of trip planning. We conducted a user study with 256 participants to explore our research questions across four experimental conditions varying in level of complexity (high or low) and uncertainty (high or low). Our results revealed that task complexity and task uncertainty significantly influence the performance and dynamics of human-AI group decision-making. Specifically, we found that in tasks with high complexity or high uncertainty, group performance diminishes significantly compared to tasks with low complexity or low uncertainty. AI advice was also found to positively impact decision-making performance, but this effect was statistically significant in conditions of high complexity. This positive impact does not necessarily imply that groups always agree with AI advice, as individual and group factors may still influence the decision-making process. On the other hand, we have shown that task complexity and uncertainty can have varying effects on the efficiency of human-AI group decision-making. While higher task complexity tends to increase efficiency, higher task uncertainty can decrease efficiency and prolong the time needed to reach a consensus. Overall, this study highlights the importance of considering task complexity and uncertainty in human-AI group decision-making, and the need for tailored strategies and guidelines to optimize the integration of AI systems in group decision-making, especially in uncertain environments.

Future studies should further investigate the dynamics of trust and reliance in group decision-making with AI systems to gain a more comprehensive understanding of the factors that influence group behaviour and decision-making processes. Moreover, future research should explore how AI systems can be designed and utilized to mitigate the negative effects of task complexity and uncertainty, as well as develop strategies to enhance collaboration and communication between humans and AI systems to improve group decision-making outcomes. Additionally, it would be valuable to explore other factors that may influence human-AI group decision-making, such as different task contexts and features, group sizes, group diversity, group dynamics, and the impact of potential biases in the group setting.

III

Improving Information Access: The cases of Web Search and Databases

6

On the Impact of Entity Cards on Learning-Oriented Search Tasks

Entity cards are a common occurrence in today's web Search Engine Results Pages (SERPs). SERPs provide information on a complex information object in a structured manner. Typically, they combine data from several search verticals. They have been shown to: (i) increase users' engagement with the SERP; and (ii) improve decision making for certain types of searches (such as health searches). In this chapter, we investigate whether the benefits of showing entity cards also extend to the Search as Learning (SAL) domain. Do learners learn more when entity cards are present on the SERP? To answer this question, we designed a series of learning-oriented search tasks (with a minimum search time of 15 minutes), and conducted a crowdsourced Interactive Information Retrieval (IIR) user study ($N = 144$*) with four interface conditions: (i) a control with no entity cards; (ii) displaying relevant entity cards; (iii) displaying somewhat relevant entity cards; and (iv) displaying non-relevant entity cards. Our results show that (i) entity cards do not have an effect on participants' learning, but (ii) they do significantly impact participants' search behaviours across a range of dimensions (such as the dwell time and search session duration).*

6.1. Introduction

Learning is an important aspect of our lives. Thanks to the prevalence of the *World Wide Web (WWW)*, learning is today often achieved in an informal way, with *web search engines* acting as the information source. [285] defined these search episodes as a part of *exploratory search*. Known as *Search as Learning (SAL)* [88] today, SAL is an iterative

This chapter is based on a peer-reviewed paper: **Sara Salimzadeh**, David Maxwell, Claudia Hauff. 2021. On the Impact of Entity Cards on Learning-Oriented Search Tasks. In Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3471158.3472255.

process where the goal of the learner is to gain knowledge about their specific *information need*, or *learning objective*. A large body of research now encompasses the SAL domain [87, 89, 125, 145, 151, 223, 270, 271, 317, 389, 431, 433, 451], with most of these works concerning the analysis of query logs to attain insights into how those subjected to learning-oriented search tasks behave. Another prominent research direction is how to measure learning in a scalable manner, as only with cheap to compute metrics derived from observable search behaviours at scale will we be able to fulfil the vision of a search interface that adapts to a user's learning needs. In fact, the adaptation of the search system itself—either at the front-end or the back-end—have largely been left unexplored in the SAL domain. Exceptions to this are a small number of works that propose retrieval functions that surface documents suitable for learning [387–389], and works that designed and evaluated search engine result page widgets for learning purposes [69, 82, 352].

Modern web search engines do not provide interfaces that are explicitly designed for learning-oriented searches, though they have changed remarkably over their lifespan. Until a few years ago, the *ten blue links* paradigm dominated the look and feel of SERPs. Contrasted to contemporary SERPs, results are now shown from multiple modalities and search verticals. One prominent result type is the *entity card*. Each entity card (or *information card*) contains a summary of the *entity* (e.g., the name, description, associated images, and related entities)—and thus often helps users find information without the need to interact with other search results.

Although research into the usability and usefulness of entity cards is somewhat limited, several studies [50, 214, 309] have shown that entity cards can enhance the search experience in several ways. Entity cards provide concise content corresponding to the user query by merging information from various information sources, such as images, maps, Wikipedia, or social media [50]. They assist users in accomplishing their task [242, 309], and increase users' engagement with organic search results [50].

Despite these advantages however, entity cards have not been *evaluated* in the SAL context. To this end, *we investigate in this chapter whether entity cards are beneficial to users that undertake learning-oriented search tasks* in terms of the achieved learning outcomes. We conduct an *Interactive Information Retrieval (IIR)* study, and design four SERP variants: *(i)* the control condition which provides a standard SERP without an entity card (`No-EC`); *(ii)* a SERP with an entity card relevant to the query (`Good-EC`); *(iii)* a SERP with an entity card that is somewhat relevant to the query (`Fair-EC`); and *(iv)* a SERP with a non-relevant entity card (`Bad-EC`)[1]. We implemented these variants on top of the `SearchX` framework [335], and conducted a between-group study with $N = 144$ participants. Each participant was assigned to one of the four conditions to assess how different variants of entity cards impact human learning while searching. Concretely, our research questions are as follows.

RQ1: Does the inclusion of entity cards of various quality impact the amount of learning taking place during a learning-oriented search task?

[1] As a concrete example from our query log, for the query `radioactivity`, a good entity card is *radioactive decay*, a fair one is *radionuclide* and a poor one is *time*.

RQ2: Does the inclusion of entity cards of various quality impact users' search behaviours during a learning-oriented search task?

Our main findings can be summarised as follows. *(i)* The inclusion (or not) of an entity card has no discernible impact on participants' learning gains. *(ii)* In contrast, the quality of the entity card with respect to the query has a significant effect on participants' search behaviour across a range of dimensions (such as the dwell time and search session duration).

This is the first work to begin to shed light on the influence of entity cards on users' learning gain. Despite observed changes in search behaviour led to no positive changes in learning gain, these findings point to many open issues in terms of entity card design optimised for human learning.

6.2. Related Work

Entity Cards

Despite the fact that entity cards are ubiquitous in web search engines today, there is a limited amount of research published about them. Most research focuses on exploring the impact of entity cards on users' search behaviour. [309] undertook a user study to determine the impact a non-linear SERP layout has on eye and mouse movement behaviours. They were able to show that users spend more time on relevant entity cards than their non-relevant counterparts. When entity cards are relevant, they are beneficial to reduce the task completion time (at least sometimes). This is because the information need can be directly answered by the card's content. [**NLagun_2014**] interleaved entity cards within organic search results and carried out a user study in a mobile setting. In line with [309], they found that in the presence of non-relevant entity cards, users gloss over them. Upon not finding an answer, they continue to examine results below, leading to an increased amount of time spent further down the SERP. [50] explored how entity cards affect users' search behaviours and perceived workload. While they went in-depth into generating different types of entity cards (i.e., on-topic and off-topic), results generally showed that participants were more likely to interact (in terms of clicks and mouse hovers) with cards that are relevant to their information need. Furthermore, the presence of entity cards on search result pages increases the users' engagement with organic search result pages. Relevant entity cards also do not significantly increase users' workloads.

Apart from behavioural aspects, prior works have considered how to generate and present entity cards on SERPs. [178] examined the content of entity cards, introduced the task of *dynamic entity summarisation*, and proposed an approach to generate query-dependent entity summaries. Their user study found participants to favour dynamic summaries over static ones. Recently, [214] have shown in the *health IR* setting that when searching for information about a particular condition, users typically consider the entity card presented first—and then continue to the remainder of the SERP. In addition, they proposed an entity-focused SERP with *multiple* entity cards, and showed that the presence of relevant entity cards regardless of the interface type (i.e., one or multiple cards) leads to a higher probability of making correct decisions. Lastly, the SERP variant with multiple entity cards shown at once allowed participants to make health decisions with significantly less effort as measured by the number of clicks.

Keyphrase Extraction

In order to determine which entity card(s) to show for a given query (and without access to a large query log for training), we rely on the top retrieved documents for that query. As a first step, we need to extract the *keyphrases* from each of those documents. The task of keyphrase extraction can be defined as *"automatically selecting a small set of phrases that best describe a given free text document"* [35]. Here, we only focus on unsupervised methods, as they are most suitable for our user study due to their domain independence and no required training data. Unsupervised algorithms are divided into two primary groups: *(i) corpus-dependent approaches* [123, 126, 289, 336, 414, 421, 448] which rely on the entire corpus that the current document may be linked to; and *(ii) corpus-independent approaches* [35, 51, 72, 137, 230, 250, 251, 294, 319, 374, 414, 422], which rely on the current document only. Within the corpus-independent category, approaches follow different strategies such as: *(i) graph-based methods* [51, 137, 294, 374, 414, 422] which exploit graph-based language representations to detect keywords; *(ii) embedding-based approaches* [35, 230, 250, 251, 319]; *(iii) statistical-based methods* [72] which rely on statistical features of the text. For our work, we picked one graph-based [374], one embedding-based [35], and one statistical-based approach [72]—each reporting state-of-the-art effectiveness within their category. We picked the best model for our use case in a validation study, as described in the following section.

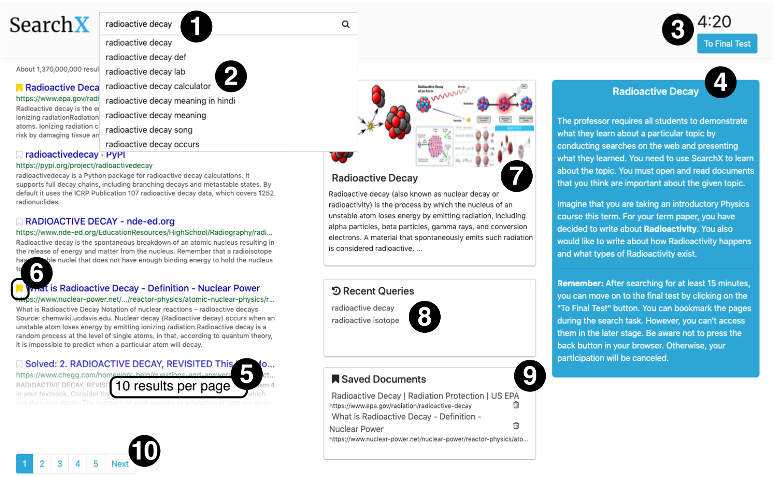
Search As Learning (SAL)

SAL is concerned with exploring how search engines can aid users in learning, in both the formal and informal setting. Prior studies [145, 351] explored the impact of domain expertise on learning. [145] observed that participants who are less familiar with a particular topic achieve slightly greater knowledge than users already familiar with the topic (though it is not yet clear whether this finding is mainly an artefact of the topics and the manner of how learning is measured). [351] noticed the difference between experts and non-experts in terms of their learning toward the end of the search task. Previous research also suggests that domain experts employ different search strategies (in terms of queries posed, documents viewed, etc.) to find what they are looking for compared to non-experts [351, 431].

An important aspect of SAL are cheap and easy to measure user behaviours that allow us to estimate the amount of learning taking place—this in turn would allow us to adapt search algorithms and interfaces on-the-fly. [125] studied the flow of evolving expertise within search sessions purely based on users' search behaviours. It was shown that SERP snippets and documents viewed inspire users' queries and reveal information about users' domain knowledge. Other proxies for learning explored include: eye movement patterns [87]; documents saved and opened [14, 145, 462]; as well as SERP clicks [14, 89]. While most studies focus on lower cognitive levels, [223] studied how search behaviours correlate with information needs at different cognitive levels. They found that users' search interactions with the SERP increased as participants moved towards tasks with higher cognitive levels of complexity.

To *explicitly* measure the learning gain (instead of inferring it from search behaviours), many lab-based user studies assess the knowledge of users before and after the search sessions via vocabulary tests, mind maps and the writing of summaries [89, 270, 271, 317, 389, 433]. Following this setup, we investigate in this work the impact that entity cards have on users' vocabulary learning and search behaviour during a learning-oriented search task.

6.3. Entity Card Implementation



radioactive decay

radioactive decay
radioactive decay def
radioactive decay lab
radioactive decay calculator
radioactive decay meaning in hindi
radioactive decay meaning
radioactive decay song
radioactive decay occurs

4:20

About 1,370,000,000 resul

Radioactive Deca
https://www.epa.gov/radi
Radioactive decay is the e
ionizing radiationRadiation
atoms. Ionizing radiation c
risk by damaging tissue an

radioactivedecay · PyPI
https://pypi.org/project/radioactivedecay
radioactivedecay is a Python package for radioactive decay calculations. It supports full decay chains, including branching decays and metastable states. By default it uses the ICRP Publication 107 radioactive decay data, which covers 1252 radionuclides.

RADIOACTIVE DECAY - nde-ed.org
https://www.nde-ed.org/EducationResources/HighSchool/Radiography/radi...
Radioactive decay is the spontaneous breakdown of an atomic nucleus resulting in the release of energy and matter from the nucleus. Remember that a radioisotope has ble nuclei that does not have enough binding energy to hold the nucleus to

hat is Radioactive Decay - Definition - Nuclear Power
ps://www.nuclear-power.net/.../reactor-physics/atomic-nuclear-physics/r...
What is Radioactive Decay Notation of nuclear reactions – radioactive decays Source: chemwiki.ucdavis.edu. Nuclear decay (Radioactive decay) occurs when an unstable atom loses energy by emitting ionizing radiation.Radioactive decay is a random process at the level of single atoms, in that, according to quantum theory, it is impossible to predict when a particular atom will decay.

Solved: 2. RADIOACTIVE DECAY, REVISITED This Mo...
https://www.chegg.com/homework-help/questions-and-answers ct...
RADIOACTIVE DECAY, REVISIT em 4
in your textbook. Consider the which
can then also decay. The numbers of each species as a function of time are given

Radioactive Decay

Radioactive decay (also known as nuclear decay or radioactivity) is the process by which the nucleus of an unstable atom loses energy by emitting radiation, including alpha particles, beta particles, gamma rays, and conversion electrons. A material that spontaneously emits such radiation is considered radioactive. ...

Recent Queries

radioactive decay
radioactive isotope

Saved Documents

Radioactive Decay | Radiation Protection | US EPA
https://www.epa.gov/radiation/radioactive-decay
What is Radioactive Decay - Definition - Nuclear Power
https://www.nuclear-power.net/nuclear-power/reactor-physics/ato...

Radioactive Decay

The professor requires all students to demonstrate what they learn about a particular topic by conducting searches on the web and presenting what they learned. You need to use SearchX to learn about the topic. You must open and read documents that you think are important about the given topic.

Imagine that you are taking an introductory Physics course this term. For your term paper, you have decided to write about **Radioactivity**. You also would like to write about how Radioactivity happens and what types of Radioactivity exist.

Remember: After searching for at least 15 minutes, you can move on to the final test by clicking on the "To Final Test" button. You can bookmark the pages during the search task. However, you can't access them in the later stage. Be aware not to press the back button in your browser. Otherwise, your participation will be canceled.

Figure 6.1: **The `SearchX` interface as used for this study. Included in this screenshot at the 10 superimposed annotation marks: (1) the query box with (2) autocompletion; (3) the timer that indicates the time spent in the search session so far; (4) the task description; (5) the ten search results per page which can be (6) saved to the (9) *Saved Documents* box; (7) the entity card; (8) the list of *Recent Queries*; and finally (10) pagination. Note that this figure shows an entity card from the `Good-EC` condition.**

The present study was undertaken using `SearchX` [335], an open-source, modular retrieval framework that allows one to undertake crowdsourced IIR experiments. Out-of-the-box, `SearchX` provides quality assurances and basic logging functionalities, ensuring that only high-quality participants complete a study—and that the necessary interactions are logged. As entity cards are not yet supported by `SearchX`, we implemented a novel entity card component for it.

Figure 6.1 demonstrates the user interface that was used by the participants of our study. Users can issue their queries in the query box which also offers query autocompletion provided by the *Bing Autosuggest API*[2]. We present ten search result snippets per page, drawn from the *Bing Search API*. Pagination is provided at the bottom of the SERP. Participants can easily bookmark documents and access them in the *Saved Documents* box. In addition, *Recent Queries* a user issued are also shown in a separate

[2] All Bing APIs used can be found at `https://www.microsoft.com/en-us/bing/`—all URLs listed in this chapter were last accessed on April 27th, 2021.

box. The description of the search task appears in the top right corner of the SERP. The timer above the task box helps users gauge the elapsed time. Our entity card is always presented at the position shown in Figure 6.1 and presents concise information regarding one significant entity within the query and search results. The remainder of this section discusses the structure of our entity cards, and the three variants of entity cards we evaluate—good (`Good-EC`), fair (`Fair-EC`) and poor/bad (`Bad-EC`) quality cards.

6.3.1. Entity Card Structure

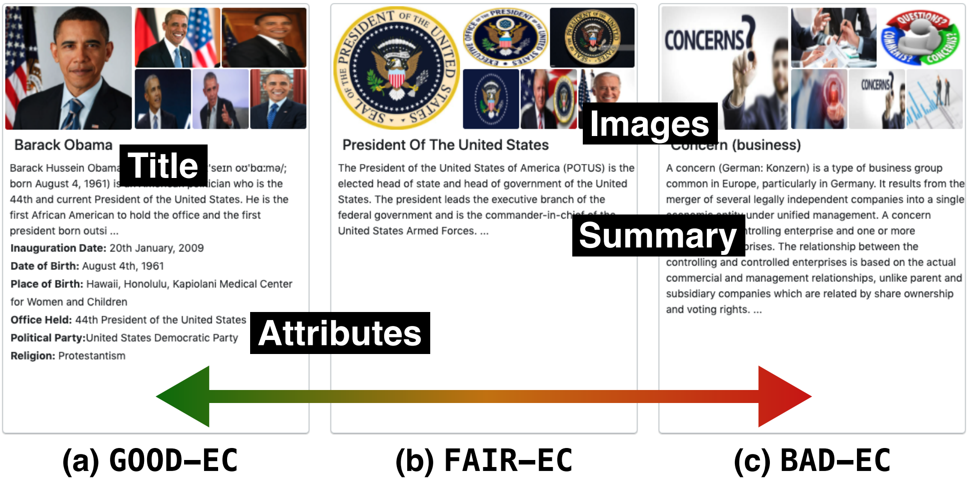
Figure 6.2 illustrates the structure of our entity cards. Each entity card consists of *up to* four components: *(i)* a set of *images*, which were obtained from the *Bing Image Search API*; *(ii)* the entity's *title*; *(iii)* its' *Wikipedia-based summary*; and *(iv)* multiple *attributes* whose existence and number are dependent on *DBPedia's* open knowledge graph, which contains structured content of various *Wikipedia* projects. In Figure 6.2, only attributes for the entity *Barack Obama* are shown. This is *not* the consequence of the experimental condition, but instead due to our decision of filtering out rare attributes. More concretely, we processed DBpedia version 2016-10[3], and remove all attributes that occurred in fewer than 20% of attributes of a particular type to avoid distracting users by the presence of unusual attributes (such as *eye colour* for entities of type *Person*).

6.3.2. Entity Card Rankers

The most important question in the setup of our study is how to determine the ranking of entities: for a given query, once a ranking of entities has been established, we are able to determine the good, fair and bad entities for a query by considering the ranks at which entities are retrieved. For each query a user submits, we concatenate the user query and the top 10 search results snippets. We opted to not include the actual document content in this step as this would require an additional ten HTTP requests, slowing down our SERP's responsiveness significantly (and a slow responsiveness is known to decrease user engagement [249]).

After setting up the context as the concatenation of the query and top ten retrieved document snippets, we then need to retrieve the ranking of the entities through keyword extraction methods. As described in Section 6.2, several unsupervised approaches for keyphrase extraction exist. Besides the already noted advantages of unsupervised approaches, we also aim to detect keyphrases on-the-fly, thus requiring a fast algorithm (and inference of a large neural network for instance has significant speed constraints). Based on prior works, we selected three keyphrase extraction approaches that are all corpus-independent: `Yake` [72], `RaKUn` [374] and `EmbedRank` [35]. We select these algorithms as: *(i)* they have functioning open-source implementations; *(ii)* they are lightweight, unsupervised algorithms that produce output in a timely manner; and *(iii)* they are robust (i.e., they do not degrade in effectiveness significantly) to changes in collections and domains. For each of these algorithms, we provide our query and document snippets as input, and consider the resulting top 20 ranking of keyphrases. Highly-ranked keyphrases have the highest relevance score with respect to query and document snippets. In order to convert the ranking of these 20 keyphrases into a ranking of 20 en-

[3] https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10



Barack Obama

Barack Hussein Obama 'seɪn oʊ'bɑːmə/; born August 4, 1961) is ician who is the 44th and current President of the United States. He is the first African American to hold the office and the first president born outsi ...

Inauguration Date: 20th January, 2009

Date of Birth: August 4th, 1961

Place of Birth: Hawaii, Honolulu, Kapiolani Medical Center for Women and Children

Office Held: 44th President of the United States

Political Party: United States Democratic Party

Religion: Protestantism

President Of The United States

The President of the United States of America (POTUS) is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander-in-ch United States Armed Forces. ...

Concern (business)

A concern (German: Konzern) is a type of business group common in Europe, particularly in Germany. It results from the merger of several legally independent companies into a single under unified management. A concern ntrolling enterprise and one or more prises. The relationship between the controlling and controlled enterprises is based on the actual commercial and management relationships, unlike parent and subsidiary companies which are related by share ownership and voting rights. ...

(a) GOOD–EC **(b) FAIR–EC** **(c) BAD–EC**

Figure 6.2: **Demonstration of the three types of entity cards (for the query** `barack obama`**). From left to right: *(a)*** `Good-EC`**, a high-quality, on-topic entity card; *(b)*** `Fair-EC`**, another somewhat relevant entity card, but not the first choice; and *(c)*** `Bad-EC`**, an entity card not relevant to the query.**

tities, we employ the TagMe API[4]. This API links each keyphrase to at least one pertinent Wikipedia page. In any cases, TagMe returns at least one output. We chose the output of TagMe with the highest probability score and fixed this as the entity corresponding to the keyphrase.

For simplicity, we refer to the keyphrase extraction algorithms now as our *entity rankers*, as the procedure to convert the extracted keyphrases to entity rankings (via the TagMe API) is the same for all three. Next, we describe the user study we conducted to determine which of the three entity rankers provides us with the best ranking.

6.3.3. Comparison of Entity Card Rankers

First, we fixed a list of ten topics[5] randomly drawn from the *TREC 2019 Decision (Health Misinformation) Track*[6]. We asked ten volunteers of a computer science lab to provide up to five queries for each of the topics, whose TREC topic description we provided to them. This resulted in between 10 and 27 unique queries per topic, with a median number of 12 unique queries. Each of these queries was submitted to the Bing Search API, from which the top ten result snippets were extracted. Based on this input, we retrieved the entity rankings from Yake, RaKUn and EmbedRank respectively.

As in our actual experiment, we only show one entity card per query as we are most interested in the top-ranked entity retrieved by each algorithm. For this reason, we now focus on the top ranked entity. We discard queries for which at least two of the three algorithms produced the same entity, leaving us with 70 queries—and the three respective

[4] https://sobigdata.d4science.org/web/tagme/tagme-help

[5] The ten TREC topic titles are *acupuncture insomnia, ear drops remove ear wax, honey wound, melatonin jet lag, magnesium muscle cramps, insulin gestational diabetes, vaccine common cold, antibiotics children pneumonia, caffeine asthma*, and *surgery obesity*.

[6] https://trec.nist.gov/data/misinfo2019.html

Table 6.1: **An example query chain, as drawn from a search session under the *Glycolysis* topic. Corresponding entity cards that were presented for each query across conditions Good-EC, Fair-EC, and Bad-EC are shown in their respective columns.**

Query	Good-EC	Fair-EC	Bad-EC
1. glycolysis	glycolysis	nicotinamide adenine dinucleotide	river source
2. pyruvate pyruvic acid	lysosomal acid lipase	pyruvic acid	ration
3. cellular respiration	cellular respiration	adenosine triphosphate	art
4. major phases of glycolysis	glycolysis	sugar	steps

top-retrieved entities.

We then randomly selected three topics from our initial list of 10 topics. For each topic, we randomly drew four queries from the collected queries and assigned them to 32 volunteers to judge which of the three top-ranked entity cards from Yake, RaKUn, and EmbedRank respectively are *useful* given the information need (i.e., the TREC topic). They could select multiple options or select *None* to signify that they do not consider any of the presented entity cards to be useful. Overall, EmbedRank's top retrieved entity was selected as the useful entity for 35% of the queries, in contrast to RaKUn's and Yake's 17% and 15% respectively. For 21% of queries, none of the algorithms returned a useful top-ranked entity.

Based on these results, we opted to take EmbedRank forward as our entity ranker throughout the remainder of the experiment discussed in this chapter.

6.3.4. Entity Card Types

Given a query and an entity ranking produced by EmbedRank, we create three types of entity cards.

- Good-EC The top ranked entity is selected.
- Fair-EC The entity at the rank five is selected.
- Bad-EC The entity at rank 20 is selected.

To provide the reader with an impression of the entities retrieved for the three entity card conditions, we refer to Table 6.1. For example, for query *cellular respiration* (the third query), the entities in descending order of quality are as follows: *cellular respiration* (Good-EC); *adenosine triphosphate* (Fair-EC); and *art* (Bad-EC). For completeness, the No-EC condition is our control condition: here, no entity card is presented on the SERP.

Note that besides retrieving different entities for the different conditions, we do not alter the way the entity card looks for each entity type. In particular, what type of attributes and imagery is shown depends on the information available on Wikipedia/DBPedia, and not the entity card type. As stated previously, Figure 6.2 provides an example query and the three entity cards generated.

Evaluation of Entity Cards. In order to determine whether our intended quality levels of the entity cards were in fact correct, we manually evaluated the quality of the entity cards the participants received in response to their queries across all conditions.

In our manual labelling effort, we labelled entity cards as *good* when the entity card title exists in the query. The label *fair* is given to entity cards when their title aligns with any concepts related to the query. Lastly, we mark entirely off-topic entity cards as *bad*. The manual annotation of 743 participants' queries (and the corresponding entity card) led to the following results: 87.2% of entity cards shown in the `Good-EC` condition were annotated as *good*, 82.4% of entity cards shown in the `Fair-EC` condition were labelled as *fair*, and 99.2% of entity cards shown in the `Bad-EC` condition were labelled as *bad*. This gives us confidence that the entity cards presented to participants during the experiment fell into line with our expectations.

Additionally, we asked the participants post-test to evaluate the relevance of the entity cards to their queries. 82.8% of the participants in the `Good-EC` condition asserted that entity cards are *Mostly/Always Relevant* to the queries, while the proportion is 76.1% for the participants of the `Fair-EC` condition, and 7.5% for participants in the `Bad-EC` condition.

Table 6.2: **Example annotations of participants' definitions of vocabulary terms for the topic *Glycolysis*.**

Vocabulary term **pyruvate**	
Correct	*A compound that is produced via glycolysis and is related to pyruvic acid.*
Partially correct	*It is a product of glycol is it can help with fat burning.*
Incorrect	*A molecular unit of sugar.*
Vocabulary term **krebs cycle**	
Correct	*The Krebs cycle is also called the citric acid cycle. It's a series of chemical reactions which require oxygen and get energy from food. It can only be aerobic. It produces ATP and also other compounds used by the electron transport chain.*
Partially correct	*Also known as they citric acid cycle.*
Incorrect	*A cellular process that helps an organism live.*

6.4. Study Design

We now describe our user study in more detail. We go over the search topics, how to measure learning, and the workflow the participants of the study followed.

6.4.1. Topics

We employ three of the topics introduced by [301]'s search as learning study: *Glycolysis*, *Radioactive Decay*, and *Qubits*. Each of these topics comes with a list of 10 vocabulary terms that have been manually curated by the authors.

For example, for the *Glycolysis* topic, vocabulary terms include *krebs cycle*, *electron transport chain*, and *cellular respiration*. These vocabulary terms are terms that: *(i)* were mentioned in a specific video lecture about the topic at least once; and that *(ii)* do not frequently occur outside of this domain-specific context. In their work, [301] proposed a list of in total 10 topics. We chose the three listed above based on the availability of entity cards: concretely, we received the query log of [301], submitted all the queries for each topic to the Bing Search API, and ran `EmbedRank` to retrieve the respective entity rankings. We then selected the three topics with the largest number of relevant entities. Specifically, the topic *Radioactive Decay* (with a rate of 6.02 entities per query) has the greatest number of entities per query, followed by topics *Glycolysis* and *Qubits* with 5.4 and 4.7 entities per query, respectively.

6.4.2. LEARNING GAIN

We measure our participants' learning gain by measuring their difference in knowledge in a pre-test (conducted right before the search session) and a post-test (conducted right after the search session) in line with [**Yu_2018**, 145, 301, 351, 388, 389]. As in [301, 351], we employ the slightly modified *Vocabulary Knowledge Scale (VKS)* test [378, 389, 429], which demonstrate the incremental stage of stages of word learning [97]. For every vocabulary term, our participants are asked about their knowledge across four levels:

(1) I don't remember having seen this term/phrase before.

(2) I have seen this term/phrase before, but I don't think I know what it means.

(3) I have seen this term/phrase before, and I think it means ___ .

(4) I know this term/phrase. It means ___.

Note that for levels *(3)* and *(4)*, we require participants to write their definition of the term. The difference between the two is in the certainty of the participants' knowledge: in level *(3)* the uncertainty is high; with level *(4)*, participants are certain about their knowledge.

Again, in line with prior works [90, 301, 387, 389], we employ *Realised Potential Learning (RPL)* to measure the learning gain which normalises *Absolute Learning Gain (ALG)* by the *Maximum possible Learning Gain (MLG)*. ALG is an aggregated difference in knowledge level before and after the search session across all vocabulary terms—with the added proviso that *knowledge cannot degrade over the time of the search session (between the pre- and post-tests).*

Here, $vks^{pre}(v_i)$ and $vks^{post}(v_i)$ indicate the scores assigned to vocabulary term v_i in the pre- and post-test, respectively. We set the vks score to 0 knowledge levels *(1)* or *(2)*. We also assign the score of 1 for both knowledge levels *(3)* and *(4)*, which is in line with the binary setup employed in [301]. RPL is computed as follows.

$$ALG = \frac{1}{n}\sum_{i=1}^{n}\mathbf{max}(0, vks^{post}(v_i) - vks^{pre}(v_i))$$

$$MLG = \frac{1}{n}\sum_{i=1}^{n} maxScore - vks^{pre}(v_i)$$

$$RPL = \frac{ALG}{MLG}$$

6.4.3. PROCEDURE

When a participant enters the study, the online learning experiences questionnaire is presented, consisting of seven questions. These questions are inspired by [350], and focus on online learning experiences with the goal to prime participants for the upcoming task. Then, we present the pre-test for the three topics to each participant. For each topic (in addition to the 10 vocabulary knowledge questions), we include three more general questions to probe the participants.

- *How much do you know about this topic?*
- *How interested are you to learn more about this topic?*
- *How difficult do you think it will be to search for information about this topic?*

Thus, in the pre-test, each participant answers a total of $7 + 3 \times 13 = 46$ questions. Subsequently, the participants move on to the search phase where they are randomly assigned to one of our four experimental conditions. For the topic, the one with the least amount of prior knowledge (computed from the answers to their pre-test questions) is selected. Before starting the search task, a tutorial is shown to the participant providing information about how to interact with different interface components. The search task presented to the participants is the following (the underlined phrases are specific to each search topic).

> Imagine that you are taking an introductory Physics course this term. For your term paper, you have decided to write about Radioactivity. You also would like to write about how Radioactivity happens and what types of Radioactivity exist.

The minimum search time was fixed to fifteen minutes to provide sufficient time to search and learn while alleviating fatigue. We relied on the Bing Search API as our search backend, and filtered out any search results originating from Wikipedia or any of its mirrored pages. As we aim for our participants to search in order to learn, we removed this source of information to avoid participants spending their search time reading a single Wikipedia document.

During the search session, participants can search, view, and bookmark documents. We disable copy and paste options and limit the tab changes to a maximum of two to avoid participants searching the web to answer our questions. At three browser tab changes, a participant is disqualified from the study.

The experiment ends with a post-test, which contains the same vocabulary knowledge test as the pre-test this time though only focused on the one topic assigned to the participant. Additionally, participants are tasked with writing a summary with a minimum of 100 words, and the term paper's outline as indicated in the search task description. Lastly, we include 10 questions regarding the entity cards, their experience working with our search system, their perceived learning, and perceived search success.

6.4.4. PARTICIPANTS

We conducted our user study on the *Prolific* Academic Platform. We required our $N = 144$ participants to: *(i)* have at least 15 accepted Prolific task submissions; *(ii)* be native English speakers (limiting participants to be from only the United Kingdom); and *(iii)* have a minimum approval rate of 85%. The study took approximately 40 minutes to complete. We paid our participants GBP£6.43 per hour for the experiment. Among our participants, 64.5% were female, and 35.5% were male. We report a mean age of 32.4 (minimum 18 years, maximum 74 years). Due to the nature of crowd-sourced studies, we continued to add more participants to our Prolific task until we reached 36 participants for each condition.

6.4.5. Vocabulary Knowledge Assessment

In total, participants provided us with 394 concept definitions (across both the pre-test definitions written for the topic that was eventually selected for the respective participant, and the post-test definitions) when self-assessing their knowledge as level *(3)* or *(4)* (see §6.4.2). We manually evaluated all provided concept definitions and labelled each one as either *correct*, *partially correct*, or *incorrect*. Examples of definitions and the labels we assigned to them are provided in Table 6.2. More formally, we employed the following criteria to judge each definition provided by a participant.

(2) **Correct** If a participant explains one related concept without any errors, their definition was assigned the highest score. Furthermore, the highest score was given to the participant's definition which explains multiple related concepts, while leeway was given if an error was in *one* of the concepts.

(1) **Partially Correct** The participant's definition describing one related concept with any errors was given a score of 1. This score also applied to participants whose definition provided a correct synonym for the term. For example, the *Krebs cycle* is also known as the *citric acid cycle*.

(0) **Incorrect** Definitions that are either entirely incorrect or trivial (e.g., *'beta-minus decay is a kind of decay'*).

As a first step in our annotation, we randomly sampled 50 of the vocabulary term definitions (13% of the total available terms). The authors then annotated them independently according to the above correctness criteria. *Inter-annotator agreement*, computed as *Cohen's kappa*, is 0.83. With this high rate of agreement, we then split the remaining definitions and annotated them independently. In contrast to prior works [69, 301, 352], we did *not* rely on self-assessments of knowledge. Instead, we instead manually verified to what extent these self-assessments were correct. We found that for knowledge level *(3)* (see §6.4.2): 31% of the provided term definitions were identified as being correct; 38% were partially correct; with the remaining 31% incorrect. From the vocabulary terms self-assessed as knowledge level *(4)*: 48% were correct; 25% were partially correct; with the remaining 27% incorrect.

6.5. Results

To address our two research questions as outlined in Section 6.1, measures were analysed by using a two-way ANOVA. These were conducted considering both the conditions and topics as factors; main effects were examined where $\alpha = 0.05$. For post-hoc analysis, the TukeyHSD pairwise test was used. For results in all tables, ± values denote the standard deviation.

Entity Cards and Learning (RQ1)

RQ1 asks to what extent entity cards of varying quality impact the amount of learning taking place. Table 6.3, row **X**, presents the RPL across the four experimental conditions. To complement this, rows **XI**-**XII** also report the RPL achieved over each of the three topics. As these measures only provide a high-level overview of the learning gain, Figure 6.3

Table 6.3: **Mean (± standard deviations) of RPL and search behaviour measures across all participants over each of the four experimental conditions. A † indicates two-way Anova significance, while ^{G,F,B,N} reveals post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) increases vs. `Good-EC`, `Fair-EC`, `Bad-EC`, and `No-EC` conditions, respectively.**

			`Good-EC`	`Fair-EC`	`Bad-EC`	`No-EC`
Behavioural	**I**	Number of participants	36	36	36	36
	II	Search session duration (mm:ss)†	16:43 (± 4:22)	16:15 (± 2:04)	16:58 (± 3:06)^{N}	15:44 (± 0:45)^{B}
	III	Number of queries	4.89(±2.23)	5.18(±2.41)	5.62(±2.69)	6.03(±2.89)
	IV	Fraction of entity card terms within the subsequent query†	0.69(±0.30)^{FB}	0.34(±0.29)^{GB}	0.02(±0.05)^{GF}	-
	V	Average time between queries (secs)	221.08(±187.32)	197.56(±162.22)	187.99(±113.41)	164.5(±80.87)
	VI	Number of hovers over entity cards	14.94(±9.01)	16.62(±9.73)	13.94(±6.35)	-
	VII	Average time between documents (secs)	79.41(±82.75)	63.96(±43.36)	73.57(±60.10)	67.43(±49.68)
	VIII	Average document dwell time (secs)†	126(±69.6)	117(±65.4)	151.2(±93.6)^{N}	113.4(±48.6)^{B}
	IX	Number of unique documents viewed	8.25(±4.22)	8.56(±3.8)	8.68(±3.24)	9.75(±3.98)
Learning	**X**	RPL (over all topics)	0.19(±0.21)	0.22(±0.22)	0.17(±0.22)	0.18(±0.17)
	XI	RPL for topic *Radioactive Decay*	0.16(±0.19)	0.12(±0.17)	0.09(±0.14)	0.13(±0.16)
	XII	RPL for topic *Qubits*	0.15(±0.19)	0.22(±0.23)	0.11(±0.12)	0.15(±0.11)
	XIII	RPL for topic *Glycolysis*	0.26(±0.26)	0.35(±0.20)	0.34(±0.31)	0.25(±0.22)

Table 6.4: **Summary statistics for the three topics used in our study (± standard deviations). A † indicates two-way Anova significance, while ^{R,Q,G} indicate post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) vs. *Radioactive Decay*, *Qubits*, and *Glycolysis*, respectively.**

			Radioactive Decay	***Qubits***	***Glycolysis***
#	**I**	Total participants	48	48	48
	II	# in conditions `Good-EC`, `Fair-EC`, `Bad-EC`, and `No-EC`	12	12	12
Behavioural	**III**	Average number of queries	5.0(±2.40)	5.90(±2.44)	5.39(±2.93)
	IV	Median number of queries	4	5.5	4
	V	Average time between queries (sec)†	210.56(±102.47)^{Q}	145.48(±55.35)^{RG}	224.99(±215.84)^{Q}
	VI	Median time between queries (sec)	183.53	142.07	148.54
	VII	Average number of bookmarks	2.88(±2.83)	3.15(±3.67)	3.25(±2.97)
	VIII	Median number of bookmarks	2.5	2	3
	IX	Average number of unique documents viewed†	7.90(±3.66)^{G}	8.83(±3.72)	9.80(±3.99)^{R}
	X	Median number of unique documents viewed	7	8.5	9
RPL	**XI**	RPL†	0.12(±0.16)^{G}	0.16(±0.16)^{G}	0.30(±0.25)^{RQ}
	XII	Median RPL	0.10	0.11	0.30

plots the *distribution* of learning gain across participants for each of the four conditions and each of the three topics respectively.

We first focus on the learning gain across four experimental conditions. For our control condition (`No-EC`), the average *RPL* is 0.18, which means that participants gain on average 18% of the knowledge they could have gained at best. When comparing `No-EC` with the other conditions, we do not observe significant differences in learning gain. We also observed that the learning gain for the `Bad-EC` is the lowest compared to other conditions; lower than even `No-EC`. Additionally, Figure 6.3(a) shows that for both `Good-EC` and `Fair-EC`, the variability in RPL scores across participants is larger than for the other two conditions. Although there is no significant difference across conditions, these findings suggest that (at least partially) relevant entity cards may improve learning gain, but only marginally. In contrast, poor entity cards could negatively impact

Table 6.5: **Source of terms for query reformulations. A † indicates two-way Anova significance, while ^{G,F,B,N} indicate post-hoc significance (TukeyHSD pairwise test, p < 0.05) vs. `Good-EC`, `Fair-EC`, `Bad-EC`, and `No-EC` conditions, respectively.**

		`Good-EC`	`Fair-EC`	`Bad-EC`	`No-EC`
I	Fraction of query terms from prior snippets†	0.29(±0.25)	0.18(±0.19)^{N}	0.25(±0.23)	0.34(±0.17)^{F}
II	Fraction of query terms from prior documents†	0.50(±0.33)^{N}	0.35(±0.33)^{N}	0.41(±0.36)^{N}	0.58(±0.21)^{GFB}
III	Fraction of query terms from prior entity card titles†	0.24(±0.17)^{FB}	0.12(±0.14)^{GB}	0.01(±0.03)^{GF}	-
IV	Fraction of query terms from prior entity card summaries†	0.53(±0.24)^{FB}	0.39(±0.31)^{GB}	0^{GF}	-

on learning—with the suggestion that a bad entity card may distract participants from learning within complex topics. Rows **XI**-**XIII** of Table 6.3 also report the RPL across each condition, splitting it up by each of the three topics trialled.

Table 6.4 presents a summary of the RPL (amongst behavioural measures) from a per-topic perspective. We can see on row **XI** a large variation in the mean RPL attained over the three topics: 0.12 ± 0.16 for *Radioactive Decay*; 0.16 ± 0.16 for *Qubits*; and 0.30 ± 0.25 for *Glycolysis*. Indeed, *Glycolysis* was found to have a significantly higher level of RPL than either *Radioactive Decay* or *Qubits*. This meant that *Glycolysis* was considered the easier topic on average, with *Radioactive Decay* appearing to be the more complex. The differences between topics are also visible in Figure 6.3(b): *Glycolysis* has the highest median with the greatest variability in learning gain. What had an impact on knowledge gain is the distribution of topics among participants.

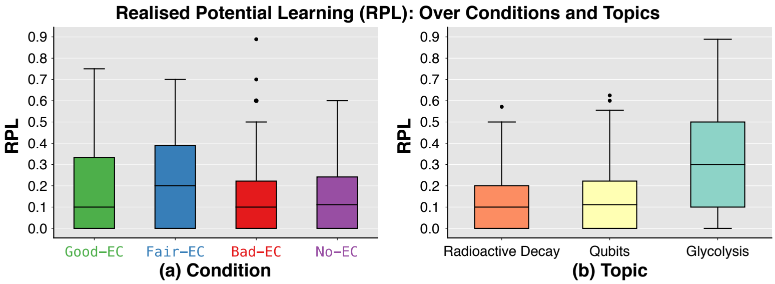


Figure 6.3: **RPL, considered over both the four experimental conditions (a), and (b) the three topics trialled.**

Given these observations, we find the presence of entity cards (no matter their quality with respect to the issued queries) to not lead to higher learning gains (thus addressing **RQ1**). However, comparing the RPL across our conditions, we can see that bad/poor entity cards (`Bad-EC`) have detrimental impact on an individual's learning. Results show that topic difficulty does play a major role, with significant differences found between the mean performance of participants when the three topics are considered separately.

Entity Cards and Search Behaviours (RQ2)

We return to Table 6.3 for insights into the search behaviours exhibited by participants over each of the four conditions trialled, as shown on rows **I**-**IX**.

We first examine the recorded search session duration reported on row **II** of Table 6.3. With results presented in minutes and seconds, we observe that for both `Good-EC`

(16:43±4:22) and Fair-EC (16:15±2:04), the mean session time is approximately one minute longer than for No-EC (15:44±0:45). We also note that participants spent significantly longer using interface Bad-EC(16:58±3:06) (with a higher variance) than on No-EC. Together, these findings suggest a slightly higher engagement with the task and interface when entity cards were present on our search interface, regardless of the quality of the cards provided. Looking deeper, we find that this pattern was repeated when considering average document dwell times, with the same patterns once again being observed (see row **VIII**). Examining the interactions with the entity cards themselves, we note that the mean number of hovers over the entity cards was found to be approximately 15 for all three conditions containing them (Table 6.3, row **VI**). No significant differences were observed. A similar number of documents were examined across all four conditions, once again without any observed significant differences (Table 6.3, row **IX**). Here, No-EC has the highest number of viewed documents on average, at 9.75 ± 3.98. This intuitively makes sense: no entity cards means the only source that participants could gain information was to go and read the linked documents. To complement this, we observe a trend: a decrease in the number of unique documents viewed as the quality of the presented entity cards increases. Here, we hypothesise that as entity card quality increased, participants had a greater likelihood of being able to satisfy their information need on the SERP without having to resort to clicking links.

In terms of the number of queries issued, participants in the No-EC condition on average issued the greatest number of queries on average (6.03 ± 2.89), though this was not significantly so (Table 6.3, row **III**). We observe a consistent increasing trend in the number of queries issued as the entity card quality drops (or the entity card is absent), starting from Good-EC (4.89 ± 2.23) and ending at No-EC (6.03 ± 2.89). When receiving (partially) relevant information from the entity cards, we speculate that participants were able to obtain important information for their information need from them. Correspondingly, in terms of the average time between queries, participants in the Good-EC condition recorded the highest time (221.08 seconds) with that time dropping as we move along the conditions towards poor entity cards (Table 6.3, row **V**).

Within the post-test, we also asked participants how much they paid attention to entity cards to ensure the impact of entity cards on their behaviour. A total of 90% of participants of Good-EC stated they examined entity cards regularly. 90.3% and 82.5% of Fair-EC and Bad-EC participants respectively also self-reported paying attention to them regularly.

In order to examine whether entity cards influenced the terms that appeared in subsequent queries in the search sessions, we examined the fraction of entity card terms occurring within the queries issued by the participants (Table 6.3, row **IV**). We found significant differences between the three conditions that presented entity cards, with a clear, increasing trend from Bad-EC (0.02 ± 0.05), through to Fair-EC (0.34 ± 0.20), up to Good-EC (0.69 ± 0.30). Significant differences existed between all conditions, suggesting that participants were able to judge the quality of the entity cards and employed them when formulating their queries (e.g., through the learning of terms to then issue to the search engine). Spurred by this finding, we examined this phenomenon in more detail.

In terms of query reformulations, we observed entity cards to have a considerable impact. In Table 6.5, we examined the *source* of participant's query terms. We report the following statistics.

- **Fraction of query terms from prior snippets** Here, we consider *previously observed* snippets as a potential source for query terms.
- **Fraction of query terms from prior documents** For each query, we consider all *previously viewed* documents, and compute the fraction of query terms that appeared in at least one of them.
- **Fraction of query terms from prior entity card titles** Here, instead of considering previously viewed documents, we consider only entity card titles.
- **Fraction of query terms from prior entity card summaries**. Finally, we consider the entity card summary text, instead of the title.

We acknowledge that this can only be considered an approximation, as we do not know whether for instance a term present in a viewed document was even read by a participant (this likely requires eye-tracking hardware and analysis, as per [124]). However, significant differences were found across all four additional measures. If we first consider the measures corresponding to the entity cards, it is unsurprising to note that the fraction of query terms from both entity card titles and summaries were significantly higher for `Good-EC` than `Bad-EC`, with `Fair-EC` once again, on average, landing in between the two extremes. From the `Good-EC` summaries, for example, the fraction of terms in participant queries jumped from 0.53 ± 0.24 down to a flat 0 for `Bad-EC`—this acts as a sanity check, confirming that `Bad-EC` entity cards always yielded entity cards that did not correspond to the given query.

Taking this analysis further, we also extracted *query chains* from our gathered interaction logs to examine what terms were actually used. Table 6.1 presents an example query chain drawn from a participant's interaction log over the *Radioactive Decay* topic. Along the first column are the queries issued by the participant, with the associated entity card titles shown for each of the three conditions. We can see that the terms that appear in the issued queries correspond closely to those in `Good-EC`, with the third query's terms matching those of the suggested `Good-EC` exactly.

These results show that there is at least some interaction effect in the search and learning process, where entity cards are priming and providing participants with query terms to assist in their query formulation patterns. Further work is required to investigate this.

6.6. Conclusions

In this chapter, we examined to what extent entity cards impact users' learning gains (**RQ1**) and search behaviours (**RQ2**) for learning-oriented search tasks.

To answer our two research questions, we conducted a crowdsourced user study, where $N = 144$ participants were assigned to one of four conditions. The conditions controlled whether entity cards were present on the SERP, and if present, dictated whether they were *good* (relevant to the query), *fair* (contained a degree of relevant information), or *bad* (not relevant to the query). We evaluated participants' knowledge with a vocabulary learning test.

Our results show that entity cards—as used in our experimental setup—do not significantly affect human learning, with RPL scores consistently low and without significant differences between conditions. On the other hand, significant differences were found when examining topic effects.

When considering the search behaviours of participants, we did observe a number of significant differences across the four conditions. For example, varying the entity cards presented significantly impacted on the dwell time spent over documents, and overall session duration. We also observed a consistent trend that with lower quality entity cards the number of queries increase, although this was not significant. Similarly, as the entity card quality decreases, the number of unique documents viewed was shown to increase consistently across conditions (though again, not significantly so). When examining query terms issued by our participants, we began to see evidence that demonstrated that participants may indeed be examining the entity cards and using them to reformulate their queries, assisting in the learning process. Significant differences were observed when considering the fraction of query terms appearing in entity card title and summaries.

Our study has several limitations related to the task (artificial in nature), evaluation regime (we only consider vocabulary learning) and study setup (we are limited to a single search session).

Our study did not regard *concept difficulty*, and instead focused purely on providing entity cards based on the entity rankings derived from `EmbedRank`. We also opted to show a single entity card, as this is the common web search setup. However, some evidence [**article**] suggests that multiple entity cards may also be suitable for a learning environment. Introducing different entity card styles (depending on a participant's prior knowledge levels or their search strategies) would also be an interesting direction for future work. Instead of simply taking a Wikipedia summary and some basic attributes for the entity in question, richer content could be included based upon prior search history. In order to gain insights into the impact of entity cards on higher-level learning, we also need to explore more complex learning tasks and move beyond a single search session setup. As continuation of work by [406], we may also want to study the effect of entity cards in different domains along various cognitive processes (apply, evaluate, create) and knowledge types (factual, conceptual, procedural).

7

Exploring the Feasibility of Crowd-Powered Decomposition of Complex User Questions in Text-to-SQL Tasks

Natural Language Interfaces to Databases (NLIDB), also known as `Text-to-SQL` models, enable users with different levels of knowledge in Structured Query Language (SQL) to access relational databases without any programming effort. By translating natural languages into SQL query, not only do NLIDBs minimize the burden of memorizing the schema of databases and writing complex SQL queries, but they also allow non-experts to acquire information from databases in natural languages. However, existing NLIDBs largely fail to translate natural languages to SQL when they are complex, preventing them from being deployed in real-world scenarios and generalizing across unseen complex databases. In this chapter, we explored the feasibility of decomposing complex user questions into multiple sub-questions — each with a reduced complexity — as a means to circumvent the problem of complex SQL generation. We investigated the feasibility of decomposing complex user questions in a manner that each sub-question is simple enough for existing NLIDBs to generate correct SQL queries, using non-expert crowd workers in juxtaposition with SQL experts. Through an empirical study on an NLIDB benchmark dataset, we found that crowd-powered decomposition of complex user questions led to

This chapter is based on a peer-reviewed paper: **Sara Salimzadeh,** Ujwal Gadiraju, Claudia Hauff, and Arie van Deursen. 2022. Exploring the Feasibility of Crowd-Powered Decomposition of Complex User Questions in Text-to-SQL Tasks. In Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '22), June 28-July 1, 2022, Barcelona, Spain. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3511095.3531282.

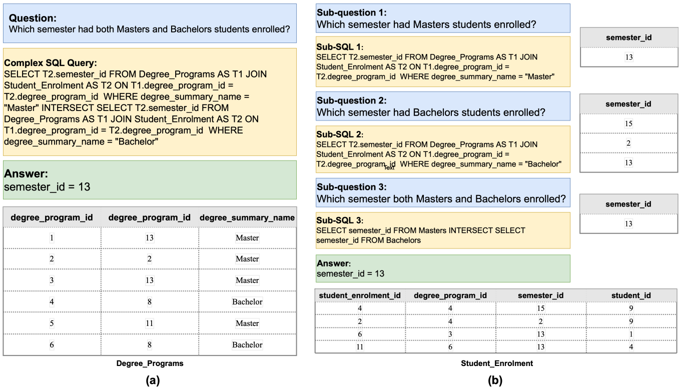
an accuracy boost of an existing `Text-to-SQL` pipeline from 30% to 59% (96% accuracy boost). Similarly, decomposition by SQL experts resulted in boosting the accuracy to 76% (153% accuracy boost). Our findings suggest that crowd-powered decomposition can be a scalable alternative to producing the training data necessary to build machine learning models that can automatically decompose complex user questions, thereby improving `Text-to-SQL` pipelines.

7.1. Introduction

Building Natural Language Interfaces to Databases (NLIDBs) has been identified as one of the most significant semantic parsing tasks for decades [36, 70, 112, 385, 395, 467]. By automatically converting text into the Structured Query Language (SQL), NLIDBs allow users to communicate with relational data in natural languages (NL) without any programming effort. These NL questions often cannot be directly answered by search engines. For example, in response to the question '*What are the total population and average area of countries in the continent of North America whose area is bigger than 3000?*', an NLIDB would return 480753000 and 1344763 for the total population and average area respectively; while a search engine would present a number of tables and leave the computation to the user. Such interfaces (also known as `Text-to-SQL` models within the NLP community) relieve users who are not proficient in query languages from the burden of learning techniques for querying databases by allowing them to pose NL questions.

Within recent years, the emergence of complex, large, and human-annotated datasets consisting of NL questions and their corresponding SQL queries has significantly developed the field. Traditionally these have included in-domain datasets such as WikiSQL [467], ATIS [96, 204], and Advising [135], more recently the family of `Spider` cross-domain datasets, including `Spider` [**yu2019spider**], SParC [453], and CoSQL [452] challenge the generalizability of models to unseen databases. Although recent studies have demonstrated the high accuracy (above 70%) of state-of-the-art `Text-to-SQL` models trained and evaluated on the `Spider` dataset, the performance of these models on complex queries is rather low, as many struggle to predict complex SQL queries, **Complex SQL Generation**. Parsing a question into a SQL query with nested queries, multiple SELECT clauses, GROUP BY, ORDER BY, UNION, INTERSECT, and EXCEPT requires a model to capture the semantic dependency between the NL question, database schema, and SQL syntax. According to the `Spider` criteria, SQL queries are classified into four difficulty levels – *easy*, *medium*, *hard*, and *extra hard*. The difficulty level is determined based on the number of SQL components, selections, conditions, nested sub-queries, column selections, aggregators, etc. Further, a question is complex when the corresponding SQL query is hard or extra hard.

Evaluating the accuracy of the top-five state-of-the-art `Text-to-SQL` models only on complex questions within the development set of `Spider` as the preliminary step, we found that their performance is below 50%. On questions with corresponding SQL queries of easy and medium difficulty levels, however, such models perform with an accuracy of over 80%. Therefore, we explore to what extent the **decomposition of complex questions**, as a **novel stage** within the `Text-to-SQL` pipeline, can bring us further in the area of `Text-to-SQL`. This is guided by our intuition that by decomposing complex



Question:
Which semester had both Masters and Bachelors students enrolled?

Complex SQL Query:

```
SELECT T2.semester_id FROM Degree_Programs AS T1 JOIN
Student_Enrolment AS T2 ON T1.degree_program_id =
T2.degree_program_id  WHERE degree_summary_name =
"Master" INTERSECT SELECT T2.semester_id FROM
Degree_Programs AS T1 JOIN Student_Enrolment AS T2 ON
T1.degree_program_id = T2.degree_program_id  WHERE
degree_summary_name = "Bachelor"
```

Answer:
semester_id = 13

degree_program_id	degree_program_id	degree_summary_name
1	13	Master
2	2	Master
3	13	Master
4	8	Bachelor
5	11	Master
6	8	Bachelor

Degree_Programs

(a)

Sub-question 1:
Which semester had Masters students enrolled?

Sub-SQL 1:

```
SELECT T2.semester_id FROM Degree_Programs AS T1 JOIN
Student_Enrolment AS T2 ON T1.degree_program_id =
T2.degree_program_id  WHERE degree_summary_name = "Master"
```

semester_id
13

Sub-question 2:
Which semester had Bachelors students enrolled?

Sub-SQL 2:

```
SELECT T2.semester_id FROM Degree_Programs AS T1 JOIN
Student_Enrolment AS T2 ON T1.degree_program_id =
T2.degree_program_id  WHERE degree_summary_name = "Bachelor"
```

semester_id
15
2
13

Sub-question 3:
Which semester both Masters and Bachelors enrolled?

Sub-SQL 3:

```
SELECT semester_id FROM Masters INTERSECT SELECT
semester_id FROM Bachelors
```

semester_id
13

Answer:
semester_id = 13

student_enrolment_id	degree_program_id	semester_id	student_id
4	4	15	9
2	4	2	9
6	3	13	1
11	6	13	4

Student_Enrolment

(b)

Figure 7.1: **(a) An example of a complex question in the `Spider` dataset. In addition to the complex question, the corresponding SQL query, the answer, and tables are shown. (b) The decomposition of the question in (a) is illustrated; Instead of feeding the complex question in (a) into `Text-to-SQL` models, we manually decompose the question into the three sub-questions. These sub-questions are classified as simpler than the original question. Executing sub-questions sequentially on the database, we can observe that answer to the complex question is the same as sub-question 3 in (a).**

questions into multiple easy and medium questions, `Text-to-SQL` models can convert them into correct SQL queries with a higher accuracy, thereby circumventing the challenge of complex SQL generation, illustrated in Figure 7.1.

Note that the proposed decomposition stage is different from standard text simplification in NLP [353], a task in which text is rewritten to make it easier to process for a given audience. The complexity of questions in the `Spider` dataset originates from the underlying SQL query and the dependency between the text and database schema as opposed to the linguistic complexity of NL questions. To verify this, we analyzed whether metrics that are popularly used in text simplification tasks such as *Flesch-Kincaid readability score*, *Flesch's reading ease score*, *Type-Token Ration*, and *Lexical variation* are effective in distinguishing levels of difficulty in complex user questions. We found that easy and medium questions have the same lexical complexity and lexical richness as hard and extra hard questions, confirming that the existing text simplification methods are ill-suited for decomposing complex user questions. In order to assess the feasibility of decomposition, we thereby raise the following research questions:

RQ1: To what extent can we leverage the decomposition of complex user questions as a means to circumvent the challenge of complex SQL generation facing existing `Text-to-SQL` pipelines?

RQ2: To what extent can non-expert crowd workers aid in the decomposition of complex user questions in `Text-to-SQL` tasks in comparison to SQL experts?

To assess the potential benefit of decomposing complex questions, we first manually decomposed the questions and corresponding queries within the development set of the `Spider` dataset serving as an **oracle decomposition**. We then compared the accuracy gained by `Text-to-SQL` models with the new pipeline in which the oracle decomposition was augmented, realizing an increase in accuracy by over 163% (i.e., from 30% to 79%). Despite the promise of decomposition, to develop ML models that can (semi) automate the decomposition of complex user questions in a generalizable fashion, we would require a substantial amount of training data. Since hiring groups of experts is a costly endeavour [355], the viability of decomposing complex user questions at a beneficial scale hinges on its cost-effectiveness. Crowdsourcing has proved to be a reliable, effective, and efficient approach in many tasks [175, 290, 360] and across different domains [366], including within the NLP field [**Zheng_20150**, 195, 422]. Thus, we explored whether non-expert crowd workers (recruited from the Prolific crowdsourcing platform) can power such a cost-effective alternative. In comparison to the accuracy boost of 153% as a result of the decomposition carried out by a small group of SQL experts ($N = 5$), decomposition by non-expert crowd workers ($N = 83$) led to an accuracy boost of over 96%. Our findings show that crowd workers can effectively decompose complex user questions and thereby aid in circumventing the challenge of complex SQL generation in `Text-to-SQL` pipelines.

Our experiments pave the way towards extending crowd-powered decomposition on available `Text-to-SQL` datasets to gather a substantial amount of training data. This is a crucial prerequisite for building ML-based automatic decomposition models integrated into the existing `Text-to-SQL` pipeline to circumvent the challenge of complex SQL generation.

7.2. Background and Related Work

Context Independent Text-to-SQL Parsing

Generating SQL queries from natural language questions has been an active field of study for a long period in both database and NLP communities [10, 184, 261, 265, 332, 426, 438, 458, 459]. Previous `Text-to-SQL` parsers employed either expert-designed rules [393, 397, 437] or statistical techniques [240, 396, 458]. Over the past few years, driven by the development of a large in-domain context-independent dataset, WikiSQL [467], many deep learning models proposed by researchers have shown promising results for this task [168, 183, 386, 467]. All of these studies focus on mapping a single query to the corresponding SQL query which is known as context-independent parsing. Deep learning models generally adapt an encoder-decoder framework to solve the `Text-to-SQL` problem as a sequence-to-sequence problem [112, 113, 234, 386, 467]. To show and test the limitations of the `Text-to-SQL` models on generalizability on various domains and databases, [454] proposed a complex cross-domain dataset called `Spider`. In addition to the sequence-to-sequence paradigm, namely the generation-based methods, state-of-the-art neural models leverage more strategies such as sketched-based techniques (generates a SQL skeleton first and then fills the skeleton with database schema tokens) [83, 111, 183, 201, 279, 280, 443, 456, 467], data augmentation [416, 438], various attentional architectures for question/schema encoding such intermediate representation for decoding [149, 166, 186, 455], graph representation of databases in schema encoding [47, 73, 76], schema linking (correctly identify column and value mentions in an natural language questions and link them to the given database schema) [47, 66, 104, 112, 114, 166, 259, 272, 274, 369, 417, 456].

While there are some attempts to tackle the complex SQL generation issue, it is still a significant challenge for Text-to-SQL models [135, 259, 454]. For instance, schema linking methods by capturing the alignment between text and table indirectly address this challenge. On the other hand, intermediate representation approaches are designed to bridge the gap between text and SQL. Furthermore, some studies have examined decomposing complex SQL queries within the decoder to generate multiple clauses and sub-queries [252, 455]. Unlike these studies, in this work, we investigate the potential performance gain by adding a decomposition stage in the Text-to-SQL pipeline to decompose complex natural language questions before submitting them to Text-to-SQL models.

Context Dependent Text-to-SQL Parsing

Recently, context-dependent Text-to-SQL parsing has drawn a lot of attention. Compared to benchmarks with single-turn questions, ATIS, a simple in-domain context-dependent benchmark, was proposed first. The models evaluated on ATIS leveraged the sequence-to-sequence framework[8]. Later, to overcome the lack of generalizability of models, two large-scale context-dependent datasets were introduced for the Text-to-SQL task, SParC [453] and CoSQL [452] modelling conversational dependencies between questions. The Text-to-SQL models, also known as conversational Text-to-SQL models, require understanding the context of sequentially related questions compared to single-turn models. Several studies were conducted on these two benchmarks that proposed EditSQL [461], IGSQL [68], IST-SQL [423], R^2SQL [199], RAT-SQL-TC [264] models. In addition to employing strategies in the previous section to tackle the problem of translating Text to SQL, these models track dialogue states to generate SQL queries according to the context. [264] conducted an exploratory study within context-dependent parsing to determine how far we are from effective context modelling. In this work, we employed R^2SQL as the baseline to assess the accuracy gain of decomposing complex questions in the Text-to-SQL pipeline. It was the first open-source context-dependent model in the SParC leaderboard [1] at the time of carrying out the experiments in this chapter.

Text-to-SQL Datasets

The growing interest in Text-to-SQL applications has led to various datasets including in-domain datasets ATIS [96, 204], GeoQuery [332, 458], Restaurants [332, 392], Scholar [204], Advising [135], Academic [261], Yelp [444], IMDB [444] which have been studied for decades. WikiSQL is among the first large-scale datasets with relatively simple questions and single tables extracted from Wikipedia. Although WikiSQL contains 80654 questions and SQL pairs for 24241 databases, it is generated from a limited set of templates and only covers the single SELECT column, aggregation, and WHERE clause. Furthermore, keywords like JOIN, GROUP BY, and ORDER BY are not included. The family of Spider datasets, Spider [454], SparC [453], and CoSQL [452] contain the most difficult questions having nested queries, covering many SQL syntaxes, and multiple table joins. These datasets evaluate the Text-to-SQL models to generalize not only new SQL

[1] https://yale-lily.github.io/sparc

queries and database schemas but also new domains. Spider contains 10,181 questions and 5,693 unique complex SQL queries on 200 databases with multiple tables, covering 138 domains. It also supports a wide range of SQL syntax. Thus, we examined our proposed solution on the development set of the Spider dataset.

Data Annotation & Crowdsourcing

Natural Language Processing research has been spurred on by the growing number of annotated corpora [195, 422, 466]. Such corpora are leveraged to train, evaluate, and compare NLP algorithms. However, annotating data is an expensive and time-consuming process [355]. The emergence of crowdsourcing [131] platforms such as MTurk [2] has led to a widespread adoption of crowd-powered workflows to create annotated corpora [40, 74, 93, 146, 159, 254, 290, 310, 322, 331, 357, 376]. Crowdsourcing has been shown to be a cheaper and faster alternative compared to expert annotation [142, 415]. In addition to data labelling, crowdsourcing proved to be a reliable approach in many tasks [175, 290, 360] and domains [366]. Several prior works have proposed methods to improve the effectiveness of crowd-sourced data acquisition [133, 174, 337]. Although much research is conducted to quality control and quality assurance in crowdsourcing [98, 148, 194], several studies have also shown the benefit of employing experts to provide higher quality labels [21, 434]. Prior works have proposed augmenting crowd worker labels with those from experts to optimize the cost and quality of data labelling [222, 341, 368, 446]. We employ both domain experts and crowd workers for data annotation in this work. Our findings suggest the potential benefit of leveraging crowd workers to create training data and then build ML-based decomposition model in the future.

7.3. Gold Standard for Decomposition of Complex Questions

This section introduces the steps for developing a gold standard for the decomposition, creating SpiderDec serving as the **oracle decomposition**. We then employ a Text-to-SQL model to assess the potential accuracy boost by decomposition. Note that the accuracy is measured based on comparing the execution result of each SQL query with the corresponding gold query.

SpiderDec: Extension of the Spider Dataset

In the Spider dataset, data is split into training, development, and a hidden test set. We manually decomposed the questions, and corresponding queries within the development set of the Spider dataset on questions with hard and extra hard SQL queries, thereby creating SpiderDec[3]. State-of-the-art Text-to-SQL models have over 80% execution accuracy for SQL queries with easy and medium hardness levels, while the performance is less than 50% for hard and extra hard SQL queries. So, decomposing hard and extra hard questions into multiple easy and medium questions can lead to a higher accuracy of Text-to-SQL pipeline. For simplicity, we refer to the hard and extra hard

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

[3] https://github.com/sarasal/decomposition

questions of Spider dataset as complex questions. We limited our approach to the development set first to explore the potential benefit of the decomposition task; we leave the annotation of the training set as future work in case of accuracy boost in the Text-to-SQL pipeline. Our rationale behind considering the Spider dataset as a lens to circumvent the problem of complex user questions is governed by the scale and diversity of the Spider dataset compared to others. Furthermore, the cross-domain setup of Spider allows Text-to-SQL models to use different databases for training and testing. Within the Spider dataset, there are 332 complex training examples over 20 databases in total. Each example consists of a natural language question and its corresponding SQL gold query. In the remainder of the chapter, we refer to each instance in the dataset as a pair of the NL question and the SQL query. We annotated Spider development set in two stages: sub-SQL annotation and sub-question annotation. As the complexity of the Text-to-SQL task derives from the underlying SQL queries, we created SpiderDec from a SQL-centered perspective, annotating SQL queries and questions.

SQL ANNOTATION

In the SpiderDec decomposition, we first broke down each complex SQL query into multiple subsequent easy or medium SQL sub-queries. Based on our rubric inspired by prior work [454], each sub-SQL meets one of the conditions in Table 7.1 to be considered as easy or medium. Among 332 pairs of the NL question and the SQL query, 26.8% of SQL queries contain these keywords: *EXCEPT*, *UNION*, and *NOT IN* on which decomposition to medium or easy is not applicable. Therefore, we only decomposed their nested sub-queries into simpler ones and kept the keyword without the necessity of having all the SQL sub-queries with the easy or medium level of difficulty.

Table 7.1: **Criteria to identify whether a SQL query is easy or medium used as a guideline for decomposition.**

Easy or Medium SQL Query	
Condition 1	1) one SELECT column, 2)maximum one aggregator, 3)maximum one keyword from [WHERE, GROUP BY, ORDER BY, LIMIT, JOIN, OR, LIKE, HAVING], 3) no keywords from [EXCEPT, UNION, INTERSECT, IN, NOT IN]
Condition 2	1) maximum two conditions from [number of aggregator > 1, number of SELECT columns > 1, number of WHERE conditions > 1, number of GROUP BY clauses > 1], 2) maximum one keyword from [WHERE, GROUP BY, ORDER BY, LIMIT, JOIN, OR, LIKE, HAVING], and 3) no keywords from [EXCEPT, UNION, INTERSECT, IN, NOT IN]
Condition 3	1) maximum one condition from [number of aggregator > 1, number of SELECT columns > 1, number of WHERE conditions > 1, number of GROUP BY clauses > 1], 2) two keywords from [WHERE, GROUP BY, ORDER BY, LIMIT, JOIN, OR, LIKE, HAVING], and 3) no keywords from [EXCEPT, UNION, INTERSECT, IN, NOT IN]

SUB-QUESTIONS ANNOTATION

Given decomposed SQL sub-queries per SQL query from the previous stage, we assigned a natural language sub-question to each of the annotated sub-queries. In order to determine whether sub-questions are semantically equivalent to their associated complex questions, two experts manually evaluated them and resolved any conflicts with each other.

Evaluation of SpiderDec

To investigate the potential accuracy boost achievable by adding the decomposition stage to the Text-to-SQL pipeline, we are required to measure the performance of Text-to-SQL models on the newly generated dataset. Instead of complex questions, we gave decomposed sub-questions to pre-trained models as input data. We then calculated the execution accuracy gained on the entire development set and separately per hardness category. To this end, we leveraged R^2SQL [199], a context-dependent BERT-based Text-to-SQL model trained on SParC [453] dataset. We then assessed the execution results of predicted sub-SQLs by R^2SQL and compared them with the result obtained from the original development set of Spider (existing Text-to-SQL pipeline). R^2SQL can effectively model contextual questions and database schemas. SParC dataset is built on top of the Spider, providing rich contextual phenomena and thematic relations between the questions. Because the sub-questions are thematically dependent on each other acting as contextual utterances, we adapted the context-dependent Text-to-SQL model, which maps the entire sub-questions to the corresponding SQL queries. Furthermore, the R^2SQL is the first open-source model on the leaderboard at the time of experimenting[4].

7.4. Crowd-Powered Decomposition

We now describe our crowd-powered study in more detail. We go over the annotation tool, the task, participants, the workflow. We then explain our measurement to evaluate participants' decomposition.

7.4.1. Annotation Tool

We developed an annotation tool on top of the R^2SQL pre-trained model for crowd workers to decompose complex questions. We first created a Text-to-SQL API from R^2SQL, translating contextual natural language questions into SQL queries. We leveraged Vue.js JavaScript framework [5] for the frontend and Flask [6] for the backend. Within the annotation process, the question to be decomposed and its associated database are presented to participants. They can easily interact with tables, search an item, sort rows, and scroll them. They can also execute the predicted SQL corresponding to their sub-questions.

In creating SpiderDec, we decomposed SQL queries. We then assigned NL questions to the queries (SQL-centered decomposition), while the crowd workers only access the NL questions and decompose them (question-centered decomposition). In the real-world scenarios, we do not necessarily have the gold SQL queries and labeled data, so we designated our crowd-powered study to investigate the feasibility of decomposing natural language questions and explore to what extent the question-centered decomposition result in accuracy boost compared to SpiderDec.

[4] https://yale-lily.github.io/spider

[5] https://vuejs.org

[6] https://flask.palletsprojects.com/en/2.0.x/

7.4.2. Participants

In our study, participants included SQL experts and non-expert crowd workers. We hired five computer science students with at least two years of experience with SQL as experts. Due to the high cost of hiring experts, we limited the number of experts to five students. Since the number of students was below sufficient samples to carry out statistical comparisons, we assigned them the entire set to gain more insight into their decomposition performance and quality. Each student spent between 12-30 hours for the whole corpus in the development set of `Spider`. According to the institutional regulations, the participants were paid between 22-30 € per hour based on their course credits. In addition to experts, 83 non-experts were employed through the Prolific Academic Platform. With the Prolific platform, we required the participants to (i) have at least 100 accepted Prolific task submissions, (ii) to be native English speakers, and (iii) and have a minimum approval rate of 90%. The study took approximately 50 minutes to decompose six complex questions. We also paid our participants 7.5 £ (9 €) per hour for the experiment. For simplicity, we refer to Prolific participants as non-experts in the remainder of the chapter.

SQL Knowledge

We measured the SQL knowledge of participants in a post-test conducted right after the decomposition task to avoid cognitive biases [117]. To this end, we manually designed our survey as no standard SQL assessment test is available in the literature. The survey took 10 minutes to complete and consists of 10 questions. First, participants were asked one question to self-report on their SQL proficiency, followed by seven questions regarding the key concepts of SQL [7]. Inspired by prior work [378, 389, 429], we employed the modified VKS test to measure participants' knowledge across four levels. Our questions are related to key concepts of SQL, which are used in our dataset, including relational databases, primary key, foreign key, SELECT statement, WHERE clause, JOIN tables, and Aggregate functions. Participants were asked to write their concept definitions for levels (3) and (4). Finally, participants were given a simple question, *Write a SQL query that returns the name of the 3 youngest winners across all matches found in the table matches.*, with a schema of the database to write down a SQL for. This question helps us to investigate their knowledge in practice.

(1) I don't remember having seen this term/phrase before.

(2) I have seen this term/phrase before, but I don't think I know what it means.

(3) I have seen this term/phrase before, and I think it means ___ .

(4) I know this term/phrase. It means ___.

English Proficiency

In addition to SQL proficiency, we hypothesized that the participants' proficiency in reading and writing could affect their performance. Decomposition a NL question first requires understanding the questions -associated with participants' reading skill- and

[7] https://www.interviewbit.com/sql-interview-questions/#sql

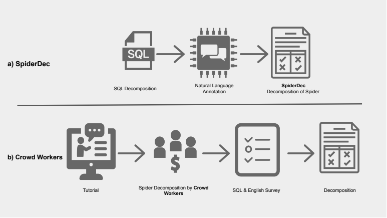


Figure 7.2: **Overview of the flow of the user study and `SpiderDec` creation**

then paraphrasing them in multiple sub-questions- connected with participants' writing skill. Therefore, we leveraged the self-assessment grid of CEFR scales, Common European Framework of Reference for Languages: Learning, Teaching, Assessment [8]. CEFR is an international standard describing reading, listening, speaking, and writing skills on a six-point scale, starting from A1 as a beginner to C2 as a master. Our task was only dependent on participants' reading and writing skills; we, therefore, included the self-assessment questions of these skills. As we included native English speakers in the study, we assumed their knowledge level is above A2 and excluded A1 and A2 from the options. In total, the participants answered three questions related to their reading and writing skill in English.

7.4.3. Procedure

When participants entered the study, a 15-minute tutorial video provided information about their task and how to interact with different annotation tool components to decompose questions. All explained concepts were simplified, avoiding any technical burden for participants. Furthermore, the study was also elaborated on two examples within two stages: 1) how to decompose a complex question into multiple sub-questions and 2) how to work with the annotation tool. Subsequently, the participants moved on to the training phase, where they were given those two examples again to work with the annotation tool and learn the task in practice. Participants could stay in this stage as long as they wish to. Participants were then redirected to the actual decomposition task by clicking the respective button in the training stage— all 332 complex questions from the development set of `Spider` were randomly assigned to experts, while non-experts were given six questions. They could also skip a question if they were not certain about how to decompose it.

The experiment ended with a post-test where the SQL knowledge survey and English proficiency self-assessment were given to participants. We set these surveys as the post-test to avoid cognitive biases [117] such as *Anchoring Effect*-where the participants may overlay focus on answering the survey question rather than the actual task-, *Overconfidence or Optimism Bias*- where the participants overestimate their ability to perform the task when they can answer all the questions in the survey-, and *Loss Aversion Bias*- when the participants suspect that the answers to the questions may affect their payment. Lastly, we included five questions regarding the annotation tool and tutorial,[9]

[8] http://ebcl.eu.com/wp-content/uploads/2011/11/CEFR-all-scales-and-all-skills.pdf
[9] https://www.ueq-online.org

their experience working with our annotation tool, and the perceived performance. The workflow corresponding to data annotation by participants is illustrated in Figure 7.2.

7.4.4. Evaluation of Decomposition

In total, experts provided us with 1515 sub-questions. These sub-questions are associated with 623 decomposed questions, indicating that each question on average contains 2.43 sub-questions. Non-experts created 1082 sub-questions in total for 453 decomposed questions, showing on average 2.37 sub-questions per question.

Table 7.2: **Example evaluation of a participant's sub-questions**

Question: What is the country with the most number of TV channels and how many does it have?	
Correct	Which country has the most number of TV channels? What is target country and how many TV channels does it have?
Partially Correct	Which country appears in the list of TV channels the most times? How many TV channels does this country have?
Incorrect	What are TV channels in the countries?

Assessment of Sub-questions

We manually evaluated whether sub-questions were semantically equivalent to questions or not. Among 1515 sub-questions provided by experts, we randomly sampled 312 questions with the confidence interval of 95% from the population size while we evaluated all sub-questions generated by non-experts. As we assured the quality of data generated by experts, we randomly sampled experts' sub-questions rather than checking all of them. We labeled the entire block of sub-questions as either correct, partially correct, or incorrect. Examples of sub-questions and the labels we assigned to them are provided in Table 7.2. We employed the following criteria to judge the equivalency of sub-questions to the original complex question.

- ***(2)*** **Correct.** If a participant's sub-questions include all concepts that appear in a question, it indicates that the question is semantically equivalent to the sub-questions. In that case, the entire block of sub-questions is assigned the highest score of **2**.
- ***(1)*** **Partially Correct.** If the participant's decomposition misses one concept from the original question, a score of 1 is given to that. For example, if the question asks about the *name* and *birth date* and the participant only included *name*, we labelled the decomposition as 1.
- ***(0)*** **Incorrect.** Sub-questions that are either entirely incorrect, incomplete, or missed more than one concept from the original question.

Assessment of Sub-SQL's

To examine whether participants' decomposition leads to the `Text-to-SQL` pipeline performance boost, we employed our baseline, R^2SQL, which is also in line with our approach in Section 7.3. First, the baseline predicted the sub-SQLs. We then compared the execution result of the block of sub-questions with the execution result of the gold

Table 7.3: **Accuracy of the `Text-to-SQL` pipeline on `Spider` and `SpiderDec` reported on hard and extra hard questions.**

	Dataset	Total	Hard	Extra
I	`Spider`	0.3	0.37	0.23
II	`SpiderDec`	0.79	0.82	0.76
III	Diff.	0.49	0.45	0.53

Table 7.4: **Number of correct SQL predictions out of 332 complex questions on `Spider` and `SpiderDec` in the `Text-to-SQL` pipeline.**

	Dataset	Ques.	# Hard	# Extra
I	`Spider`	100	62	38
II	`SpiderDec`	265	138	127
III	Diff.	165	76	89

SQL query. We labeled the sub-SQL's either correct or incorrect. The block of the sub-SQL's is correct if its execution result is equivalent to the execution result of the gold SQL; otherwise, it is incorrect.

7.5. Results

Performance of the baseline on SpiderDec (RQ1)

In RQ1, we examine to what extent the decomposition of complex questions affects the performance of the `Text-to-SQL` pipeline facing the challenge of complex SQL generation. We return to Table 7.3 and Table 7.4 for insights into the performance of the R^2SQL (cf. Section 7.3) over the original dataset and decomposed set which is elaborated on each difficulty level.

The baseline predicted 100 questions correctly out of 332 questions in the development set, 62 questions among the hard and 38 questions from the extra hard division. On the other hand, after decomposing complex questions, the model predicted 165 more correct questions leading to 265 questions in total, including 138 hard questions and 127 extra hard questions (cf. Table 7.4).

By comparing the execution accuracy of the baseline on `Spider` and `SpiderDec`, we observed that the accuracy on complex data raised from 0.3 to 0.79, (cf. Table 7.3). We note that the contribution of decomposition on performance gain to each division of data is nearly the same, with 76 and 89 more correct questions for hard and extra hard, respectively.

In Section 7.3, we discussed that sub-questions for keywords NOT IN, EXCEPT, and UNION are still hard, being less difficult than the original question. To gain insight on the impact of decomposition on complex keywords, we focus on Table 7.5. For the NOT IN keyword, the baseline predicted 18 questions out of 46 from the original `Spider` while this number increased to 38 considering the `SpiderDec`. Similarly, this number raised from 4 to 21 for the EXCEPT keyword, and from 0 to 5 for the UNION. Although the decomposition did not lead to the sub-questions with an easy or medium difficulty level for these complex data, the baseline outperformed significantly on `SpiderDec`. Given these observations, we can see the benefit of decomposition on all types of complex questions.

Looking deeper, we also examined the cases where the `Text-to-SQL` model failed to predict the correct SQL query even after the decomposition had applied, which is 67 questions in total. We classified the majority of errors into two groups. Table 7.7 shows more examples for each category. This is understandable since the decomposition task only simplifies questions by breaking them down into multiple questions. As mentioned

Table 7.5: **Distribution of complex keywords in `Spider` dev. set. The number of questions predicted correctly in the `Text-to-SQL` pipeline w/o decomposition is reported.**

	Keywords	Total	`Spider`	`SpiderDec`
I	NOT IN	46	18	38
II	EXCEPT	32	4	21
III	UNION	11	0	5
IV	Total	89	22	64

Table 7.6: **Errors generated by experts and non-experts**

Error Type	Experts	Non-Experts
Complex Sub-Questions	5	49
Missed Final Sub-Questions	2	24
Missed One Keyword	8	19
Different Interpretation of Questions	5	0
Other	7	44
Total	27	136

earlier, as it does not add any additional knowledge to sub-questions, they do not contribute to any solutions for the following issues.

- **Implicit Column Names:** Within this group of questions, some of the column names in the SQL query are implicitly mentioned in the question, so the `Text-to-SQL` model requires to infer them. For instance, we have this question. *Which airlines have departing flights from both APG and CVO airports?* The column *SourceAirport* should be inferred from the phrase *departing flights*
- **General Knowledge or Table Content:** This group of questions includes one or multiple values of the tables. Sometimes these table values are considered general knowledge. Within this example, *What is the name of a country that has the shortest life expectancy in Asia?*, *Asia* is the continent, so the model needs to know this general knowledge or recognize it as the table content.

Table 7.7: **Example of errors remain after adding the decomposition stage to the `Text-to-SQL` pipeline**

Implicit Column Names	
Src. or Dest. Airport	Which city has most number of *departing flights?* `SELECT T1.City FROM airports AS T1 JOIN flights AS T2 ON T1.AirportCode = T2.SourceAirport GROUP BY T1.City ORDER BY count(*) DESC LIMIT 1`
Current Address	What are the last name of the students who *live in* North Carolina `SELECT T1.last-name FROM Students AS T1 JOIN Addresses AS T2 ON T1.current-address-id = T2.address-id WHERE T2.state-province-county="NorthCarolina"`
General Knowledge or Table Content	
Continent	What is the name of country that has the shortest life expectancy in *Asia*? `SELECT Name FROM country WHERE Continent = "Asia" ORDER BY LifeExpectancy LIMIT 1`
Language	Which cities are in European countries where **English** is the *official language*? `SELECT T3.Name FROM country AS T3 JOIN countrylanguage AS T4 ON T3.Code = T4.CountryCode WHERE T4.IsOfficial = "T" AND T4.Language = "English"`

Performance of Crowd Workers on Decomposition Task (RQ2)

RQ2 investigates to what extent crowd workers can decompose complex question compared to the oracle decomposition. We first report the result of SQL knowledge survey

Table 7.8: **Decomposition performance of experts and non-experts reported in percentage**

		Total	Hard	Extra
I	Expert	61.8	64.9	59.5
II	Non-Experts	48	55.6	40

Table 7.9: **Accuracy of the `Text-to-SQL` pipeline on `Spider` dev. set and decomposed data created by experts and non-experts**

		Total	Hard	Extra
I	Spider	0.3	0.37	0.23
II	Experts	0.76	0.75	0.77
III	Non-Experts	0.59	0.69	0.5

and English proficiency. Then, we examine to what extent the decomposition leads to an accuracy boost with decomposition compared to existing `Text-to-SQL` pipeline.

In the SQL knowledge survey, all five SQL experts assessed themselves as level (4). By manually evaluating the concept definitions, we verified that all of the answers were correct and the experts had sufficient SQL knowledge to carry out the task. In terms of reading and writing skills in English, all experts had the highest levels, C2.

In total, 83 non-experts provided us with 67 SQL concept definitions when self-assessing their knowledge as level (3) or (4). All 67 definitions were labeled as incorrect. This result suggests that our non-experts group indeed did not have any background knowledge in SQL. Evaluating non-experts' English proficiency in reading and writing, we found that their skills were distributed within the level of B1 to C2. We observed 26.8% of non-experts with level B (B1, B2) and the remaining 73.1% with level C (C1, C2) in reading. In writing, we reported these numbers as 29.2% and 70.7% with level B and level C, respectively. Regarding the demographic data, among non-experts, 35.7% were female, and 62.4% were male. The ean age of participants was 32, with a minimum of 18 years and a maximum of 55 years.

Table 7.8 illustrates the decomposition performance of experts. We applied our decomposition approach to the hard and extra hard division of the `Spider` development set and evaluated the experts' performance in each division separately. Experts were able to decompose 61.8% of questions correctly contributed to 64.9% on hard division and 59.5% on extra hard. Although the performance on hard division is higher than the extra hard, the low difference between these two numbers suggest that the difficulty of questions does not impact the experts' decomposition performance. On the other hand, we reported that non-experts decompose 48% of questions, with 55.6% and 40% separately on hard and extra hard questions. As the performance of non-experts on hard questions is higher than extra hard questions, we can see that non-experts perceived the extra hard question as more difficult than the hard ones.

In addition to decomposition performance, we examined the potential benefit of experts' decomposition on `Text-to-SQL` pipeline accuracy. The accuracy on the original development set of `Spider` is calculated as 0.3, particularly 0.37 and 0.23 on hard and extra hard questions. Table 7.9 presents the accuracy of the baseline on the `Spider` and the decomposed questions by crowd workers. In terms of the accuracy boost, experts' decomposition led to 0.76 accuracy on complex questions split to 0.75 and 0.77 for hard and extra hard. We can also see that experts contributed more to improving the accuracy on extra hard questions from 0.23 to 0.77 (+0.54). Non-experts decomposition also prompted 0.59 accuracy on the complex questions, with 0.69 and 0.5 accuracies on hard and extra hard questions. Furthermore, we found that experts outperformed non-experts (accuracy 0.76 vs. 0.59), which is also in line with our finding regarding the decomposition performance. In contrast to experts, non-experts impact more on hard data, with 0.32 and 0.27 boost on hard and extra hard, respectively. Experts decomposition remarkably increased the accuracy for extra hard questions while non-experts decomposition contributed more to hard questions. In other words, the results demonstrate that when question difficulty increases, non-experts' performance deviates from the experts.

Taking these analyses further, we can see that the performance of experts (0.76) is in line with the accuracy boost achieved by `SpiderDec` (0.79). `SpiderDec` is created according to guideline in 7.3 based on SQL gold query while decomposition by experts only applied on natural language questions.

In terms of English proficiency, we also found that the reading and writing skill of non-experts significantly affected their decomposition performance. As the number of experts was limited, we only analyzed the impact of reading and writing factors in the non-experts group measured by a two-way ANOVA test. The test considered reading and writing as factors; the main effects were examined where $\alpha = 0.05$. For post-hoc analysis, the Tukey HSD pairwise test was used. We found that non-experts with level B1 in writing and reading significantly had lower performance than other levels. These results suggest that we can gain higher performance if we pre-screen the participants and reject those with reading and writing skills of B1.

Among the data created by experts, 37.6% of the provided decomposition were identified as errors, 6% were partially correct, and the remaining 31.4% were incorrect. On the other hand, among the decomposed questions created by non-experts, we observed 51.8% of the decompositions were error, with 13% and 38.8% were labeled as partially correct and incorrect, respectively. Taking this analysis further, we also examined different types and frequencies of errors produced by experts and non-expert, illustrated in Table 7.6. We subjectively categorized the error types into two groups: recoverable errors and costly errors.

- **Recoverable Errors**: These are errors that can be fixed through a relatively simple post-hoc analysis without modifying the decomposed queries substantially, either through expert intervention or through algorithmic interventions.
- **Costly Errors**: These errors cannot be fixed easily through post-hoc analysis without modifying the decomposed queries significantly. Experts would need to rewrite one or more complete sub-queries to fix such errors.

We also manually checked the errors and classified them into five groups. We first introduce these types. We then determine which of them are recoverable and which are costly. Examples for each categories are shown in Table 7.10.

- **Complex Sub-Questions**: Sub-questions in this group have the same difficulty level as the corresponding questions. Participants cannot identify how to break down the questions to make it less difficult, so they only paraphrase the question or write down sub-questions as complex as the questions. This type of error can be easily detected automatically by comparing the difficulty level of the sub-SQL queries to the gold SQL queries. However, it is a costly error. An expert is required to revise this decomposition or rewrite it from scratch, having monetary and time costs. As expected, this type of error mainly occurred for non-experts as they do not have sufficient knowledge to determine how complex their sub-questions are. An example, in Row **I** in Table 7.10, sub-questions are only the paraphrased form of the question. Converting them to SQL, their SQL queries are as complex as the question. Alternative sub-questions could be (i) which student owns a cat as a pet? (ii) which students are not among them? (iii) return their age and major.

- **Missed Final Sub-Questions**: Sub-questions in this category are nearly correct, only missing the last sub-question required to return the target result set. By comparing the execution result of the question and the sub-questions, we can determine the existence of some errors. However, it is difficult to identify whether the errors fit in this group. An expert is needed to identify this error category manually. Similarly, the experts should write down the final sub-question. Therefore, this error type is among costly errors, while their recovery leads to remarkable performance gain. This error is more frequent among non-experts than experts which also intuitively make sense. To resolve the sub-questions in the Row **II**, we need to add this sub-question: which airports are not among those lists? Find their name.

- **Missed One Keyword**: Sub-questions are one keyword away from the related questions. When Participants created decomposition data, they skipped or modified one keyword of the questions in their sub-questions. This type of error could mistakenly happen, or participants may not understand the role of the keyword, so they did not take that keyword into account. This error can be automatically discovered and revised by adapting an attention model to identify keywords. So, they are recoverable errors. This group of errors is the most frequent error type among experts. The sub-questions in the example of Row **III** skipped the word *Currently*. According to the tables associated with the question, students both have a current and permanent address. So, the term *Currently* plays an important role in distinguishing which columns to return.

- **Different Interpretation of Questions**: For some questions, participants interpret them differently from what existed in the gold standard. Although these interpretations are valid, we mark them as an error because of different execution results with the corresponding questions. As resolving this error requires rewriting the sub-questions, we classify them as a costly error. Only Experts generate this type of error. Looking into the example of Row **IV**, we can see that the word *predominant* can be interpreted differently. Does it mean that the language is official? Does it mean the language is spoken with the highest percentage? Although both interpretations could be correct, the second meaning is incorporated into the dataset.

- **Other**: Sub-questions within this error category are partly correct or thoroughly incorrect. Participants' sub-questions are not satisfied with the condition of being semantically equivalent to the corresponding questions. Finding and resolving such errors are not only time-consuming, but also they require the cost of expert interventions.

Table 7.10: **Examples of error generated by experts and non-experts**

Examples	
I. Complex Sub-Questions	**Question:**What major is every student who does not own a cat as a pet, and also how old are they? **Sub-Questions** • How old are the students who do not own a cat as a pet? • What major is each student which does not own a cat?
II. Missed Final Sub-Questions	**Question:** Find the name of airports which do not have any flight in and out? **Sub-Questions** • Which airports are source airports? • Which airports are destination airports?
III. Missed One Keyword	**Question:** Find the last name of the students who currently live in the state of North Carolina but have not registered in any degree program. **Sub-Questions** • List the address ids of all addresses in North Carolina. • Find the student whose address corresponds with the list in North Carolina. • Find which of these students have not enrolled in any degree program? • Find the surnames associated with these students who have not enrolled and live in North Carolina.
IV. Different Interpretation of Questions	**Question:** Count the number of countries for which Spanish is the predominantly spoken language. **Sub-Questions** • What countries speak Spanish? • Given Spanish countries, count the number of officials.

7.6. Discussion

7.6.1. Implications of Our Work

This study shows that decomposition leads to the accuracy boost for the `Text-to-SQL` pipeline on complex questions. This can begin to shed light on the influence of decomposition as a promising approach in improving the accuracy of `Text-to-SQL` tasks. As the follow-up study, we can build a fully automatic ML-based decomposition model integrated into the existing `Text-to-SQL` pipeline. For training such a model, it is crucial to collect a substantial amount of labeled data, such as decomposition of the family of `Spider` dataset. Our findings support the evidence of employing crowd workers for this task as a scalable method. According to our error analysis in 7.5, a fully automatic decomposition model might face several challenges. Among five error groups discussed in 7.5, the major challenges for automatic decomposition can be error types of *Complex Sub-Questions* and *Missed One Keyword*. Circumventing the challenge of *Complex Sub-Questions* is difficult since verifying the sufficient level a question should be broken down is difficult for even a human. There is a trade-off between the granularity of sub-questions and their complexity. The fine-grained the sub-questions, the less complex sub-questions we have. However, we might oversimplify questions that are not required at all. In terms of *Missed One Keyword* challenge, we demand an attention model to identify the keywords within the questions and evaluate whether those keywords existed in the sub-questions. Many of these keywords are dependent on the schema of tables. However, the attention model in training would not be enough.

7.6.2. Caveats and Limitations

We had an imbalance number of SQL experts, restricting us from gaining insight into their performance individually and employing any statistical tests. We also did not consider workflows to optimize decomposition such as aggregation of crowd-workers' answers and double-checking their decomposition answers by experts, which means that it would be possible to achieve higher accuracy than what we observed in our work when optimized. Furthermore, we only leveraged one `Text-to-SQL` model in our study. Although our decomposition approach is independently defined of any `Text-to-SQL` models, a comprehensive analysis of state-of-the-arts `Text-to-SQL` models can give us a better insight into the impact of decomposition on different models.

7.7. Conclusions and Future Work

This chapter explores the feasibility of decomposing complex user questions within the `Text-to-SQL` pipeline as a means to circumvent one of the significant shortcomings of `Text-to-SQL` models in complex SQL generation (RQ1). We first adapted the decomposition on the development set of the `Spider` dataset, breaking complex questions down into simpler sub-questions in a way that `Text-to-SQL` models can convert them correctly to corresponding SQL queries. We then investigated the feasibility of leveraging crowd workers to produce sufficient training data for building a ML-based model decomposing complex questions automatically (RQ2).

We defined the decomposition task for complex questions in which a complex question is split into multiple subsequent sub-questions. Having assessed the decomposition approach on complex questions in `Spider` dev. set (`SpiderDec`), we found that the accuracy raised remarkably from 30% to 79%. Our results support the evidence of decomposition as a promising approach to boost the performance of existing `Text-to-SQL` pipelines.

We then examined the performance of 88 crowd workers on decomposing the natural language questions within the development set of `Spider`. Compared to the accuracy boost of 153% (30% to 76%) as a result of the decomposition carried out by a small group of SQL experts ($N = 5$), decomposition by non-expert crowd workers ($N = 83$) led to an accuracy boost of over 96% (30% to 59%). Our findings show that crowd workers can effectively decompose complex user questions and thereby aid in creating training data at a beneficial scale for generalization of decomposition in `Text-to-SQL` pipelines.

8

Conclusions

This thesis contributes to understanding the various task-related contextual factors and stages involved in decision-making, highlighting their significant impact on decision outcomes and individuals' behavior. The insights from this research can guide future empirical research to refine their methodologies, improve the design of decision-making processes, and develop interventions and AI systems that enhance decision outcomes. Continuous assessment of socio-technical systems allows researchers, designers, and practitioners to pinpoint areas for improvement and refine current methods and frameworks to adapt them to particular decision-making contexts and the complexity of real-world situations. This can ultimately lead to complementary human-AI collaboration and the development of decision support systems that leverage the strengths of both humans and AI, resulting in improved decision outcomes. In our effort to delve into decision-making environments, we have underscored the significance of rigorous and replicable research by presenting frameworks and approaches for examining decision-making procedures.

In this concluding chapter, we revisit our research questions to summarize our key findings and contributions. We then discuss the implications of our research from technical, methodological, and theoretical perspectives, followed by the limitations of our work. Finally, we highlight future research directions that can further build upon these theories and methodologies to advance the field of human-AI decision-making.

8.1. Summary of Findings

With the findings and discussion presented in three parts of this thesis, we revealed the importance of considering task-related contextual elements and group settings in decision-making scenarios. Adapting empirical research, AI system design, and intervention strategies according to these factors can lead to a better understanding of decision processes and improved decision outcomes. This section outlines the key findings and contributions of this research based on three main themes:

Part I: Understanding the Role of Task Context

In the first part of this thesis, we sought to *evaluate task context in human-AI decision-making* (**RQ1**), specifically by narrowing down our focus to task complexity and uncertainty. We proposed a conceptual framework for evaluating decision-making contexts through the lens of complexity to enable comparison and evaluation of empirical studies. Using this theoretical framework, we compared various decision-making tasks in the literature and synthesized current patterns and challenges. We applied the framework practically, along with an additional task-related contextual factor - uncertainty - to explore differences in decision-making behavior and outcomes across diverse real-world settings.

Using the framework outlined in Chapter 2, we discovered that tasks with lower levels of complexity share common contextual factors that can impact decision-making results. These tasks are typically low-stakes with no prior domain knowledge required. On the other hand, complex tasks tend to be high-stakes and resemble some real-world decision-making scenarios requiring specialized knowledge and expertise. While comparing relatively simple tasks through the lens of complexity is beneficial, it is equally important to adapt the framework to address the nuances and challenges associated with more complex decision-making scenarios where there is greater potential for human reliance. Furthermore, our research highlighted a significant theoretical and empirical gap in studying task complexity within human-AI decision-making processes. Existing studies primarily focus on simpler scenarios, overlooking complex decision-making tasks that also reflect real-world challenges.

In Chapter 3, we empirically assessed how task complexity and uncertainty affect individuals' behavior and decision outcomes through a series of experiments conducted in a real-world context. Our results indicate that complex and uncertain tasks can lead to sub-optimal outcomes, especially when people over-rely on AI systems without carefully assessing their suggestions. We also found that despite the influence of complexity and uncertainty on individuals' behavior, their level of trust in AI systems remained relatively stable across various decision-making contexts. These findings emphasize the need to consider the specific characteristics of decision-making tasks when designing empirical research and developing decision-support systems tailored to the needs and challenges of complex decision-making environments. We also proposed our configurable framework as a starting point for future empirical research, allowing for customization and adaptation to different decision-making contexts and settings, and increasing the reproducibility of the findings.

Part II: Adding Groups to the Mix: Human-AI group Decision-Making

In the second part of this thesis, we focused *on the impact of task-related contextual factors, task complexity, and uncertainty on group decision-making processes and outcomes* (**RQ2**). We particularly explored the effects of AI systems on group dynamics, including coordination, communication, consensus-building, efficiency, and overall decision quality. Furthermore, we investigated how task complexity and uncertainty influence the usage of AI systems in group settings and how they differ from individual decision-making. To this end, we conducted experiments in group decision-making scenarios with varying levels of task complexity and uncertainty.

Chapter 4 demonstrates that task complexity and uncertainty significantly affect group AI decision-making processes, leading to lower performance in high-complexity or uncertain situations. Compared to individual setups, AI systems have also shown a positive effect on group performance, especially in more complex tasks. The results show that integrating AI systems into group decision-making can improve outcomes by fostering collaboration and discussion among group members, leading to well-informed decisions. We also found that the efficiency of groups increased in complex tasks, yet the overall decision accuracy remained lower compared to simpler tasks. Conversely, in uncertain scenarios, the performance and efficiency of groups declined, underscoring the necessity for a more nuanced design approach when incorporating AI systems into group decision-making.

We should understand and analyze contextual factors before integrating AI systems into decision-making processes, as the success of AI integration depends on the specific task complexity and level of uncertainty. Blindly adopting AI without considering specific contextual elements may impede effective decision-making and lead to suboptimal outcomes. Our study also emphasizes the need to provide tailored support and interventions to help groups navigate decision-making tasks effectively amid uncertainty and complexity. Future research should consider the context within which group decision-making occurs, as these factors greatly influence the success of interactions between groups and AI systems in decision-making processes.

Part III: Improving Information Access: The cases of Web Search and Databases

In the final part of this thesis, we focused on *improving information access, specifically through search engines and databases* to ultimately enhance decision-making processes(**RQ3**). This part focuses on information access that enables decision-making, rather than on decision-making directly. We proposed using entity cards to summarize relevant information from search results, which search engines can integrate into their interface. We designed and evaluated various entity card variations and assessed their impact on users' learning and behavior change through an empirical user study. Additionally, we suggested utilizing a decomposition approach to enhance the performance of natural language interfaces for databases in complex situations without modifying the underlying systems. Then, we assessed the viability and efficiency of this approach through a series of experiments.

Our findings from Chapter 7 indicate that entity cards can have a significant impact on users' learning, particularly for simpler subjects. It has been demonstrated that low-quality entity cards can have a negative effect on users' learning outcomes, underlining the importance of designing accurate and informative entity cards. Additionally, we observed that entity cards play a significant role in shaping user engagement and behavior during the information-seeking process. Users tend to spend more time interacting with and exploring both the interface and search results than using a traditional interface. Furthermore, we found that entity cards assist users in refining their search queries and exploring related concepts.

Chapter 8 has focused on enhancing natural language interfaces for databases

(NLIDB) by employing the decomposition approach. This approach can significantly improve NLIDBs' performance in complex real-world scenarios, enabling a wide range of users to access accurate information efficiently. By evaluating the feasibility of an automated decomposition mechanism, we showed that utilizing a crowd-based approach harnessing collective intelligence can successfully produce training data for building agnostic decomposition models and improve the overall accuracy and effectiveness of NLIDBs regardless of their underlying structure.

Overall, these findings highlight the importance of customizing and optimizing user interfaces in complex scenarios to facilitate information access. By integrating entity cards into search engine interfaces and improving the performance of NLIDBs through decomposition, we can enhance the information access and ultimately enable users to make more informed decisions.

8.2. Implications and Methodological Insights

The key implication of this research is encouraging the rigorous design of future research studies to better understand the underlying mechanisms and factors that influence decision-making outcomes, ultimately improving decision-making in various contexts. It also provides valuable insights for research communities, designers, and practitioners, which can be utilized to enhance the development of interventions and AI systems that aid decision-making in accordance with the contextual needs of decision-makers. We outline several key implications derived from this research:

Challenges of Decision-Making in Complex Scenarios

In our study of complex contexts, we have discovered that making decisions and searching for information can be especially demanding. Complexity in such environments stems from a range of factors, including the unpredictable nature of the environment, the multitude of available information sources, and the need to consider multiple perspectives and resources in order to make informed decisions. Limited cognitive resources of individuals may further exacerbate these challenges and hinder the ability to navigate complex decision-making contexts effectively. In these scenarios, individuals often heavily rely on AI systems and interfaces to help them acquire, process, and organize relevant information without critically evaluating its quality or considering alternative perspectives. This is primarily due to the fact that using these tools incurs lower costs compared to manually evaluating and processing all the accessible information, often resulting in sub-optimal results. While integrating AI systems and adapting user interfaces can offer potential solutions, further research is needed to explore tailored approaches and interventions that can effectively support decision-making in complex scenarios. These strategies should aim at reducing cognitive overload associated with evaluating given suggestions and improving information comprehensibility and accessibility while promoting critical thinking and consideration of alternative perspectives.

We also sought to identify and define elements that may increase the complexity of decision-making situations. In particular, we explored various aspects of task complexity, uncertainty, and topic intricacy. While these factors have had a significant impact on decision-making procedures and information access, we acknowledge that further research is necessary to identify other elements and dimensions of complexity that may influence decision-making outcomes. For instance, the complexity of decision-making could also stem from factors such as the dynamic nature of the decision context, the cognitive load of decision-making, the interdependence of different variables, and the time limitations placed on decision-makers. Future research should investigate the effect of these factors on decision-making outcomes and explore strategies to mitigate their negative impact. Furthermore, we focused on the complexity from the perspective of decision context rather than focusing on individual decision-makers. By examining the perceived complexity of decision-making contexts in future work, we can better understand the challenges faced by decision-makers and develop strategies to adapt their decision-making processes accordingly.

Integration of AI Systems into Different Contexts

The integration of AI systems into decision-making scenarios is intended to enhance the outcome of decision-making processes by capitalizing on the unique strengths of humans and AI systems. The benefits of incorporating AI into group decision-making processes are apparent in complex situations, as AI systems can effectively augment the collective intelligence and diverse perspectives of group members to enhance decision-making outcomes. These results indicate that the advantages of AI systems differ significantly based on the particular task environment. Therefore, successfully integrating them into decision-making scenarios depends on an in-depth understanding of the specific challenges and needs within each setting.

Before integrating AI systems into a decision-making scenario, it is crucial to carefully consider these three questions: Is there a need to add AI systems to the decision-making process? What are the potential benefits and risks of incorporating AI into the context? How can AI be effectively integrated into decision-making to improve outcomes and address the unique challenges individual decision-makers face? Answering these questions requires careful consideration of the potential role of AI systems in each specific context and how they can be adapted to meet the unique needs and constraints of each situation. It is also essential to contemplate what constitutes an appropriate AI system for a given context and how to optimize the interaction between humans and AI systems to achieve desired outcomes. This optimization can be employed by developing appropriate AI algorithms, designing user-friendly interfaces, and providing clear guidelines and training for individuals tailored explicitly to the context and needs of decision-makers.

Towards Rigorous Empirical Studies in HCI

Our retrospective review revealed a lack of systematic research on identifying and analyzing contextual factors in human-AI decision-making scenarios. While we briefly addressed task-related contextual factors like task complexity, task uncertainty, and group dynamics, future studies should recognize additional specific contextual factors that could impact decision-making behavior. Our findings highlight that these factors should be considered and controlled for in rigorous empirical studies to ensure accurate and reliable results. Neglecting to account for these variables may lead to confounding effects and erroneous conclusions in human-AI decision-making research. Understanding the effect of these factors can also contribute to shaping the design of AI systems that are tailored to specific decision-making contexts and user needs.

Conducting rigorous field studies is essential for assessing the success of human decision-making in real-world settings. Our retrospective review revealed that most existing studies involve relatively simple decision scenarios, often based on hypothetical settings, which deviate from the complexities of actual decision-making contexts. This lack of ecological validity limits the generalizability and transferability of findings, undermining their practical significance in real-world decision-making. To address these limitations, we designed our studies to examine decision-making in a naturalistic setting, considering the complexities and uncertainties that decision-makers face. Future research should focus on conducting studies that closely replicate real-world decision-making scenarios to improve the applicability and relevance of research findings.

The reproducibility issue in HCI research also highlights the need for rigorous empirical studies in decision-making situations where we clearly define the relevant variables and factors that can affect decision outcomes. Researchers should strive to provide transparent methodologies, detailed descriptions of experimental conditions, and open access to data and analysis code. This can enhance the credibility and reliability of research findings and promote replication and further exploration of decision-making processes on top of existing knowledge. In our efforts to address these concerns, we proposed our modular and configurable framework for conducting such studies to encourage rigorous empirical studies in real decision-making contexts. This framework serves as the groundwork for researchers to design and implement robust experiments and extend them to new domains and contexts.

8.3. Limitations and Future Work

Despite our efforts to design rigorous empirical studies, we acknowledge the limitations of our work. Future research can address these limitations and further explore and refine methods for studying decision-making in real-world settings.

More contextual elements to be investigated. In this thesis, we assessed how contextual factors, task complexity, and uncertainty affect decision-making outcomes. However, some limitations need to be addressed in future research. Although we operationalized these factors using real-world data, alternative approaches or measurements could offer additional insights. For instance, task uncertainty can be determined by various elements, including resource availability, time constraints, and limited understanding of decision implications. Future research should explore integrating these various operationalizations into the study design and analysis to gain a better understanding of decision-making processes. We narrowed our task scenarios to low-risk contexts, which may not fully capture the challenges and decision dynamics in high-stakes situations where complexity and uncertainty are amplified. Future empirical studies should explore our research questions in high-stakes contexts to better understand the strategies and cognitive processes involved. Our studies did not explicitly examine several task-related elements, such as stake, time pressure, information overload, information loss, cognitive load, and decision-maker expertise level. While it is not possible to investigate and control all potential factors in a few studies, future research should explore these additional variables and their interplay to inform the design of tailored decision-making interventions that address specific challenges encountered in these decision-making contexts. Furthermore, we restricted our research to a single decision-making scenario. Future studies should seek to validate our results in various decision-making areas to explore the broader applicability and specific contexts of the observed impacts.

Group decision-making requires further examinations. Studying the impact of group dynamics, we focused on small groups with members paired randomly. The complexity of group dynamics becomes more evident in larger settings, where diverse perspectives and potential power dynamics come into play. Additionally, it's important to note that the randomly formed groups may not fully mirror the dynamics found in real-world group settings, where members are often selected based on specific expertise, roles, or relationships. The small groups we considered in our study served as a starting point for examining group decision-making processes. However, future research should also investigate the implications and dynamics of decision-making in larger or more specifically defined groups. We also acknowledge that numerous factors can influence the effectiveness of decision-support tools in a group context, which were not examined in this study. These include leadership dynamics, communication patterns, varied hierarchical structures, individual biases, conflict resolution strategies, and decision-making norms. Future research should, therefore, explore and integrate these factors to develop an in-depth understanding of decision-making processes in group settings.

Sociotechnical factors need to be assessed collectively over time. While we aimed to create a naturalistic setting for decision-making, there are still limitations in terms of external validity and generalizability. Our studies may not have fully captured numerous factors that can influence decision-making tasks in the real world. We did not consider individual differences in decision-making styles, personality traits, or cognitive abilities, which could impact the strategies employed by decision-makers and ultimately determine their decision-making outcomes. In the pursuit of constructing sociotechnical systems tailored to various contexts, we should examine the task contexts, decision-makers, and AI systems in scenarios that emulate real-world conditions while taking into account a variety of individual, contextual, and social factors simultaneously. Moreover, influential factors can change over time, requiring continuous adjustments to decision-making situations. For instance, the approaches and methods for making decisions may evolve based on the knowledge gained by decision-makers during a specific session. The dynamic nature of decision-making scenarios necessitates empirical research to observe the processes and outcomes of decision-making over a prolonged period. Therefore, future research should focus on studying decision-making in real-world contexts, considering individual and contextual factors, and examining the long-term effects of AI adoption on decision outcomes.

Evaluation criteria need to be enhanced. Through our empirical studies, we measured different behavioral indicators and performance metrics to assess the success of human-AI teams in decision-making. We mainly focused on measuring accuracy, efficiency, trust, and reliance as crucial metrics and comparing them with the ideal criteria like appropriate reliance and complementary performance. However, these metrics alone may not capture the full complexity and nuances of decision-making outcomes. Depending on different contexts, the success of human-AI decision-making may also be determined by factors such as stakeholder satisfaction, ethical considerations, long-term consequences, and adaptability to changing conditions. For instance, the efficiency of the decision-making process may be prioritized in specific scenarios. In contrast, performance metrics or ethical considerations may take precedence in others, such as high-risk situations or morally sensitive decision-making contexts. Therefore, a range of metrics should be considered based on the specific objectives and needs of each decision-making scenario. Furthermore, we had access to ground truth data to evaluate the performance of human-AI teams. However, in real-world decision-making contexts, there may be limited or no availability of such data. Therefore, indicators for appropriate reliance on AI systems and the quality and consequences of decisions may not be immediately apparent or even become absent. This highlights the need for future studies to explore robust evaluation criteria that can accurately assess decision-making outcomes in dynamic environments with uncertain or unavailable ground truth data.

Additional methodologies for deeper insights. We examined the decision outcomes of individuals and groups and analyzed their behavioral and cognitive processes using established questionnaires, experimental tasks, and log data. However, this approach may have limitations in capturing all the steps involved in decision-making processes and understanding the rationale behind each action taken by the decision-makers. Therefore, future research should incorporate more detailed and context-specific methods such as thinking-aloud protocols and tailored interviews to gain deeper insights into decision-making processes. These insights may reveal the cognitive biases, heuristics, and reasoning patterns that impact decision-making outcomes. Additionally, they can aid in a better understanding of users' diverse needs and inform the customization of AI systems to address those specific requirements.

Expert decision-making requires a different approach. Within this thesis, we sought to target general decision-making contexts where no expertise or specific domain knowledge is required. This scoping was chosen to ensure that the findings and recommendations from this research can benefit a wide range of decision-makers in various contexts. Furthermore, experts often rely on their experience and knowledge rather than predominantly engaging in analytical decision-making processes or information-seeking. Therefore, the influence of AI systems on experts could be mediated by the level of expertise and domain knowledge, which should be further explored in future research.

Decision-making in the age of Generative AI. The emergence of large language models in recent years has significantly altered the adoption of AI technologies in decision-making settings, particularly impacting people's daily lives. Generative AI systems have brought many challenges and concerns regarding transparency, trust, reliance on these systems, and explainability. These challenges are magnified as AI systems become more complex and the broader range of user needs and preferences come into play. However, concerns about human-AI interaction have been long-standing in the HCI domain, which has emphasized the necessity of explainable, transparent, and reliable systems. For example, recommender systems have great potential to support decision-making processes. However, their lack of transparency and explainability has raised questions about their reliability. The impact of search engines and social media algorithms on information access and decision-making is also significant. The lack of transparency and potential bias in the ranking algorithms can greatly affect the quality and diversity of the information available to users, ultimately shaping individuals' behaviors and decisions. These issues highlight the importance of considering the behavior of the users and their understanding of the AI system's capabilities, limitations, and the consequences of relying on the system's decisions. Inspired by solutions to similar challenges in the field of HCI, future studies should continue evaluating user behaviors and decision-making outcomes with AI assistance in today's rapidly changing technological landscape. This can help identify specific challenges and opportunities that influence the design of AI systems while also increasing awareness about the implications of using AI in different decision-making scenarios.

8.4. Ethical Considerations and Challenges

As humans play a critical role in decision-making processes, we should consider the ethical implications and potential challenges associated with the use of AI systems. These may include privacy and confidentiality issues and biases that may arise in AI-driven decision-making systems.

We emphasized the importance of interventions to assist decision-makers in navigating complex and uncertain scenarios, aiming to mitigate over-reliance on AI systems and highlighting the continued value of human judgment. However, to ensure the effectiveness of these interventions, user behaviors and performance need to be carefully monitored and assessed, which may conflict with privacy concerns and GDPR. Consequently, striking a balance between harvesting the benefits of AI systems and safeguarding user privacy is essential for enhancing decision-making outcomes and maintaining ethical standards. In addition, potential biases that may arise in AI-driven decision-making systems should be thoroughly examined and addressed to ensure fair and equitable outcomes. These biases can stem from the data used to train AI algorithms, potentially leading to unfair outcomes or perpetuating existing societal inequalities. To address these challenges and ensure fairness and equity in AI-driven decision-making, transparency and interpretability should be incorporated into AI systems. This will enable users to comprehend the AI system's decision-making process and understand the rationale behind its actions, thus promoting accountability and allowing for the detection and mitigation of biases.

8.5. Concluding Remarks

This thesis adds to the current body of literature on human-AI decision-making by emphasizing the significance of contextual factors in influencing decision-making outcomes. It also offers methodological guidance for understanding the impact of AI systems on decision-making, thereby informing empirical research design in this field. By proposing a configurable framework, we encourage future research to systematically investigate more factors that contribute to the success of human-AI decision-making while ensuring the reproducibility and generalizability of findings. In several empirical studies, we have found that contextual factors such as task complexity, task uncertainty, and group dynamics can significantly influence decision-making outcomes. High levels of uncertainty and complexity in tasks often cause individuals to rely too heavily on AI systems, as they exceed their cognitive capacity. Similarly, groups tend to follow similar patterns, with collective decision-making becoming increasingly influenced by the AI system's recommendations in more complex tasks. Over-reliance on AI systems has been observed in such complex decision-making contexts, leading to sub-optimal outcomes and reduced critical thinking abilities. Our results also indicate that integrating AI systems is more beneficial for groups than individuals, as the collective intelligence and diverse perspectives within a group can foster critical thinking and enhance decision-making outcomes.

Future studies should aim to identify additional influential factors and develop interventions that help users make more informed decisions when using AI systems in different domains and contexts. Understanding the impact of AI assistance on individuals and groups will enable us to design AI support that improves decision-making outcomes and encourages critical thinking tailoring to the specific needs and characteristics of decision-making contexts. We hope this research contributes to these goals by shedding light on the complex relationship between human factors, AI systems, and contextual elements. These goals can be achieved through the combined efforts of researchers, practitioners, and designers who recognize the importance of understanding the potential benefits and limitations of AI systems in decision-making.

Bibliography

[1] Saranya A. and Subhashini R. “A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends”. In: *Decision Analytics Journal* 7 (2023), p. 100230. ISSN: 2772-6622. DOI: https://doi.org/10.1016/j.dajour.2023.100230. URL: https://www.sciencedirect.com/science/article/pii/S277266222300070X.

[2] Tahir Abbas et al. “Making Time Fly: Using Fillers to Improve Perceived Latency in Crowd-Powered Conversational Systems”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9.1 (Oct. 2021), pp. 2–14. DOI: 10.1609/hcomp.v9i1.18935. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/18935.

[3] Tahir Abbas et al. “Understanding User Perceptions of Response Delays in Crowd-Powered Conversational Systems”. In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2 (Nov. 2022). DOI: 10.1145/3555765. URL: https://doi.org/10.1145/3555765.

[4] Ashraf Abdul et al. “COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376615. URL: https://doi.org/10.1145/3313831.3376615.

[5] Abdullah Almaatouq et al. “Task complexity moderates group synergy”. In: *Proceedings of the National Academy of Sciences* 118.36 (2021), e2101062118.

[6] Ahmed Alqaraawi et al. “Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 275–285. ISBN: 9781450371186. DOI: 10.1145/3377325.3377519. URL: https://doi.org/10.1145/3377325.3377519.

[7] Irwin Altman and Dalmas Taylor. “The development of interpersonal relationships”. In: *New York, Rinehart & Winston* (1973).

[8] Saleema Amershi et al. “Guidelines for Human-AI Interaction”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300233. URL: https://doi.org/10.1145/3290605.3300233.

[9] Saleema Amershi et al. "Power to the People: The Role of Humans in Interactive Machine Learning". In: *AI Magazine* 35.4 (Dec. 2014), pp. 105–120. DOI: 10.1609/aimag.v35i4.2513. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2513.

[10] I. Androutsopoulos, G.D. Ritchie, and P. Thanisch. "Natural language interfaces to databases – an introduction". In: *Natural Language Engineering* 1.1 (1995), pp. 29–81. DOI: 10.1017/S135132490000005X.

[11] Alessa Angerschmid et al. "Fairness and Explanation in AI-Informed Decision Making". In: *Machine Learning and Knowledge Extraction* 4.2 (2022), pp. 556–579. ISSN: 2504-4990. DOI: 10.3390/make4020026. URL: https://www.mdpi.com/2504-4990/4/2/26.

[12] Ariful Islam Anik and Andrea Bunt. "Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445736. URL: https://doi.org/10.1145/3411764.3445736.

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

[14] J. Arguello and R. Capra. "The Effects of Aggregated Search Coherence on Search Behavior". In: *ACM Trans. Inf. Syst.* 35.1 (Sept. 2016). ISSN: 1046-8188.

[15] Dan Ariely and Simon Jones. *Predictably irrational*. HarperCollins New York, 2008.

[16] Arthur Aron et al. "The experimental generation of interpersonal closeness: A procedure and some preliminary findings". In: *Personality and social psychology bulletin* 23.4 (1997), pp. 363–377.

[17] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376638. URL: https://doi.org/10.1145/3313831.3376638.

[18] Zahra Ashktorab et al. "AI-Assisted Human Labeling: Batching for Efficiency without Overreliance". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449163. URL: https://doi.org/10.1145/3449163.

[19] Omid Askarisichani et al. "Expertise and confidence explain how social influence evolves along intellective tasks". In: *arXiv preprint arXiv:2011.07168* (2020).

[20] Omid Askarisichani et al. "Predictive models for human–AI nexus in group decision making". In: *Annals of the New York Academy of Sciences* 1514.1 (2022), pp. 70–81.

[21] Sinem Aslan et al. “Human Expert Labeling Process (HELP): Towards a Reliable Higher-order User State Labeling Process and Tool to Assess Student Engagement”. In: *Educational Technology archive* 57 (2017), pp. 53–59.

[22] Richard Badham, Chris Clegg, and Toby Wall. “Socio-technical theory”. In: *Handbook of ergonomics* (2000).

[23] Bahador Bahrami et al. “Optimally interacting minds”. In: *Science* 329.5995 (2010), pp. 1081–1085.

[24] Mihalj Bakator and Dragica Radosav. “Deep learning and medical diagnosis: A review of literature”. In: *Multimodal Technologies and Interaction* 2.3 (2018), p. 47.

[25] Nikola Banovic et al. “Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust”. In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW1 (Apr. 2023). DOI: 10.1145/3579460. URL: https://doi.org/10.1145/3579460.

[26] Gagan Bansal et al. “Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 2–11. DOI: 10.1609/hcomp.v7i1.5285. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5285.

[27] Gagan Bansal et al. “Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445717. URL: https://doi.org/10.1145/3411764.3445717.

[28] Gagan Bansal et al. “Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff”. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI’19/IAAI’19/EAAI’19. Honolulu, Hawaii, USA: AAAI Press, 2019. ISBN: 978-1-57735-809-1. DOI: 10.1609/aaai.v33i01.33012429. URL: https://doi.org/10.1609/aaai.v33i01.33012429.

[29] John A Bargh and Melissa J Ferguson. “Beyond behaviorism: on the automaticity of higher mental processes.” In: *Psychological bulletin* 126.6 (2000), p. 925.

[30] Daniel Barkoczi and Mirta Galesic. “Social learning strategies modify the effect of network structure on group performance”. In: *Nature communications* 7.1 (2016), p. 13109.

[31] Robert S Baron. “So right it’s wrong: Groupthink and the ubiquitous nature of polarized group decision making”. In: *Advances in experimental social psychology* 37.2 (2005), pp. 219–253.

[32] Joshua Becker, Devon Brackbill, and Damon Centola. “Network dynamics of social influence in the wisdom of crowds”. In: *Proceedings of the national academy of sciences* 114.26 (2017), E5070–E5076.

[33] Emma Beede et al. “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–12. ISBN: 9781450367080. DOI: 10.1145/3313831.3376718. URL: https://doi.org/10.1145/3313831.3376718.

[34] Andrew Bell et al. “It’s Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy”. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 248–266. ISBN: 9781450393522. DOI: 10.1145/3531146.3533090. URL: https://doi.org/10.1145/3531146.3533090.

[35] K. Bennani-Smires et al. “Simple Unsupervised Keyphrase Extraction using Sentence Embeddings”. In: *Proc. 22nd CoNLL*. Association for Computational Linguistics, Oct. 2018, pp. 221–229.

[36] Jonathan Berant et al. “Semantic Parsing on Freebase from Question-Answer Pairs”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1533–1544. URL: https://aclanthology.org/D13-1160.

[37] Niels van Berkel and Kasper Hornbæk. “Implications of Human-Computer Interaction Research”. In: *Interactions* 30.4 (June 2023), pp. 50–55. ISSN: 1072-5520. DOI: 10.1145/3600103. URL: https://doi.org/10.1145/3600103.

[38] Niels van Berkel et al. “Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445365. URL: https://doi.org/10.1145/3411764.3445365.

[39] Niels van Berkel et al. “Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445365. URL: https://doi.org/10.1145/3411764.3445365.

[40] Michael S. Bernstein et al. “Soylent: A Word Processor with a Crowd Inside”. In: *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*. UIST ’10. New York, New York, USA: Association for Computing Machinery, 2010, pp. 313–322. ISBN: 9781450302715. DOI: 10.1145/1866029.1866078. URL: https://doi.org/10.1145/1866029.1866078.

[41] Tilmann Betsch and Susanne Haberstroh. “The routines of decision making”. In: 2005. URL: https://api.semanticscholar.org/CorpusID:142741226.

[42] Sudeep Bhatia. “Associations and the accumulation of preference.” In: *Psychological review* 120.3 (2013), p. 522.

[43] Reuben Binns et al. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3173951. URL: https://doi.org/10.1145/3173574.3173951.

[44] Or Biran and Kathleen McKeown. "Human-Centric Justification of Machine Learning Predictions". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI'17. Melbourne, Australia: AAAI Press, 2017, pp. 1461–1467. ISBN: 9780999241103.

[45] Geoff Boeing. "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks". In: *Computers, Environment and Urban Systems* 65 (2017), pp. 126–139. ISSN: 0198-9715. DOI: https://doi.org/10.1016/j.compenvurbsys.2017.05.004. URL: https://www.sciencedirect.com/science/article/pii/S0198971516303970.

[46] Angie Boggust et al. "Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501965. URL: https://doi.org/10.1145/3491102.3501965.

[47] Ben Bogin, Matt Gardner, and Jonathan Berant. "Global Reasoning over Database Structures for Text-to-SQL Parsing". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3659–3664. DOI: 10.18653/v1/D19-1378. URL: https://aclanthology.org/D19-1378.

[48] Silvia Bonaccio and Reeshad S Dalal. "Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences". In: *Organizational behavior and human decision processes* 101.2 (2006), pp. 127–151.

[49] Nigel Bosch and Sidney K. D'Mello. "Can Computers Outperform Humans in Detecting User Zone-Outs? Implications for Intelligent Interfaces". In: *ACM Trans. Comput.-Hum. Interact.* 29.2 (Jan. 2022). ISSN: 1073-0516. DOI: 10.1145/3481889. URL: https://doi.org/10.1145/3481889.

[50] H. Bota, K. Zhou, and J.M. Jose. "Playing Your Cards Right: The Effect of Entity Cards on Search Behaviour and Workload". In: *Proc. 1st ACM CHIIR*. 2016, pp. 131–140.

[51] A. Bougouin, F. Boudin, and B. Daille. "TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction". In: *Proc. 6th AACL-IJCNLP*. 2013, pp. 543–551.

[52] Amy Bruckman. "Research ethics and HCI". In: *Ways of Knowing in HCI* (2014), pp. 449–468.

[53] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449287. URL: https://doi.org/10.1145/3449287.

[54] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449287. URL: https://doi.org/10.1145/3449287.

[55] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449287. URL: https://doi.org/10.1145/3449287.

[56] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449287. URL: https://doi.org/10.1145/3449287.

[57] Zana Buçinca et al. "Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI '20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 454–464. ISBN: 9781450371186. DOI: 10.1145/3377325.3377498. URL: https://doi.org/10.1145/3377325.3377498.

[58] Zana Buçinca et al. "Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI '20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 454–464. ISBN: 9781450371186. DOI: 10.1145/3377325.3377498. URL: https://doi.org/10.1145/3377325.3377498.

[59] David V Budescu and Eva Chen. "Identifying expertise to extract the wisdom of crowds". In: *Management science* 61.2 (2015), pp. 267–280.

[60] Tung X Bui and A Co-oP. "A group decision support system for cooperative multiple criteria group decision making". In: *Lecture Notes in Computer Science, Berlin Heidelberg, Germany, Springer-Verlag* (1987).

[61] Jason Burton, Mari-Klara Stein, and Tina Blegind Jensen. "A systematic review of algorithm aversion in augmented decision making". In: *Journal of Behavioral Decision Making* (2019). URL: https://api.semanticscholar.org/CorpusID:210439660.

[62] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. "A systematic review of algorithm aversion in augmented decision making". In: *Journal of behavioral decision making* 33.2 (2020), pp. 220–239.

[63] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. “The Effects of Example-Based Explanations in a Machine Learning Interface”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 258–262. ISBN: 9781450362726. DOI: 10.1145/3301275.3302289. URL: https://doi.org/10.1145/3301275.3302289.

[64] Carrie J. Cai et al. “"Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359206. URL: https://doi.org/10.1145/3359206.

[65] Carrie J. Cai et al. “Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–14. ISBN: 9781450359702. DOI: 10.1145/3290605.3300234. URL: https://doi.org/10.1145/3290605.3300234.

[66] Ruichu Cai et al. “SADGA: Structure-Aware Dual Graph Aggregation Network for Text-to-SQL”. In: *CoRR* abs/2111.00653 (2021). arXiv: 2111.00653. URL: https://arxiv.org/abs/2111.00653.

[67] Wanling Cai, Yucheng Jin, and Li Chen. “Impacts of Personal Characteristics on User Trust in Conversational Recommender Systems”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. <conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517471. URL: https://doi.org/10.1145/3491102.3517471.

[68] Yitao Cai and Xiaojun Wan. “IGSQL: Database Schema Interaction Graph Based Neural Model for Context-Dependent Text-to-SQL Generation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6903–6912. DOI: 10.18653/v1/2020.emnlp-main.560. URL: https://aclanthology.org/2020.emnlp-main.560.

[69] A. Câmara et al. “Searching to Learn with Instructional Scaffolding”. In: *Proc. 6th ACM CHIIR*. Canberra ACT, Australia, 2021, pp. 209–218.

[70] Giovanni Campagna et al. “Almond: The Architecture of an Open, Crowdsourced, Privacy-Preserving, Programmable Virtual Assistant”. In: *Proceedings of the 26th International Conference on World Wide Web*. WWW ’17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 341–350. ISBN: 9781450349130. DOI: 10.1145/3038912.3052562. URL: https://doi.org/10.1145/3038912.3052562.

[71] Donald J. Campbell. “Task Complexity: A Review and Analysis”. In: *The Academy of Management Review* 13.1 (1988), pp. 40–52. ISSN: 03637425. URL: http://www.jstor.org/stable/258353 (visited on 09/16/2022).

[72] R. Campos et al. “YAKE! Keyword extraction from single documents using multiple local features”. In: *Inf. Sci.* 509 (2020), pp. 257–289.

[73] Ruisheng Cao et al. “LGESQL: Line Graph Enhanced Text-to-SQL Model with Mixed Local and Non-Local Relations”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 2541–2555. DOI: 10.18653/v1/2021.acl-long.198. URL: https://aclanthology.org/2021.acl-long.198.

[74] Tommaso Caselli and Oana Inel. “Crowdsourcing StoryLines: Harnessing the Crowd for Causal Relation Annotation”. In: *Proceedings of the Workshop Events and Stories in the News 2018*. Santa Fe, New Mexico, U.S.A: Association for Computational Linguistics, Aug. 2018, pp. 44–54. URL: https://aclanthology.org/W18-4306.

[75] Siew H. Chan, Qian Song, and Lee J. Yao. “The moderating roles of subjective (perceived) and objective task complexity in system use and performance”. In: *Computers in Human Behavior* 51 (2015), pp. 393–402. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2015.04.059. URL: https://www.sciencedirect.com/science/article/pii/S0747563215003544.

[76] Zhi Chen et al. “ShadowGNN: Graph Projection Neural Network for Text-to-SQL Parser”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 5567–5577. DOI: 10.18653/v1/2021.naacl-main.441. URL: https://aclanthology.org/2021.naacl-main.441.

[77] Hao-Fei Cheng et al. “Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300789. URL: https://doi.org/10.1145/3290605.3300789.

[78] Chun-Wei Chiang and Ming Yin. “Exploring the Effects of Machine Learning Literacy Interventions on Laypeople’s Reliance on Machine Learning Models”. In: *27th International Conference on Intelligent User Interfaces*. IUI ’22. Helsinki, Finland: Association for Computing Machinery, 2022, pp. 148–161. ISBN: 9781450391443. DOI: 10.1145/3490099.3511121. URL: https://doi.org/10.1145/3490099.3511121.

[79] Chun-Wei Chiang and Ming Yin. “You’d Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift”. In: *13th ACM Web Science Conference 2021*. WebSci ’21. Virtual Event, United Kingdom: Association for Computing Machinery, 2021, pp. 120–129. ISBN: 9781450383301. DOI: 10.1145/3447535.3462487. URL: https://doi.org/10.1145/3447535.3462487.

[80] Chun-Wei Chiang et al. “Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581015. URL: https://doi.org/10.1145/3544548.3581015.

[81] Shih-Yi Chien et al. “The Effect of Culture on Trust in Automation: Reliability and Workload”. In: *ACM Trans. Interact. Intell. Syst.* 8.4 (Nov. 2018). ISSN: 2160-6455. DOI: 10.1145/3230736. URL: https://doi.org/10.1145/3230736.

[82] B. Choi et al. “OrgBox: A Knowledge Representation Tool to Support Complex Search Tasks”. In: *Proc. 6th ACM CHIIR*. 2021, pp. 219–228.

[83] DongHyun Choi et al. “RYANSQL: Recursively Applying Sketch-based Slot Fillings for Complex Text-to-SQL in Cross-Domain Databases”. In: *Computational Linguistics* 47.2 (June 2021), pp. 309–332. DOI: 10.1162/coli_a_00403. URL: https://aclanthology.org/2021.cl-2.12.

[84] Michael Chromik et al. “I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI”. In: *26th International Conference on Intelligent User Interfaces*. IUI ’21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 307–317. ISBN: 9781450380171. DOI: 10.1145/3397481.3450644. URL: https://doi.org/10.1145/3397481.3450644.

[85] Robert B Cialdini and Robert B Cialdini. *Influence: The psychology of persuasion*. Vol. 55. Collins New York, 2007.

[86] Edward T Cokely et al. “Decision making skill: From intelligence to numeracy and expertise”. In: *Cambridge handbook of expertise and expert performance* 2018 (2018), pp. 476–505.

[87] M.J. Cole et al. “Inferring user knowledge level from eye movement patterns”. In: *IP&M* 49.5 (2013), pp. 1075–1091.

[88] K. Collins-Thompson, P. Hansen, and C. Hauff. “Search as learning (dagstuhl seminar 17092)”. In: *Dagstuhl reports*. Vol. 7. 2. 2017.

[89] K. Collins-Thompson et al. “Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies”. In: *Proc. 1st ACM CHIIR*. 2016, pp. 163–172.

[90] H. Colt et al. “Measuring learning gain during a one-day introductory bronchoscopy course”. In: *Surgical endoscopy* 25 (Jan. 2011), pp. 207–16.

[91] Thomas D Cook, Donald Thomas Campbell, and William Shadish. *Experimental and quasi-experimental designs for generalized causal inference*. Vol. 1195. Houghton Mifflin Boston, MA, 2002.

[92] Sam Corbett-Davies et al. “Algorithmic Decision Making and the Cost of Fairness”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’17. Halifax, NS, Canada: Association for Computing Machinery, 2017, pp. 797–806. ISBN: 9781450348874. DOI: 10.1145/3097983.3098095. URL: https://doi.org/10.1145/3097983.3098095.

[93] Jonathan Corney et al. “Towards crowdsourcing translation tasks in library cataloguing, a pilot study”. In: *4th IEEE International Conference on Digital Ecosystems and Technologies*. 2010, pp. 572–577. DOI: 10.1109/DEST.2010.5610593.

[94] Victor Coscrato and Derek Bridge. “Estimating and Evaluating the Uncertainty of Rating Predictions and Top-n Recommendations in Recommender Systems”. In: *ACM Trans. Recomm. Syst.* 1.2 (Apr. 2023). DOI: 10.1145/3584021. URL: https://doi.org/10.1145/3584021.

[95] Anita Crescenzi et al. “Adaptation in Information Search and Decision-Making under Time Constraints”. In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. CHIIR ’21. Canberra ACT, Australia: Association for Computing Machinery, 2021, pp. 95–105. ISBN: 9781450380553. DOI: 10.1145/3406522.3446030. URL: https://doi.org/10.1145/3406522.3446030.

[96] Deborah A. Dahl et al. “Expanding the Scope of the ATIS Task: The ATIS-3 Corpus”. In: *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. 1994. URL: https://aclanthology.org/H94-1010.

[97] E. Dale. “Vocabulary Measurement: Techniques and Major Findings”. In: *Elementary English* 42.8 (1965), pp. 895–948.

[98] Florian Daniel et al. “Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions”. In: *ACM Comput. Surv.* 51.1 (Jan. 2018). ISSN: 0360-0300. URL: https://doi.org/10.1145/3148148.

[99] Devleena Das and Sonia Chernova. “Leveraging Rationales to Improve Human Task Performance”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 510–518. ISBN: 9781450371186. DOI: 10.1145/3377325.3377512. URL: https://doi.org/10.1145/3377325.3377512.

[100] Devleena Das and Sonia Chernova. “Leveraging Rationales to Improve Human Task Performance”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 510–518. ISBN: 9781450371186. DOI: 10.1145/3377325.3377512. URL: https://doi.org/10.1145/3377325.3377512.

[101] Jeffrey Dastin. “Amazon scraps secret AI recruiting tool that showed bias against women”. In: *Ethics of data and analytics*. Auerbach Publications, 2022, pp. 296–299.

[102] Clintin P Davis-Stober et al. “When is a crowd wise?” In: *Decision* 1.2 (2014), p. 79.

[103] Bart A De Jong, Kurt T Dirks, and Nicole Gillespie. “Trust and team performance: A meta-analysis of main effects, moderators, and covariates.” In: *Journal of applied psychology* 101.8 (2016), p. 1134.

[104] Xiang Deng et al. “Structure-Grounded Pretraining for Text-to-SQL”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 1337–1350. DOI: 10.18653/v1/2021.naacl-main.105. URL: https://aclanthology.org/2021.naacl-main.105.

[105] Michael Desmond et al. “Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface”. In: *26th International Conference on Intelligent User Interfaces*. IUI ’21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 392–401. ISBN: 9781450380171. DOI: 10.1145/3397481.3450698. URL: https://doi.org/10.1145/3397481.3450698.

[106] Dennis J Devine and Jennifer L Philips. “Do smarter teams do better: A meta-analysis of cognitive ability and team performance”. In: *Small group research* 32.5 (2001), pp. 507–532.

[107] Ap Dijksterhuis. “Think different: the merits of unconscious thought in preference development and decision making.” In: *Journal of personality and social psychology* 87.5 (2004), p. 586.

[108] Murat Dikmen and Catherine Burns. “The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending”. In: *International Journal of Human-Computer Studies* 162 (2022), p. 102792.

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

[110] Jonathan Dodge et al. “Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 275–285. ISBN: 9781450362726. DOI: 10.1145/3301275.3302310. URL: https://doi.org/10.1145/3301275.3302310.

[111] Li Dong and Mirella Lapata. “Coarse-to-Fine Decoding for Neural Semantic Parsing”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 731–742. DOI: 10.18653/v1/P18-1068. URL: https://aclanthology.org/P18-1068.

[112] Li Dong and Mirella Lapata. “Language to Logical Form with Neural Attention”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 33–43. DOI: 10.18653/v1/P16-1004. URL: https://aclanthology.org/P16-1004.

[113] Li Dong, Chris Quirk, and Mirella Lapata. “Confidence Modeling for Neural Semantic Parsing”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 743–753. DOI: 10.18653/v1/P18-1069. URL: https://aclanthology.org/P18-1069.

[114] Zhen Dong et al. “Data-Anonymous Encoding for Text-to-SQL Generation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5405–5414. DOI: 10.18653/v1/D19-1543. URL: https://aclanthology.org/D19-1543.

[115] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[116] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[117] Tim Draws et al. “A Checklist to Combat Cognitive Biases in Crowdsourcing”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9.1 (Oct. 2021), pp. 48–59. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/18939.

[118] Julia Dressel and Hany Farid. “The accuracy, fairness, and limits of predicting recidivism”. In: *Science Advances* 4.1 (2018), eaao5580. DOI: 10.1126/sciadv.aao5580. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aao5580. URL: https://www.science.org/doi/abs/10.1126/sciadv.aao5580.

[119] Jaimie Drozdal et al. “Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 297–307. ISBN: 9781450371186. DOI: 10.1145/3377325.3377501. URL: https://doi.org/10.1145/3377325.3377501.

[120] Susan Dumais et al. “Understanding user behavior through log data and analysis”. In: *Ways of Knowing in HCI* (2014), pp. 349–372.

[121] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. “AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517443. URL: https://doi.org/10.1145/3491102.3517443.

[122] Serge Egelman, Ed H Chi, and Steven Dow. “Crowdsourcing in HCI research”. In: *Ways of Knowing in HCI*. Springer, 2014, pp. 267–289.

[123] K. Eichler and G. Neumann. “DFKI KeyWE: Ranking Keyphrases Extracted from Scientific Articles”. In: *Proc. 5th SemEval*. July 2010, pp. 150–153.

[124] C. Eickhoff, S. Dungs, and V. Tran. “An eye-tracking study of query reformulation”. In: *Proc. 38th ACM SIGIR*. 2015, pp. 13–22.

[125] C. Eickhoff et al. “Lessons from the Journey: A Query Log Analysis of within-Session Learning”. In: WSDM ’14. 2014, pp. 223–232.

[126] G. Ercan and I. Cicekli. “Using lexical chains for keyword extraction”. In: *Inf. Process. Manag.* 43 (2007), pp. 1705–1714.

[127] Alexander Erlei, Abhinav Sharma, and Ujwal Gadiraju. “Understanding Choice Independence and Error Types in Human-AI Collaboration”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. New York, NY, USA: Association for Computing Machinery, 2024.

[128] Alexander Erlei et al. “For What It’s Worth: Humans Overwrite Their Economic Self-Interest to Avoid Bargaining With AI Systems”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517734. URL: https://doi.org/10.1145/3491102.3517734.

[129] Alexander Erlei et al. “For what it’s worth: Humans overwrite their economic self-interest to avoid bargaining with AI systems”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–18.

[130] Alexander Erlei et al. “Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8.1 (Oct. 2020), pp. 43–52. DOI: 10.1609/hcomp.v8i1.7462. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7462.

[131] Enrique Estellés-Arolas and Fernando González-Ladrón-De-Guevara. “Towards an Integrated Crowdsourcing Definition”. In: *J. Inf. Sci.* 38.2 (Apr. 2012), pp. 189–200. ISSN: 0165-5515. DOI: 10.1177/0165551512437638. URL: https://doi.org/10.1177/0165551512437638.

[132] Angela Fagerlin et al. “Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale”. In: *Medical Decision Making* 27.5 (Sept. 2007), pp. 672–680. DOI: 10.1177/0272989X07304449. URL: https://ideas.repec.org/a/sae/medema/v27y2007i5p672-680.html.

[133] Shaoyang Fan et al. “CrowdCO-OP: Sharing Risks and Rewards in Crowdsourcing”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–24.

[134] C Melissa Fender and Lisa T Stickney. “When two heads aren’t better than one: conformity in a group activity”. In: *Management Teaching Review* 2.1 (2017), pp. 35–46.

[135] Catherine Finegan-Dollak et al. “Improving Text-to-SQL Evaluation Methodology”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 351–360. DOI: 10.18653/v1/P18-1033. URL: https://aclanthology.org/P18-1033.

[136] Martin Fishbein and Icek Ajzen. *Predicting and changing behavior: The reasoned action approach*. Psychology press, 2011.

[137] C. Florescu and C. Caragea. “A Position-Biased PageRank Algorithm for Keyphrase Extraction”. In: *AAAI*. 2017.

[138] Brian J Fogg. “Persuasive technology: using computers to change what we think and do”. In: *Ubiquity* 2002.December (2002), p. 2.

[139] Riccardo Fogliato et al. “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. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1362–1374. ISBN: 9781450393522. DOI: 10.1145/3531146.3533193. URL: https://doi.org/10.1145/3531146.3533193.

[140] Raymond Fok and Daniel S Weld. “In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making”. In: *arXiv preprint arXiv:2305.07722* (2023).

[141] Donelson R Forsyth. “Group dynamics”. In: (2011).

[142] Karën Fort, Gilles Adda, and K. Bretonnel Cohen. “Last Words: Amazon Mechanical Turk: Gold Mine or Coal Mine?” In: *Computational Linguistics* 37.2 (June 2011), pp. 413–420. DOI: 10.1162/COLI_a_00057. URL: https://aclanthology.org/J11-2010.

[143] Thomas Franke, Christiane Attig, and Daniel Wessel. “A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale”. In: *International Journal of Human–Computer Interaction* 35.6 (2019), pp. 456–467. DOI: 10.1080/10447318.2018.1456150. eprint: https://doi.org/10.1080/10447318.2018.1456150. URL: https://doi.org/10.1080/10447318.2018.1456150.

[144] Simon French, John Maule, and Nadia Papamichail. *Decision Behaviour, Analysis and Support*. Cambridge University Press, 2009.

[145] U. Gadiraju et al. “Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web”. In: Proc. 3rd ACM CHIIR. 2018, pp. 2–11.

[146] Ujwal Gadiraju et al. “Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd”. In: *Evaluation in the crowd. Crowdsourcing and human-centered experiments*. Springer, 2017, pp. 6–26.

[147] Krzysztof Z. Gajos and Lena Mamykina. "Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning". In: *27th International Conference on Intelligent User Interfaces*. IUI '22. Helsinki, Finland: Association for Computing Machinery, 2022, pp. 794–806. ISBN: 9781450391443. DOI: 10.1145/3490099.3511138. URL: https://doi.org/10.1145/3490099.3511138.

[148] Edwin Gamboa et al. "The Crowd Thinks Aloud: Crowdsourcing Usability Testing with the Thinking Aloud Method". In: *HCI International 2021 - Late Breaking Papers: Design and User Experience*. Ed. by Constantine Stephanidis et al. Cham: Springer International Publishing, 2021, pp. 24–39. ISBN: 978-3-030-90238-4.

[149] Yujian Gan et al. "Natural SQL: Making SQL Easier to Infer from Natural Language Specifications". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2030–2042. DOI: 10.18653/v1/2021.findings-emnlp.174. URL: https://aclanthology.org/2021.findings-emnlp.174.

[150] Darren Gergle and Desney S Tan. "Experimental research in HCI". In: *Ways of Knowing in HCI*. Springer, 2014, pp. 191–227.

[151] S. Ghosh, M. Rath, and C. Shah. "Searching as Learning: Exploring Search Behavior and Learning Outcomes in Learning-related Tasks". In: *Proc. 3rd ACM CHIIR* (2018), pp. 22–31.

[152] Gerd Gigerenzer. "Fast and frugal heuristics: The tools of bounded rationality". In: *Blackwell handbook of judgment and decision making* 62 (2004), p. 88.

[153] Gerd Gigerenzer. "Gut feelings: The intelligence of the unconscious". In: *Proceedings of the European Cognitive Science Conference 2007*. Taylor & Francis. 2017, p. 3.

[154] Gerd Gigerenzer and Peter M Todd. *Simple heuristics that make us smart*. Oxford University Press, USA, 1999.

[155] Ella Glikson and Anita Woolley. "Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals (in press)". In: *The Academy of Management Annals* (Apr. 2020).

[156] Oscar Gomez et al. "ViCE: Visual Counterfactual Explanations for Machine Learning Models". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI '20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 531–535. ISBN: 9781450371186. DOI: 10.1145/3377325.3377536. URL: https://doi.org/10.1145/3377325.3377536.

[157] Mitchell L. Gordon et al. "Jury Learning: Integrating Dissenting Voices into Machine Learning Models". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3502004. URL: https://doi.org/10.1145/3491102.3502004.

[158] Kazjon Grace et al. “Q-Chef: The Impact of Surprise-Eliciting Systems on Food-Related Decision-Making”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501862. URL: https://doi.org/10.1145/3491102.3501862.

[159] Catherine Grady and Matthew Lease. “Crowdsourcing Document Relevance Assessment with Mechanical Turk”. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. Los Angeles: Association for Computational Linguistics, June 2010, pp. 172–179. URL: https://aclanthology.org/W10-0727.

[160] Ben Green and Yiling Chen. “Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3479562. URL: https://doi.org/10.1145/3479562.

[161] Ben Green and Yiling Chen. “Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts”. In: *arXiv preprint arXiv:2012.05370* (2020).

[162] Ben Green and Yiling Chen. “The Principles and Limits of Algorithm-in-the-Loop Decision Making”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359152. URL: https://doi.org/10.1145/3359152.

[163] Ben Green and Yiling Chen. “The Principles and Limits of Algorithm-in-the-Loop Decision Making”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359152. URL: https://doi.org/10.1145/3359152.

[164] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. “Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359280. URL: https://doi.org/10.1145/3359280.

8

[165] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. “Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359280. URL: https://doi.org/10.1145/3359280.

[166] Jiaqi Guo et al. “Towards Complex Text-to-SQL in Cross-Domain Database with Intermediate Representation”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4524–4535. DOI: 10.18653/v1/P19-1444. URL: https://aclanthology.org/P19-1444.

[167] Lijie Guo et al. “Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules”. In: *27th International Conference on Intelligent User Interfaces*. IUI ’22. Helsinki, Finland: Association for Computing Machinery, 2022, pp. 537–548. ISBN: 9781450391443. DOI: 10.1145/3490099.3511111. URL: https://doi.org/10.1145/3490099.3511111.

[168] Tong Guo and Huilin Gao. *Content Enhanced BERT-based Text-to-SQL Generation*. 2020. arXiv: 1910.07179 [cs.CL].

[169] Rotem D. Guttman et al. "Play for Real(Ism) - Using Games to Predict Human-AI Interactions in the Real World". In: *Proc. ACM Hum.-Comput. Interact.* 5.CHI PLAY (Oct. 2021). DOI: 10.1145/3474655. URL: https://doi.org/10.1145/3474655.

[170] Didier Guzzoni et al. "Active, A Tool for Building Intelligent User Interfaces". In: ASC. 2007.

[171] J. Richard Hackman. "Toward understanding the role of tasks in behavioral research". In: *Acta Psychologica* 31 (1969), pp. 97–128. ISSN: 0001-6918. DOI: https://doi.org/10.1016/0001-6918(69)90073-0. URL: https://www.sciencedirect.com/science/article/pii/0001691869900730.

[172] Sophia Hadash et al. "Improving Understandability of Feature Contributions in Model-Agnostic Explainable AI Tools". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517650. URL: https://doi.org/10.1145/3491102.3517650.

[173] THORVALD HÆREM, BRIAN T. PENTLAND, and KENT D. MILLER. "TASK COMPLEXITY: EXTENDING A CORE CONCEPT". In: *The Academy of Management Review* 40.3 (2015), pp. 446–460. ISSN: 03637425. URL: http://www.jstor.org/stable/43700530 (visited on 09/16/2022).

[174] Lei Han et al. "The impact of task abandonment in crowdsourcing". In: *IEEE Transactions on Knowledge and Data Engineering* 33.5 (2019), pp. 2266–2279.

[175] Christopher G. Harris. "You're hired! an examination of crowdsourcing incentive models in human resource tasks". In: *in WSDM Workshop on Crowdsourcing for Search and Data Mining (CSDM*, pp. 15–18.

[176] Galen Harrison et al. "An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 392–402. ISBN: 9781450369367. DOI: 10.1145/3351095.3372831. URL: https://doi.org/10.1145/3351095.3372831.

[177] Nichole Harvey and Colin A Holmes. "Nominal group technique: an effective method for obtaining group consensus". In: *International journal of nursing practice* 18.2 (2012), pp. 188–194.

[178] F. Hasibi, K. Balog, and S.E. Bratsberg. "Dynamic Factual Summaries for Entity Cards". In: *Proc. 40th ACM SIGIR*. 2017, pp. 773–782.

[179] Gaole He and Ujwal Gadiraju. "Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making". In: *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22)*. 2022.

[180] Gaole He and Ujwal Gadiraju. "Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making". In: *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22)*. 2022.

[181] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. "Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581025. URL: https://doi.org/10.1145/3544548.3581025.

[182] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. "Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. New York, NY, USA: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581025. URL: https://doi.org/10.1145/3544548.3581025.

[183] Pengcheng He et al. *X-SQL: reinforce schema representation with context*. 2019. arXiv: 1908.08113 [cs.CL].

[184] Gary G. Hendrix et al. "Developing a Natural Language Interface to Complex Data". In: *ACM Trans. Database Syst.* 3.2 (June 1978), pp. 105–147. ISSN: 0362-5915. DOI: 10.1145/320251.320253. URL: https://doi.org/10.1145/320251.320253.

[185] David Dryden Henningsen, Michael G Cruz, and Mary Lynn Miller. "Role of social loafing in predeliberation decision making." In: *Group Dynamics: Theory, Research, and Practice* 4.2 (2000), p. 168.

[186] Jonathan Herzig et al. "Unlocking Compositional Generalization in Pre-trained Models Using Intermediate Representations". In: *ArXiv* abs/2104.07478 (2021).

[187] Daniel Herzog and Wolfgang Wörndl. "A User Study on Groups Interacting with Tourist Trip Recommender Systems in Public Spaces". In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP '19. Larnaca, Cyprus: Association for Computing Machinery, 2019, pp. 130–138. ISBN: 9781450360210. DOI: 10.1145/3320435.3320449. URL: https://doi.org/10.1145/3320435.3320449.

[188] Michael Hilb. "Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance". In: *Journal of Management and Governance* 24 (2020), pp. 851–870.

[189] Randy Y Hirokawa and Marshall Scott Poole. *Communication and group decision making*. Vol. 77. Sage, 1996.

[190] Kevin Anthony Hoff and Masooda Bashir. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust". In: *Human Factors* 57.3 (2015). PMID: 25875432, pp. 407–434. DOI: 10.1177/0018720814547570. eprint: https://doi.org/10.1177/0018720814547570. URL: https://doi.org/10.1177/0018720814547570.

[191] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. “Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 8. 2020, pp. 63–72.

[192] Lu Hong and Scott E Page. “Groups of diverse problem solvers can outperform groups of high-ability problem solvers”. In: *Proceedings of the National Academy of Sciences* 101.46 (2004), pp. 16385–16389.

[193] Sujin K Horwitz and Irwin B Horwitz. “The effects of team diversity on team outcomes: A meta-analytic review of team demography”. In: *Journal of management* 33.6 (2007), pp. 987–1015.

[194] Tobias Hossfeld et al. “Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing”. In: *IEEE Transactions on Multimedia* 16.2 (2014), pp. 541–558. DOI: 10.1109/TMM.2013.2291663.

[195] Eduard Hovy et al. “OntoNotes: The 90% Solution”. In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York City, USA: Association for Computational Linguistics, June 2006, pp. 57–60. URL: https://aclanthology.org/N06-2015.

[196] Christopher K Hsee. “Attribute evaluability and its implications for joint-separate evaluation reversals and beyond”. In: (2000).

[197] George P Huber and Kyle Lewis. “Cross-understanding: Implications for group cognition and performance”. In: *Academy of Management review* 35.1 (2010), pp. 6–26.

[198] Scott E Hudson and Jennifer Mankoff. “Concepts, values, and methods for technical human–computer interaction research”. In: *Ways of Knowing in HCI*. Springer, 2014, pp. 69–93.

[199] Binyuan Hui et al. “Dynamic Hybrid Relation Exploration Network for Cross-Domain Context-Dependent Semantic Parsing”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.14 (May 2021), pp. 13116–13124. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17550.

[200] Aya Hussein, Sondoss Elsawah, and Hussein A. Abbass. “Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions”. In: *Human Factors* 62.8 (2020). PMID: 31590574, pp. 1237–1248. DOI: 10.1177/0018720819879273. eprint: https://doi.org/10.1177/0018720819879273. URL: https://doi.org/10.1177/0018720819879273.

[201] Wonseok Hwang et al. *A Comprehensive Exploration on WikiSQL with Table-Aware Word Contextualization*. 2019. arXiv: 1902.01069 [cs.CL].

[202] Mir Riyanul Islam et al. “A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks”. In: *Applied Sciences* 12.3 (2022). ISSN: 2076-3417. DOI: 10.3390/app12031353. URL: https://www.mdpi.com/2076-3417/12/3/1353.

[203] John M Ivancevich, Michael T Matteson, and Robert Konopaske. “Organizational behavior and management”. In: (1990).

[204] Srinivasan Iyer et al. “Learning a Neural Semantic Parser from User Feedback”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 963–973. DOI: 10.18653/v1/P17-1089. URL: https://aclanthology.org/P17-1089.

[205] Maia Jacobs et al. “How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection”. In: *Translational psychiatry* 11.1 (2021), p. 108.

[206] Alon Jacovi and Yoav Goldberg. “Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://aclanthology.org/2020.acl-main.386.

[207] Alon Jacovi et al. “Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 624–635. ISBN: 9781450383097. DOI: 10.1145/3442188.3445923. URL: https://doi.org/10.1145/3442188.3445923.

[208] Anthony Jameson, Martijn C. Willemsen, and Alexander Felfernig. “Individual and Group Decision Making and Recommender Systems”. In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, and Bracha Shapira. New York, NY: Springer US, 2022, pp. 789–832. ISBN: 978-1-0716-2197-4. DOI: 10.1007/978-1-0716-2197-4_21. URL: https://doi.org/10.1007/978-1-0716-2197-4_21.

[209] Anthony Jameson et al. “Choice architecture for human-computer interaction”. In: *Foundations and Trends® in Human–Computer Interaction* 7.1–2 (2014), pp. 1–235.

[210] Irving L Janis. “Victims of Groupthink: A psychological study of foreign-policy decisions and fiascoes.” In: (1972).

[211] Mohammad Hossein Jarrahi. “Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making”. In: *Business Horizons* 61.4 (2018), pp. 577–586. ISSN: 0007-6813. DOI: https://doi.org/10.1016/j.bushor.2018.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0007681318300387.

[212] Pradthana Jarusriboonchai, Aris Malapaschas, and Thomas Olsson. “Design and Evaluation of a Multi-Player Mobile Game for Icebreaking Activity”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. San Jose, California, USA: Association for Computing Machinery, 2016, pp. 4366–4377. ISBN: 9781450333627. DOI: 10.1145/2858036.2858298. URL: https://doi.org/10.1145/2858036.2858298.

[213] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. “Foundations for an Empirically Determined Scale of Trust in Automated Systems”. In: *International Journal of Cognitive Ergonomics* 4.1 (2000), pp. 53–71. DOI: 10.1207/S15327566IJCE0401_04. eprint: https://doi.org/10.1207/S15327566IJCE0401_04. URL: https://doi.org/10.1207/S15327566IJCE0401%5C_04.

[214] G. Jimmy Zuccon, G. Demartini, and B. Koopman. “Health Cards to Assist Decision Making in Consumer Health Search”. In: *AMIA* 2019 (Mar. 2020), pp. 1091–1100.

[215] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. “Rethinking AI Explainability and Plausibility”. In: *arXiv preprint arXiv:2303.17707* (2023).

[216] raymond Soames Job and James Dalziel. “Defining Fatigue as a Condition of the Organism and Distinguishing It From Habituation, Adaptation, and Boredom”. In: Nov. 2000, pp. 466–476. ISBN: 9781410600448. DOI: 10.1201/b12791-3.2.

[217] Joseph G Johnson and Jerome R Busemeyer. “Rule-based decision field theory: A dynamic computational model of transitions among decision-making strategies”. In: *The routines of decision making*. Psychology Press, 2014, pp. 3–20.

[218] Matthew K. Miller, Martin Johannes Dechant, and Regan L. Mandryk. “Meeting You, Seeing Me: The Role of Social Anxiety, Visual Feedback, and Interface Layout in a Get-to-Know-You Task via Video Chat.” In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. <conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445664. URL: https://doi.org/10.1145/3411764.3445664.

[219] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

[220] Daniel Kahneman and Amos Tversky. “Prospect theory: An analysis of decision under risk”. In: *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013, pp. 99–127.

[221] Patricia K. Kahr et al. “The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice.” In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. IUI ’24. <conf-loc>, <city>Greenville</city>, <state>SC</state>, <country>USA</country>, </conf-loc>: Association for Computing Machinery, 2024, pp. 609–622. ISBN: 9798400705083. DOI: 10.1145/3640543.3645167. URL: https://doi.org/10.1145/3640543.3645167.

[222] Hiroshi Kajino et al. “Learning from Crowds and Experts”. In: *HCOMP@AAAI*. 2012.

[223] R. Kalyani and U. Gadiraju. “Understanding User Search Behavior Across Varying Cognitive Levels”. In: *Proc. 30th ACM HT*. 2019, pp. 123–132.

[224] Ece Kamar. “Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.” In: *IJCAI*. 2016, pp. 4070–4073.

[225] Ece Kamar, Severin Hacker, and Eric Horvitz. "Combining human and machine intelligence in large-scale crowdsourcing". In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*. AAMAS '12. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 467–474. ISBN: 0981738117.

[226] Steven J Karau and Kipling D Williams. "Social loafing: A meta-analytic review and theoretical integration." In: *Journal of personality and social psychology* 65.4 (1993), p. 681.

[227] Anna Kawakami et al. "Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517439. URL: https://doi.org/10.1145/3491102.3517439.

[228] Young Ji Kim et al. "What makes a strong team? Using collective intelligence to predict team performance in League of Legends". In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2017, pp. 2316–2329.

[229] Alex Kirlik. "Modeling Strategic Behavior in Human-Automation Interaction: Why an "Aid" Can (and Should) Go Unused". In: *Human Factors* 35.2 (1993). PMID: 8349287, pp. 221–242. DOI: 10.1177/001872089303500203. eprint: https://doi.org/10.1177/001872089303500203. URL: https://doi.org/10.1177/001872089303500203.

[230] R. Kiros et al. *Skip-Thought Vectors*. 2015.

[231] Gary A Klein. *Sources of power: How people make decisions*. MIT press, 2017.

[232] Moritz Körber. "Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation". In: *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*. Ed. by Sebastiano Bagnara et al. Cham: Springer International Publishing, 2019, pp. 13–30. ISBN: 978-3-319-96074-6.

[233] Asher Koriat. "When Are Two Heads Better than One and Why?" In: *Science* 336.6079 (2012), pp. 360–362. DOI: 10.1126/science.1216549. eprint: https://www.science.org/doi/pdf/10.1126/science.1216549. URL: https://www.science.org/doi/abs/10.1126/science.1216549.

[234] Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. "Neural Semantic Parsing with Type Constraints for Semi-Structured Tables". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 1516–1526. DOI: 10.18653/v1/D17-1160. URL: https://aclanthology.org/D17-1160.

[235] Justin Kruger and David Dunning. "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." In: *Journal of personality and social psychology* 77.6 (1999), p. 1121.

[236] Carol C Kuhlthau. “A principle of uncertainty for information seeking”. In: *Journal of documentation* 49.4 (1993), pp. 339–355.

[237] Carol C Kuhlthau. “The influence of uncertainty on the information seeking behavior of a securities analyst”. In: *Proceedings of an international conference on Information seeking in context*. 1997, pp. 268–274.

[238] Todd Kulesza et al. “Tell Me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. Austin, Texas, USA: Association for Computing Machinery, 2012, pp. 1–10. ISBN: 9781450310154. DOI: 10.1145/2207676.2207678. URL: https://doi.org/10.1145/2207676.2207678.

[239] Johannes Kunkel et al. “Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300717. URL: https://doi.org/10.1145/3290605.3300717.

[240] Tom Kwiatkowksi et al. “Inducing Probabilistic CCG Grammars from Logical Form with Higher-Order Unification”. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, Oct. 2010, pp. 1223–1233. URL: https://aclanthology.org/D10-1119.

[241] Isaac Lage et al. “Human Evaluation of Models Built for Interpretability”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 59–67. DOI: 10.1609/hcomp.v7i1.5280. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5280.

[242] D. Lagun et al. “Towards Better Measurement of Attention and Satisfaction in Mobile Search”. In: *Proc 37th ACM SIGIR*. 2014, pp. 113–122.

[243] Vivian Lai, Han Liu, and Chenhao Tan. “"Why is ’Chicago’ Deceptive?" Towards Building Model-Driven Tutorials for Humans”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376873. URL: https://doi.org/10.1145/3313831.3376873.

[244] Vivian Lai and Chenhao Tan. “On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 29–38. ISBN: 9781450361255. DOI: 10.1145/3287560.3287590. URL: https://doi.org/10.1145/3287560.3287590.

[245] Vivian Lai et al. “Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies”. In: *ArXiv* abs/2112.11471 (2021).

[246] Vivian Lai et al. “Towards a science of human-ai decision making: a survey of empirical studies”. In: *arXiv preprint arXiv:2112.11471* (2021).

[247] Vivian Lai et al. “Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies”. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’23. Chicago, IL, USA: Association for Computing Machinery, 2023, pp. 1369–1385. ISBN: 9798400701924. DOI: 10.1145/3593013.3594087. URL: https://doi.org/10.1145/3593013.3594087.

[248] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. “Interpretable Decision Sets: A Joint Framework for Description and Prediction”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1675–1684. ISBN: 9781450342322. DOI: 10.1145/2939672.2939874. URL: https://doi.org/10.1145/2939672.2939874.

[249] M. Lalmas and L. Hong. “Tutorial on Metrics of User Engagement: Applications to News, Search and E-Commerce”. In: *Proc. 11th ACM WSDM*. 2018, pp. 781–782.

[250] J.H. Lau and T. Baldwin. “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation”. In: *Proc. 1st RepL4NLP*. 2016, pp. 78–86.

[251] Quoc V. Le and Tomas Mikolov. *Distributed Representations of Sentences and Documents*. 2014. arXiv: 1405.4053 [cs.CL].

[252] Dongjun Lee. “Clause-Wise and Recursive Decoding for Complex and Cross-Domain Text-to-SQL Generation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6045–6051. DOI: 10.18653/v1/D19-1624. URL: https://aclanthology.org/D19-1624.

[253] John D Lee and Katrina A See. “Trust in automation: Designing for appropriate reliance”. In: *Human factors* 46.1 (2004), pp. 50–80.

[254] Kenton Lee et al. “Event Detection and Factuality Assessment with Non-Expert Supervision”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 1643–1648. DOI: 10.18653/v1/D15-1189. URL: https://aclanthology.org/D15-1189.

[255] Min Hun Lee et al. “A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445472. URL: https://doi.org/10.1145/3411764.3445472.

[256] Min Hun Lee et al. “Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment”. In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW2 (Oct. 2020). DOI: 10.1145/3415227. URL: https://doi.org/10.1145/3415227.

[257] Min Hun Lee et al. “Interactive Hybrid Approach to Combine Machine and Human Intelligence for Personalized Rehabilitation Assessment”. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. CHIL ’20. Toronto, Ontario, Canada: Association for Computing Machinery, 2020, pp. 160–169. ISBN: 9781450370462. DOI: 10.1145/3368555.3384452. URL: https://doi.org/10.1145/3368555.3384452.

[258] Min Kyung Lee. “Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management”. In: *Big Data & Society* 5.1 (2018), p. 2053951718756684. DOI: 10.1177/2053951718756684. eprint: https://doi.org/10.1177/2053951718756684. URL: https://doi.org/10.1177/2053951718756684.

[259] Wenqiang Lei et al. “Re-examining the Role of Schema Linking in Text-to-SQL”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6943–6954. DOI: 10.18653/v1/2020.emnlp-main.564. URL: https://aclanthology.org/2020.emnlp-main.564.

[260] Ariel Levy et al. “Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445522. URL: https://doi.org/10.1145/3411764.3445522.

[261] Fei Li and H. V. Jagadish. “Constructing an Interactive Natural Language Interface for Relational Databases”. In: *Proc. VLDB Endow.* 8.1 (Sept. 2014), pp. 73–84. ISSN: 2150-8097. DOI: 10.14778/2735461.2735468. URL: https://doi.org/10.14778/2735461.2735468.

[262] Mengyao Li et al. “No Risk No Trust: Investigating Perceived Risk in Highly Automated Driving”. In: *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. AutomotiveUI ’19. Utrecht, Netherlands: Association for Computing Machinery, 2019, pp. 177–185. ISBN: 9781450368841. DOI: 10.1145/3342197.3344525. URL: https://doi.org/10.1145/3342197.3344525.

[263] Qing Li et al. “Understanding the Effects of Explanation Types and User Motivations on Recommender System Use”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8.1 (Oct. 2020), pp. 83–91. DOI: 10.1609/hcomp.v8i1.7466. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7466.

[264] Yuntao Li et al. “Pay More Attention to History: A Context Modeling Strategy for Conversational Text-to-SQL”. In: *CoRR* abs/2112.08735 (2021). arXiv: 2112.08735. URL: https://arxiv.org/abs/2112.08735.

8

[265] Percy Liang, Michael I. Jordan, and Dan Klein. “Learning Dependency-Based Compositional Semantics”. In: *Computational Linguistics* 39.2 (June 2013), pp. 389–446. ISSN: 0891-2017. DOI: 10.1162/COLI_a_00127. eprint: https://direct.mit.edu/coli/article-pdf/39/2/389/1812365/coli_a_00127.pdf. URL: https://doi.org/10.1162/COLI%5C_a%5C_00127.

[266] Mengqi Liao, S. Shyam Sundar, and Joseph B. Walther. “User Trust in Recommendation Systems: A Comparison of Content-Based, Collaborative and Demographic Filtering”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501936. URL: https://doi.org/10.1145/3491102.3501936.

[267] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. “Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445260. URL: https://doi.org/10.1145/3411764.3445260.

[268] Zhiyuan “Jerry” Lin et al. “The limits of human predictions of recidivism”. In: *Science Advances* 6.7 (2020), eaaz0652. DOI: 10.1126/sciadv.aaz0652. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aaz0652. URL: https://www.science.org/doi/abs/10.1126/sciadv.aaz0652.

[269] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. “Explainable ai: A review of machine learning interpretability methods”. In: *Entropy* 23.1 (2020), p. 18.

[270] C. Liu and X. Song. “How Do Information Source Selection Strategies Influence Users’ Learning Outcomes’”. In: *Proc. 3rd ACM CHIIR*. 2018, pp. 257–260.

[271] H. Liu, C. Liu, and N.J. Belkin. “Investigation of users’ knowledge change process in learning-related search tasks”. In: *Proc. ASIS&T* 56.1 (2019), pp. 166–175.

[272] Haoyan Liu et al. “Leveraging Adjective-Noun Phrasing Knowledge for Comparison Relation Prediction in Text-to-SQL”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3515–3520. DOI: 10.18653/v1/D19-1356. URL: https://aclanthology.org/D19-1356.

[273] Peng Liu and Zhizhong Li. “Task complexity: A review and conceptualization framework”. In: *International Journal of Industrial Ergonomics* 42.6 (2012), pp. 553–568. ISSN: 0169-8141. DOI: https://doi.org/10.1016/j.ergon.2012.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0169814112000868.

[274] Qian Liu et al. “Awakening Latent Grounding from Pretrained Language Models for Semantic Parsing”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1174–1189. DOI: 10.18653/v1/2021.findings-acl.100. URL: https://aclanthology.org/2021.findings-acl.100.

[275] Samuele Lo Piano. “Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward”. In: *Humanities and Social Sciences Communications* 7.1 (2020), pp. 1–7.

[276] Jan Lorenz et al. “How social influence can undermine the wisdom of crowd effect”. In: *Proceedings of the national academy of sciences* 108.22 (2011), pp. 9020–9025.

[277] Zhuoran Lu and Ming Yin. “Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445562. URL: https://doi.org/10.1145/3411764.3445562.

[278] Ana Lucic, Hinda Haned, and Maarten de Rijke. “Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 90–98. ISBN: 9781450369367. DOI: 10.1145/3351095.3372824. URL: https://doi.org/10.1145/3351095.3372824.

[279] Qin Lyu et al. “Hybrid Ranking Network for Text-to-SQL”. In: *ArXiv* abs/2008.04759 (2020).

[280] Jianqiang Ma et al. “Mention Extraction and Linking for SQL Query Generation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6936–6942. DOI: 10.18653/v1/2020.emnlp-main.563. URL: https://aclanthology.org/2020.emnlp-main.563.

[281] Zilin Ma and Krzysztof Z. Gajos. “Not Just a Preference: Reducing Biased Decision-Making on Dating Websites”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517587. URL: https://doi.org/10.1145/3491102.3517587.

[282] Pattie Maes. “Agents that reduce work and information overload”. In: *Readings in human–computer interaction*. Elsevier, 1995, pp. 811–821.

[283] Keri Mallari et al. “Do I Look Like a Criminal? Examining How Race Presentation Impacts Human Judgement of Recidivism”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376257. URL: https://doi.org/10.1145/3313831.3376257.

[284] James G March. *A primer on decision making: How decisions happen*. Simon and Schuster, 1994.

[285] Gary Marchionini. “Exploratory Search: From Finding to Understanding”. In: *Commun. ACM* 49.4 (2006), pp. 41–46.

[286] Gary Marchionini and Anita Komlodi. “Design of interfaces for information seeking”. In: *Annual review of information science and technology* 33 (1998), pp. 89–130.

[287] Judith Masthoff and Amra Delić. “Group Recommender Systems: Beyond Preference Aggregation”. In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, and Bracha Shapira. New York, NY: Springer US, 2022, pp. 381–420. ISBN: 978-1-0716-2197-4. DOI: 10.1007/978-1-0716-2197-4_10. URL: https://doi.org/10.1007/978-1-0716-2197-4_10.

[288] Joseph Edward McGrath. “Groups: Interaction and performance”. In: *(No Title)* (1984).

[289] O. Medelyan, E. Frank, and I.H. Witten. “Human-competitive tagging using automatic keyphrase extraction”. In: *Proc. EMNLP*. Aug. 2009, pp. 1318–1327.

[290] Bart Mellebeek et al. “Opinion Mining of Spanish Customer Comments with Non-Expert Annotations on Mechanical Turk”. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. CSLDAMT '10. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 114–121.

[291] Gonzalo Mendez, Luis Galárraga, and Katherine Chiluiza. “Showing Academic Performance Predictions during Term Planning: Effects on Students' Decisions, Behaviors, and Preferences”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445718. URL: https://doi.org/10.1145/3411764.3445718.

[292] Stephanie M. Merritt et al. “Are Well-Calibrated Users Effective Users? Associations Between Calibration of Trust and Performance on an Automation-Aided Task”. In: *Human Factors* 57.1 (2015), pp. 34–47. DOI: 10.1177/0018720814561675. URL: https://doi.org/10.1177/0018720814561675.

[293] Stephanie M. Merritt et al. “Are Well-Calibrated Users Effective Users? Associations Between Calibration of Trust and Performance on an Automation-Aided Task”. In: *Human Factors* 57.1 (2015). PMID: 25790569, pp. 34–47. DOI: 10.1177/0018720814561675. eprint: https://doi.org/10.1177/0018720814561675. URL: https://doi.org/10.1177/0018720814561675.

[294] R. Mihalcea and P. Tarau. “TextRank: Bringing Order into Text”. In: *Proc. EMNLP*. 2004, pp. 404–411.

[295] Katherine L Milkman, Dolly Chugh, and Max H Bazerman. “How can decision making be improved?” In: *Perspectives on psychological science* 4.4 (2009), pp. 379–383.

[296] Martijn Millecamp et al. “What’s in a User? Towards Personalising Transparency for Music Recommender Interfaces”. In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’20. Genoa, Italy: Association for Computing Machinery, 2020, pp. 173–182. ISBN: 9781450368612. DOI: 10.1145/3340631.3394844. URL: https://doi.org/10.1145/3340631.3394844.

[297] David Miller et al. “Behavioral Measurement of Trust in Automation: The Trust Fall”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60.1 (2016), pp. 1849–1853. DOI: 10.1177/1541931213601422. eprint: https://doi.org/10.1177/1541931213601422. URL: https://doi.org/10.1177/1541931213601422.

[298] George A. Miller. “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information”. In: *The Psychological Review* 63.2 (Mar. 1956), pp. 81–97. URL: http://www.musanim.com/miller1956/.

[299] Jyoti Mishra, David Allen, and Alan Pearman. “Information seeking, use, and decision making”. In: *Journal of the association for information science and technology* 66.4 (2015), pp. 662–673.

[300] Tina Möckel, Christian Beste, and Edmund Wascher. “The Effects of Time on Task in Response Selection - An ERP Study of Mental Fatigue”. In: *Scientific Reports* 5 (Mar. 2015). DOI: 10.1038/srep10113.

[301] F. Moraes, S.R. Putra, and C. Hauff. “Contrasting Search as a Learning Activity with Instructor-Designed Learning”. In: *Proc. 27th ACM CIKM*. 2018, pp. 167–176.

[302] Lev Muchnik, Sinan Aral, and Sean J Taylor. “Social influence bias: A randomized experiment”. In: *Science* 341.6146 (2013), pp. 647–651.

[303] Kanchan Mukherjee. “A dual system model of preferences under risk.” In: *Psychological review* 117.1 (2010), p. 243.

[304] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. “Survey research in HCI”. In: *Ways of Knowing in HCI* (2014), pp. 229–266.

[305] Cataldo Musto et al. “Exploring the Effects of Natural Language Justifications in Food Recommender Systems”. In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’21. Utrecht, Netherlands: Association for Computing Machinery, 2021, pp. 147–157. ISBN: 9781450383660. DOI: 10.1145/3450613.3456827. URL: https://doi.org/10.1145/3450613.3456827.

[306] Meike Nauta et al. *From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI*. 2022. DOI: 10.48550/ARXIV.2201.08164. URL: https://arxiv.org/abs/2201.08164.

[307] Meike Nauta et al. “From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI”. In: *ACM Comput. Surv.* 55.13s (July 2023). ISSN: 0360-0300. DOI: 10.1145/3583558. URL: https://doi.org/10.1145/3583558.

[308] Joaquin Navajas et al. “Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds”. In: *Nature Human Behaviour* 2.2 (2018), pp. 126–132.

[309] V. Navalpakkam et al. “Measurement and Modeling of Eye-Mouse Behavior in the Presence of Nonlinear Page Layouts”. In: *Proc. 22nd WWW.* 2013, pp. 953–964.

[310] Matteo Negri et al. “Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora”. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing.* Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011, pp. 670–679. URL: https://aclanthology.org/D11-1062.

[311] Dong Nguyen. “Comparing Automatic and Human Evaluation of Local Explanations for Text Classification”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1069–1078. DOI: 10.18653/v1/N18-1097. URL: https://aclanthology.org/N18-1097.

[312] Raymond S Nickerson. “Confirmation bias: A ubiquitous phenomenon in many guises”. In: *Review of general psychology* 2.2 (1998), pp. 175–220.

[313] Mahsan Nourani, Joanie King, and Eric Ragan. “The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8.1 (Oct. 2020), pp. 112–121. DOI: 10.1609/hcomp.v8i1.7469. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7469.

[314] Mahsan Nourani et al. “Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems”. In: *26th International Conference on Intelligent User Interfaces.* IUI ’21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 340–350. ISBN: 9781450380171. DOI: 10.1145/3397481.3450639. URL: https://doi.org/10.1145/3397481.3450639.

[315] Mahsan Nourani et al. “The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 97–105. DOI: 10.1609/hcomp.v7i1.5284. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5284.

[316] Francisco Nunes et al. “Self-Care Technologies in HCI: Trends, Tensions, and Opportunities”. In: 22.6 (Dec. 2015). ISSN: 1073-0516. DOI: 10.1145/2803173. URL: https://doi.org/10.1145/2803173.

[317] H.L. O’Brien et al. “The Role of Domain Knowledge in Search as Learning”. In: CHIIR ’20. 2020, pp. 313–317.

[318] Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. “Deep learning for financial applications: A survey”. In: *Applied Soft Computing* 93 (2020), p. 106384.

[319] M. Pagliardini, P. Gupta, and M. Jaggi. “Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features”. In: *Proc. NAACL HLT*. 2018, pp. 528–540.

[320] Cecilia Panigutti et al. “Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3502104. URL: https://doi.org/10.1145/3491102.3502104.

[321] Andrea Papenmeier et al. “It’s Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI”. In: *ACM Trans. Comput.-Hum. Interact.* 29.4 (Mar. 2022). ISSN: 1073-0516. DOI: 10.1145/3495013. URL: https://doi.org/10.1145/3495013.

[322] Gabriel Parent and Maxine Eskenazi. “Clustering Dictionary Definitions Using Amazon Mechanical Turk”. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. CSLDAMT ’10. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 21–29.

[323] Joon Sung Park et al. “A Slow Algorithm Improves Users’ Assessments of the Algorithm’s Accuracy”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359204. URL: https://doi.org/10.1145/3359204.

[324] Alison Parkes. “The effect of individual and task characteristics on decision aid reliance”. In: *Behaviour & Information Technology* 36.2 (2017), pp. 165–177.

[325] Andisheh Partovi et al. “Relationship between Device Performance, Trust and User Behaviour in a Care-Taking Scenario”. In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’19. Larnaca, Cyprus: Association for Computing Machinery, 2019, pp. 61–69. ISBN: 9781450360210. DOI: 10.1145/3320435.3320440. URL: https://doi.org/10.1145/3320435.3320440.

[326] Andi Peng et al. “What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring”. In: vol. 7. 1. Oct. 2019, pp. 125–134. DOI: 10.1609/hcomp.v7i1.5281. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5281.

[327] Martin Petrin. “Corporate Management in the Age of AI”. In: *SSRN Electronic Journal* (Jan. 2019). DOI: 10.2139/ssrn.3346722.

[328] Jella Pfeiffer. *Interactive decision aids in e-commerce*. Springer Science & Business Media, 2011.

[329] Samuele Lo Piano. “Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward”. In: *Palgrave Communications* 7.1 (2020), pp. 1–7. URL: https://EconPapers.repec.org/RePEc:pal:palcom:v:7:y:2020:i:1:d:10.1057_s41599-020-0501-9.

[330] Peter Pirolli and Stuart Card. “Information foraging.” In: *Psychological review* 106.4 (1999), p. 643.

[331] Massimo Poesio et al. “Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation”. In: *ACM Trans. Interact. Intell. Syst.* 3.1 (Apr. 2013). ISSN: 2160-6455. DOI: 10.1145/2448116.2448119. URL: https://doi.org/10.1145/2448116.2448119.

[332] Ana-Maria Popescu, Oren Etzioni, and Henry Kautz. “Towards a Theory of Natural Language Interfaces to Databases”. In: *Proceedings of the 8th International Conference on Intelligent User Interfaces.* IUI ’03. Miami, Florida, USA: Association for Computing Machinery, 2003, pp. 149–157. ISBN: 1581135866. DOI: 10.1145/604045.604070. URL: https://doi.org/10.1145/604045.604070.

[333] Margaret Potter, Sandy Gordon, and Peter Hamer. “The nominal group technique: a useful consensus methodology in physiotherapy research”. In: *NZ Journal of Physiotherapy* 32.3 (2004), pp. 126–130.

[334] Forough Poursabzi-Sangdeh et al. “Manipulating and Measuring Model Interpretability”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445315. URL: https://doi.org/10.1145/3411764.3445315.

[335] S.R. Putra, F. Moraes, and C. Hauff. “SearchX: Empowering Collaborative Search Research”. In: *Proc. 41st ACM SIGIR.* SIGIR ’18. 2018, pp. 1265–1268.

[336] Minghui Qiu, Y. Li, and Jing Jiang. “Query-Oriented Keyphrase Extraction”. In: *AIRS.* 2012.

[337] Sihang Qiu et al. “Using Worker Avatars to Improve Microtask Crowdsourcing”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–28.

[338] Howard Rachlin. *The science of self-control.* Harvard University Press, 2004.

[339] Tim Rakow and Ben R Newell. “Degrees of uncertainty: An overview and framework for future research on experience-based choice”. In: *Journal of Behavioral Decision Making* 23.1 (2010), pp. 1–14.

[340] Amon Rapp, Lorenzo Curti, and Arianna Boldi. “The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots”. In: *International Journal of Human-Computer Studies* 151 (2021), p. 102630.

[341] Alexander J. Ratner et al. “Snorkel: Rapid Training Data Creation with Weak Supervision”. In: *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases* 11 3 (2017), pp. 269–282.

[342] Pamela Ravasio, Sissel Guttormsen-Sc, and Tscherter. “The Qualitative Experiment in HCI: Definition, Occurrences, Value and Use”. In: (Jan. 2004).

[343] Daniel Read et al. “Choice bracketing”. In: *Elicitation of preferences* (2000), pp. 171–202.

[344] Amy Rechkemmer and Ming Yin. “When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. <conf-loc>, <city>New Orleans</city>, <state>LA</state>, <country>USA</country>, </conf-loc>: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501967. URL: https://doi.org/10.1145/3491102.3501967.

[345] Maria Riveiro and Serge Thill. “The Challenges of Providing Explanations of AI Systems When They Do Not Behave like Users Expect”. In: *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’22. Barcelona, Spain: Association for Computing Machinery, 2022, pp. 110–120. ISBN: 9781450392075. DOI: 10.1145/3503252.3531306. URL: https://doi.org/10.1145/3503252.3531306.

[346] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. “Understanding the Role of Explanation Modality in AI-Assisted Decision-Making”. In: *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’22. Barcelona, Spain: Association for Computing Machinery, 2022, pp. 223–233. ISBN: 9781450392075. DOI: 10.1145/3503252.3531311. URL: https://doi.org/10.1145/3503252.3531311.

[347] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. “Understanding the Role of Explanation Modality in AI-assisted Decision-making”. In: *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 2022, pp. 223–233.

[348] Y. Rong et al. “Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations”. In: *IEEE Transactions on Pattern Analysis &amp; Machine Intelligence* 01 (Nov. 5555), pp. 1–20. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2023.3331846.

[349] Robert Rosenthal and Ralph L Rosnow. *Essentials of behavioral research: Methods and data analysis*. 2008.

[350] J. Rovira, Joan María Senent, and Miquel Àngel Essomba Gelabert. “Educational leadership and teacher involvement as success factors in schools in disadvantaged areas of Spain”. In: *RELIEVE* 22 (2016), p. 4.

[351] N. Roy, F. Moraes, and C. Hauff. “Exploring Users’ Learning Gains within Search Sessions”. In: *Proc. 5th ACM CHIIR*. 2020, pp. 432–436.

[352] N. Roy et al. “Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment”. In: *Proc 6th ACM CHIIR*. 2021, pp. 229–238.

[353] Horacio Saggion and Graeme Hirst. *Automatic Text Simplification*. Morgan &amp; Claypool Publishers, 2017. ISBN: 1627058680.

8

[354] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. “A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making”. In: *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’23. Limassol, Cyprus: Association for Computing Machinery, 2023, pp. 215–227. ISBN: 9781450399326. DOI: 10.1145/3565472.3592959. URL: https://doi.org/10.1145/3565472.3592959.

[355] Mike Schaekermann et al. “Expert Discussions Improve Comprehension of Difficult Cases in Medical Image Assessment”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080. URL: https://doi.org/10.1145/3313831.3376290.

[356] James Schaffer et al. “I Can Do Better than Your AI: Expertise and Explanations”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 240–251. ISBN: 9781450362726. DOI: 10.1145/3301275.3302308. URL: https://doi.org/10.1145/3301275.3302308.

[357] Arno Scharl et al. “Leveraging the Wisdom of the Crowds for the Acquisition of Multilingual Language Resources”. In: May 2012, pp. 379–383.

[358] Nicolas Scharowski et al. “Trust and Reliance in XAI–Distinguishing Between Attitudinal and Behavioral Measures”. In: *arXiv preprint arXiv:2203.12318* (2022).

[359] Max Schemmer et al. “Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making”. In: *arXiv preprint arXiv:2204.06916* (2022).

[360] Eric Schenk and Claude Guittard. “Crowdsourcing: What can be Outsourced to the Crowd, and Why ?” In: 2009.

[361] Nadine Schlicker et al. “Calibrated Trust as a Result of Accurate Trustworthiness Assessment–Introducing the Trustworthiness Assessment Model”. In: (2022).

[362] Anuschka Schmitt et al. *Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice*. en. 2021. URL: https://www.alexandria.unisg.ch/handle/20.500.14171/111308.

[363] Jakob Schoeffer et al. “On the Interdependence of Reliance Behavior and Accuracy in AI-Assisted Decision-Making”. In: *arXiv preprint arXiv:2304.08804* (2023).

[364] Patrick Schramowski et al. “Making deep neural networks right for the right scientific reasons by interacting with their explanations”. In: *Nature Machine Intelligence* 2.8 (2020), pp. 476–486.

[365] Barry Schwartz. “The paradox of choice”. In: *Positive psychology in practice: Promoting human flourishing in work, health, education, and everyday life* (2015), pp. 121–138.

[366] Armin Schwienbacher and Benjamin Larralde. “Crowdfunding of Small Entrepreneurial Ventures”. In: *The Oxford Handbook of Entrepreneurial Finance* (Sept. 2010). DOI: 10.2139/ssrn.1699183.

[367] Javier Fernández Serrano, Silvia T. Acuña, and José A. Macías. “A Review of Quantitative Empirical Approaches in Human-Computer Interaction”. In: *Proceedings of the XV International Conference on Human Computer Interaction*. Interacción ’14. Puerto de la Cruz, Tenerife, Spain: Association for Computing Machinery, 2014. ISBN: 9781450328807. DOI: 10.1145/2662253.2662309. URL: https://doi.org/10.1145/2662253.2662309.

[368] Burr Settles. “Active Learning Literature Survey”. In: 2009.

[369] Tianze Shi et al. “On the Potential of Lexico-logical Alignments for Semantic Parsing to SQL Queries”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1849–1864. DOI: 10.18653/v1/2020.findings-emnlp.167. URL: https://aclanthology.org/2020.findings-emnlp.167.

[370] Ben Shneiderman. *Human-centered AI*. Oxford University Press, 2022.

[371] Katie A Siek et al. “Field deployments: Knowing from using in context”. In: *Ways of Knowing in HCI* (2014), pp. 119–142.

[372] Herbert A Simon. “A behavioral model of rational choice”. In: *The quarterly journal of economics* (1955), pp. 99–118.

[373] Tony Simons, Lisa Hope Pelled, and Ken A Smith. “Making use of difference: Diversity, debate, and decision comprehensiveness in top management teams”. In: *Academy of management journal* 42.6 (1999), pp. 662–673.

[374] B. Škrlj, A. Repar, and S. Pollak. “RaKUn: Rank-based Keyword Extraction via Unsupervised Learning and Meta Vertex Aggregation”. In: *Stat. Lang. & Speech Proc.* 2019, pp. 311–323.

[375] Alison Smith-Renner et al. “No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13. ISBN: 9781450367080. DOI: 10.1145/3313831.3376624. URL: https://doi.org/10.1145/3313831.3376624.

[376] Rion Snow et al. “Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks”. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 254–263. URL: https://aclanthology.org/D08-1027.

[377] Aaron Springer and Steve Whittaker. “Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 107–120. ISBN: 9781450362726. DOI: 10.1145/3301275.3302322. URL: https://doi.org/10.1145/3301275.3302322.

[378] K. Stahl and M. Bravo. “Contemporary Classroom Vocabulary Assessment for Content Areas”. In: *READ TEACH* 63 (Apr. 2010), pp. 566–578.

[379] Keith E Stanovich, Richard F West, and JE Alder. “Individual differences in reasoning: Implications for the rationality debate?-Open Peer Commentary-Three fallacies”. In: *Behavioral and Brain Sciences* 23.5 (2000), pp. 665–665.

[380] Alain D. Starke, Martijn C. Willemsen, and Chris Snijders. “Using Explanations as Energy-Saving Frames: A User-Centric Recommender Study”. In: *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’21. Utrecht, Netherlands: Association for Computing Machinery, 2021, pp. 229–237. ISBN: 9781450383677. DOI: 10.1145/3450614.3464477. URL: https://doi.org/10.1145/3450614.3464477.

[381] David F Steiner et al. “Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer”. In: *The American journal of surgical pathology* 42.12 (2018), p. 1636.

[382] Ivan Dale Steiner. *Group process and productivity*. Academic press New York, 1972.

[383] Joachim Stempfle and Petra Badke-Schaub. “Thinking in design teams-an analysis of team communication”. In: *Design studies* 23.5 (2002), pp. 473–496.

[384] Constantine Stephanidis. “User interfaces for all: New perspectives into human-computer interaction”. In: *User interfaces for all-concepts, methods, and tools* 1.1 (2001), pp. 3–17.

[385] Yu Su et al. “Building Natural Language Interfaces to Web APIs”. In: CIKM ’17. Singapore, Singapore: Association for Computing Machinery, 2017, pp. 177–186. ISBN: 9781450349185. DOI: 10.1145/3132847.3133009. URL: https://doi.org/10.1145/3132847.3133009.

[386] Yibo Sun et al. “Semantic Parsing with Syntax- and Table-Aware SQL Generation”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 361–372. DOI: 10.18653/v1/P18-1034. URL: https://aclanthology.org/P18-1034.

[387] R. Syed and K. Collins-Thompson. “Exploring Document Retrieval Features Associated with Improved Short- and Long-Term Vocabulary Learning Outcomes”. In: *Proc. 3rd ACM CHIIR*. 2018, pp. 191–200.

[388] R. Syed and K. Collins-Thompson. “Optimizing search results for human learning goals”. In: *IRJ* 20 (2017), pp. 506–523.

[389] R. Syed and K. Collins-Thompson. “Retrieval Algorithms Optimized for Human Learning”. In: *Proc. 40th ACM SIGIR*. 2017, pp. 555–564.

[390] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. “Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations”. In: *26th International Conference on Intelligent User Interfaces*. IUI ’21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 109–119. ISBN: 9781450380171. DOI: 10.1145/3397481.3450662. URL: https://doi.org/10.1145/3397481.3450662.

[391] Sarah Tan et al. “Investigating human+ machine complementarity for recidivism predictions”. In: *arXiv preprint arXiv:1808.09123* (2018).

[392] Lappoon R. Tang and Raymond J. Mooney. “Using Multiple Clause Constructors in Inductive Logic Programming for Semantic Parsing”. In: *ECML*. 2001.

[393] Marjorie Templeton and John Burger. “Problems in Natural-Language Interface to DBMS with Examples from EUFID”. In: *Proceedings of the First Conference on Applied Natural Language Processing*. ANLC ’83. Santa Monica, California: Association for Computational Linguistics, 1983, pp. 3–16. DOI: 10.3115/974194.974197. URL: https://doi.org/10.3115/974194.974197.

[394] Loren Terveen, Joseph A Konstan, and Cliff Lampe. “Study, build, repeat: Using online communities as a research platform”. In: *Ways of Knowing in HCI* (2014), pp. 95–117.

[395] Jesse Thomason et al. “Learning to Interpret Natural Language Commands through Human-Robot Dialog”. In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI’15. Buenos Aires, Argentina: AAAI Press, 2015, pp. 1923–1929. ISBN: 9781577357384.

[396] C. Thompson. “Acquiring Word-Meaning Mappings for Natural Language Interfaces”. In: *Journal of Artificial Intelligence Research* 18 (Jan. 2003), pp. 1–44. ISSN: 1076-9757. DOI: 10.1613/jair.1063. URL: http://dx.doi.org/10.1613/jair.1063.

[397] Frederick B. Thompson et al. “REL: A Rapidly Extensible Language system”. In: *ACM ’69*. 1969.

[398] Nava Tintarev and Judith Masthoff. “Beyond Explaining Single Item Recommendations”. In: *Recommender Systems Handbook*. Ed. by Francesco Ricci, Lior Rokach, and Bracha Shapira. New York, NY: Springer US, 2022, pp. 711–756. ISBN: 978-1-0716-2197-4. DOI: 10.1007/978-1-0716-2197-4_19. URL: https://doi.org/10.1007/978-1-0716-2197-4_19.

[399] Suzanne Tolmeijer et al. “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 ’21. Utrecht, Netherlands: Association for Computing Machinery, 2021, pp. 77–87. ISBN: 9781450383660. DOI: 10.1145/3450613.3456817. URL: https://doi.org/10.1145/3450613.3456817.

[400] Suzanne Tolmeijer et al. “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 ’21. Utrecht, Netherlands: Association for Computing Machinery, 2021, pp. 77–87. ISBN: 9781450383660. DOI: 10.1145/3450613.3456817. URL: https://doi.org/10.1145/3450613.3456817.

[401] Richard Tomsett et al. “Rapid trust calibration through interpretable and uncertainty-aware AI”. In: *Patterns* 1.4 (2020), p. 100049.

[402] Wataru Toyokawa, Andrew Whalen, and Kevin N Laland. “Social learning strategies regulate the wisdom and madness of interactive crowds”. In: *Nature Human Behaviour* 3.2 (2019), pp. 183–193.

[403] A Trunk, H Birkel, and E Hartmann. *On the current state of combining human and artificial intelligence for strategic organizational decision making. Bus. Res. 13 (3).* 2020.

[404] Chun-Hua Tsai et al. “Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445101. URL: https://doi.org/10.1145/3411764.3445101.

[405] Aybike Ulusan et al. ““Rather Solve the Problem from Scratch”: Gamesploring Human-Machine Collaboration for Optimizing the Debris Collection Problem”. In: *27th International Conference on Intelligent User Interfaces*. IUI ’22. Helsinki, Finland: Association for Computing Machinery, 2022, pp. 604–619. ISBN: 9781450391443. DOI: 10.1145/3490099.3511163. URL: https://doi.org/10.1145/3490099.3511163.

[406] Kelsey Urgo, Jaime Arguello, and Robert Capra. “The Effects of Learning Objectives on Searchers’ Perceptions and Behaviors”. In: *Proc 6th ACM ICTIR*. 2020, pp. 77–84.

[407] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. “"At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation”. In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW2 (Oct. 2020). DOI: 10.1145/3415238. URL: https://doi.org/10.1145/3415238.

[408] Helena Vasconcelos et al. “Explanations Can Reduce Overreliance on AI Systems During Decision-Making”. In: *Proc. ACM Hum.-Comput. Interact.* 7.CSCW1 (Apr. 2023). DOI: 10.1145/3579605. URL: https://doi.org/10.1145/3579605.

[409] Michael Veale, Max Van Kleek, and Reuben Binns. “Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3174014. URL: https://doi.org/10.1145/3173574.3174014.

[410] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. “How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3476068. URL: https://doi.org/10.1145/3476068.

[411] Sruthi Viswanathan, Behrooz Omidvar-Tehrani, and Jean-Michel Renders. “What is Your Current Mindset?” In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3501912. URL: https://doi.org/10.1145/3491102.3501912.

[412] Georg Von Krogh. “Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing”. In: *Academy of Management Discoveries* 4.4 (2018), pp. 404–409.

[413] Peter P Wakker. *Prospect theory: For risk and ambiguity.* Cambridge university press, 2010.

[414] X. Wan and J. Xiao. “Single Document Keyphrase Extraction Using Neighborhood Knowledge”. In: *Proc. 23rd AAAI.* 2008, pp. 855–860.

[415] Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. “Perspectives on crowdsourcing annotations for natural language processing”. In: *Language Resources and Evaluation* 47 (2013), pp. 9–31.

[416] Bailin Wang et al. “Learning to Synthesize Data for Semantic Parsing”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Online: Association for Computational Linguistics, June 2021, pp. 2760–2766. DOI: 10.18653/v1/2021.naacl-main.220. URL: https://aclanthology.org/2021.naacl-main.220.

[417] Bailin Wang et al. “RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Online: Association for Computational Linguistics, July 2020, pp. 7567–7578. DOI: 10.18653/v1/2020.acl-main.677. URL: https://aclanthology.org/2020.acl-main.677.

[418] Dakuo Wang et al. “Human-AI Collaboration in Data Science: Exploring Data Scientists’ Perceptions of Automated AI”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359313. URL: https://doi.org/10.1145/3359313.

[419] Danding Wang et al. “Designing Theory-Driven User-Centric Explainable AI”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–15. ISBN: 9781450359702. DOI: 10.1145/3290605.3300831. URL: https://doi.org/10.1145/3290605.3300831.

[420] Guangyu Wang et al. “A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images”. In: *Nature biomedical engineering* 5.6 (2021), pp. 509–521.

[421] J. Wang, J. Liu, and C. Wang. “Keyword Extraction Based on Pagerank”. In: *PAKDD.* 2007, pp. 857–864.

[422] R. Wang. “Corpus-independent Generic Keyphrase Extraction Using Word Embedding Vectors”. In: 2015.

[423] Runze Wang et al. “Tracking Interaction States for Multi-Turn Text-to-SQL Semantic Parsing”. In: *AAAI.* 2021.

[424] Xinru Wang and Ming Yin. “Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making”. In: *26th International Conference on Intelligent User Interfaces*. IUI ’21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 318–328. ISBN: 9781450380171. DOI: 10.1145/3397481.3450650. URL: https://doi.org/10.1145/3397481.3450650.

[425] Xinru Wang and Ming Yin. “Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making”. In: *26th International Conference on Intelligent User Interfaces*. IUI ’21. College Station, TX, USA: Association for Computing Machinery, 2021, pp. 318–328. ISBN: 9781450380171. DOI: 10.1145/3397481.3450650. URL: https://doi.org/10.1145/3397481.3450650.

[426] David H.D. Warren and Fernando C.N. Pereira. “An Efficient Easily Adaptable System for Interpreting Natural Language Queries”. In: *American Journal of Computational Linguistics* 8.3-4 (1982), pp. 110–122. URL: https://aclanthology.org/J82-3002.

[427] Edmund Wascher et al. “Frontal theta activity reflects distinct aspects of mental fatigue”. In: *Biological psychology* 96 (Dec. 2013). DOI: 10.1016/j.biopsycho.2013.11.010.

[428] Linda R Weber and Allison Carter. “On constructing trust: temporality, self-disclosure, and perspective-taking”. In: *International Journal of Sociology and Social Policy* 18.1 (1998), pp. 7–26.

[429] M. B. Wesche and T. Paribakht. “Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth.” In: *Canadian Modern Lang. Review* 53 (1996), pp. 13–40.

[430] Lawrence R. Wheeless and Janis Grotz. “The Measurement of Trust and Its Relationship to Self-Disclosure”. In: *Human Communication Research* 3.3 (Mar. 2006), pp. 250–257. ISSN: 0360-3989. DOI: 10.1111/j.1468-2958.1977.tb00523.x. eprint: https://academic.oup.com/hcr/article-pdf/3/3/250/22344414/jhumcom0250.pdf. URL: https://doi.org/10.1111/j.1468-2958.1977.tb00523.x.

[431] R.W. White, S.T. Dumais, and J. Teevan. “Characterizing the Influence of Domain Expertise on Web Search Behavior”. In: *Proc. 2nd WSDM*. 2009, pp. 132–141.

[432] Clifton Wilcox. *Groupthink: an impediment to success*. Xlibris Corporation, 2010.

[433] M.J. Wilson and M.L. Wilson. “A comparison of techniques for measuring sensemaking and learning within participant-generated summaries”. In: *JASIST* 64.2 (2013), pp. 291–306.

[434] Peter Woitek, Paul Bräuer, and Holger Grossmann. “A Novel Tool for Capturing Conceptualized Audio Annotations”. In: *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*. AM ’10. Piteå, Sweden: Association for Computing Machinery, 2010. ISBN: 9781450300469. DOI: 10.1145/1859799.1859814. URL: https://doi.org/10.1145/1859799.1859814.

[435] Robert Wood. “Task complexity: Definition of the construct”. In: *Organizational Behavior and Human Decision Processes* 37 (Feb. 1986), pp. 60–82. DOI: 10.1016/0749-5978(86)90044-0.

[436] Wendy Wood and David T Neal. “A new look at habits and the habit-goal interface.” In: *Psychological review* 114.4 (2007), p. 843.

[437] W. A. Woods. “Progress in Natural Language Understanding: An Application to Lunar Geology”. In: *Proceedings of the June 4-8, 1973, National Computer Conference and Exposition*. AFIPS ’73. New York, New York: Association for Computing Machinery, 1973, pp. 441–450. ISBN: 9781450379168. DOI: 10.1145/1499586.1499695. URL: https://doi.org/10.1145/1499586.1499695.

[438] William Woods, Ronald Kaplan, and Bonnie Webber. “The Lunar Science Natural Language Information System: Final Report”. In: (Jan. 1972).

[439] Anita Williams Woolley et al. “Evidence for a collective intelligence factor in the performance of human groups”. In: *science* 330.6004 (2010), pp. 686–688.

[440] Austin P. Wright et al. “RECAST: Enabling User Recourse and Interpretability of Toxicity Detection Models with Interactive Visualization”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449280. URL: https://doi.org/10.1145/3449280.

[441] S. Paul Wright. “Adjusted P-Values for Simultaneous Inference”. In: *Biometrics* 48.4 (1992), pp. 1005–1013. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2532694 (visited on 09/09/2023).

[442] Tongshuang Wu, Daniel S. Weld, and Jeffrey Heer. “Local Decision Pitfalls in Interactive Machine Learning: An Investigation into Feature Selection in Sentiment Analysis”. In: *ACM Trans. Comput.-Hum. Interact.* 26.4 (June 2019). ISSN: 1073-0516. DOI: 10.1145/3319616. URL: https://doi.org/10.1145/3319616.

[443] Xiaojun Xu, Chang Liu, and Dawn Song. *SQLNet: Generating Structured Queries From Natural Language Without Reinforcement Learning*. 2017. arXiv: 1711.04436 [cs.CL].

[444] Navid Yaghmazadeh et al. “SQLizer: Query Synthesis from Natural Language”. In: *Proc. ACM Program. Lang.* 1.OOPSLA (Oct. 2017). DOI: 10.1145/3133887. URL: https://doi.org/10.1145/3133887.

[445] Fumeng Yang et al. “How Do Visual Explanations Foster End Users’ Appropriate Trust in Machine Learning?” In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 189–201. ISBN: 9781450371186. DOI: 10.1145/3377325.3377480. URL: https://doi.org/10.1145/3377325.3377480.

[446] Xiangli Yang et al. “A Survey on Deep Semi-supervised Learning”. In: *CoRR* abs/2103.00550 (2021). arXiv: 2103.00550. URL: https://arxiv.org/abs/2103.00550.

[447] Yi Yang, Wei Qian, and Hui Zou. “Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models”. In: *Journal of Business & Economic Statistics* 36.3 (2018), pp. 456–470.

[448] W.-T. Yih, J. Goodman, and V.R. Carvalho. “Finding Advertising Keywords on Web Pages”. In: *Proc. WWW.* 2006, pp. 213–222.

[449] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. “Understanding the Effect of Accuracy on Trust in Machine Learning Models”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300509. URL: https://doi.org/10.1145/3290605.3300509.

[450] Kun Yu et al. “Do I Trust My Machine Teammate? An Investigation from Perception to Decision”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces.* IUI ’19. Marina del Ray, California: Association for Computing Machinery, 2019, pp. 460–468. ISBN: 9781450362726. DOI: 10.1145/3301275.3302277. URL: https://doi.org/10.1145/3301275.3302277.

[451] R. Yu et al. “Predicting User Knowledge Gain in Informational Search Sessions”. In: 2018, pp. 75–84.

[452] Tao Yu et al. “CoSQL: A Conversational Text-to-SQL Challenge Towards Cross-Domain Natural Language Interfaces to Databases”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1962–1979. DOI: 10.18653/v1/D19-1204. URL: https://aclanthology.org/D19-1204.

[453] Tao Yu et al. “SParC: Cross-Domain Semantic Parsing in Context”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4511–4523. DOI: 10.18653/v1/P19-1443. URL: https://aclanthology.org/P19-1443.

[454] Tao Yu et al. “Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3911–3921. DOI: 10.18653/v1/D18-1425. URL: https://aclanthology.org/D18-1425.

[455] Tao Yu et al. “SyntaxSQLNet: Syntax Tree Networks for Complex and Cross-Domain Text-to-SQL Task”. In: *EMNLP.* 2018.

[456] Tao Yu et al. “TypeSQL: Knowledge-Based Type-Aware Neural Text-to-SQL Generation”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 588–594. DOI: 10.18653/v1/N18-2093. URL: https://aclanthology.org/N18-2093.

[457] Rachael Zehrung et al. “Vis Ex Machina: An Analysis of Trust in Human versus Algorithmically Generated Visualization Recommendations”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445195. URL: https://doi.org/10.1145/3411764.3445195.

[458] John M. Zelle and Raymond J. Mooney. “Learning to Parse Database Queries Using Inductive Logic Programming”. In: *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*. AAAI’96. Portland, Oregon: AAAI Press, 1996, pp. 1050–1055. ISBN: 026251091X.

[459] Luke S. Zettlemoyer and Michael Collins. “Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars”. In: *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. UAI’05. Edinburgh, Scotland: AUAI Press, 2005, pp. 658–666. ISBN: 0974903914.

[460] Qiaoning Zhang, Matthew L Lee, and Scott Carter. “You Complete Me: Human-AI Teams and Complementary Expertise”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New Orleans, LA, USA: Association for Computing Machinery, 2022. ISBN: 9781450391573. DOI: 10.1145/3491102.3517791. URL: https://doi.org/10.1145/3491102.3517791.

[461] Rui Zhang et al. “Editing-Based SQL Query Generation for Cross-Domain Context-Dependent Questions”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5338–5349. DOI: 10.18653/v1/D19-1537. URL: https://aclanthology.org/D19-1537.

[462] X. Zhang, M. Cole, and N. Belkin. “Predicting Users’ Domain Knowledge from Search Behaviors”. In: *Proc. 34th ACM SIGIR*. 2011, pp. 1225–1226.

[463] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. “Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 295–305. ISBN: 9781450369367. DOI: 10.1145/3351095.3372852. URL: https://doi.org/10.1145/3351095.3372852.

[464] Zelun Tony Zhang et al. “Is Overreliance on AI Provoked by Study Design?” In: *IFIP Conference on Human-Computer Interaction*. Springer. 2023, pp. 49–58.

[465] Chengbo Zheng et al. “Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581131. URL: https://doi.org/10.1145/3544548.3581131.

[466] Yudian Zheng et al. "QASCA: A Quality-Aware Task Assignment System for Crowdsourcing Applications". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. Melbourne, Victoria, Australia: Association for Computing Machinery, 2015, pp. 1031–1046. ISBN: 9781450327589. DOI: 10.1145/2723372.2749430. URL: https://doi.org/10.1145/2723372.2749430.

[467] Victor Zhong, Caiming Xiong, and Richard Socher. *Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning*. 2017. arXiv: 1709.00103 [cs.CL].

[468] Rami Zwick et al. "Consumer sequential search: Not enough or too much?" In: *Marketing Science* 22.4 (2003), pp. 503–519.

Summary

Decision-making has become increasingly intertwined with the use of AI systems to augment human capabilities. The primary goal of human-AI collaboration is to enhance outcomes by leveraging the strengths of both parties. However, as AI systems become more sophisticated, the relationship between humans and AI in decision-making has grown more complex, presenting both challenges and opportunities. For instance, the integration of AI systems can lead to over-reliance, reduced critical thinking, and sub-optimal decision-making outcomes, deviating from the intended benefits. The error rate and biases of AI systems, as well as the user's misunderstanding of the system's capabilities and limitations, can negatively impact the decision-making process and outcomes. To address these issues, researchers have examined various factors that can shape human decision-making behavior and outcomes, including human-related attributes (e.g., cognitive biases, individual differences, expertise), features of AI systems (e.g., transparency, explainability), and contextual factors (e.g., task complexity, uncertainty, time pressure). While many studies have been focused on the human-related factors and features of AI systems, less attention has been paid to the influence of contextual elements on human-AI decision-making.

This work contributes to the growing body of research on human-AI decision-making by empirically investigating the influence of contextual factors on decision-makers behaviors and outcomes. Through a series of studies, we demonstrate that factors such as task complexity, task uncertainty, and group dynamics can significantly impact the adoption of AI systems in decision-making contexts. Over-reliance on AI systems is more prevalent in complex and uncertain tasks, leading to sub-optimal outcomes and reduced critical thinking abilities. Additionally, we found that integrating AI systems can be more beneficial for groups than for individuals, as the collective intelligence and diverse perspectives within a group can enhance critical thinking and decision-making. These findings can inform the design of AI systems and the development of interventions that promote the appropriate use of AI in decision-making, tailored to the specific needs and characteristics of the context.

This thesis also informs the design of future empirical studies that aim to better understand the complex relationship between humans, AI systems, and the surrounding context. While it may not be practical to control all contextual factors in real-world settings, an awareness of their influence can guide the development of rigorous studies that can capture the dynamics of human-AI decision-making in realistic scenarios. Additionally, by proposing a configurable framework, this thesis provides a methodological toolset to enable future researchers to systematically investigate the various factors that contribute to the success of human-AI decision-making.

SAMENVATTING

Besluitvorming raakt steeds meer verweven met het gebruik van AI-systemen om menselijke capaciteiten te vergroten. Het primaire doel van de samenwerking tussen mens en AI is om de resultaten te verbeteren door gebruik te maken van de sterke punten van beide partijen. Naarmate AI-systemen geavanceerder worden, wordt de relatie tussen mens en AI in de besluitvorming echter complexer, wat zowel uitdagingen als kansen met zich meebrengt. De integratie van AI-systemen kan bijvoorbeeld leiden tot te veel vertrouwen, minder kritisch denken en suboptimale besluitvormingsresultaten, waardoor de beoogde voordelen niet worden behaald. De foutmarge en vooroordelen van AI-systemen, evenals het onbegrip van de gebruiker over de mogelijkheden en beperkingen van het systeem, kunnen het besluitvormingsproces en de resultaten negatief beïnvloeden. Om deze problemen aan te pakken, hebben onderzoekers verschillende factoren onderzocht die het menselijke besluitvormingsgedrag en de resultaten kunnen beïnvloeden, waaronder mensgerelateerde eigenschappen (zoals cognitieve bias, individuele verschillen en expertise), kenmerken van AI-systemen (zoals transparantie en uitlegbaarheid) en contextuele factoren (zoals de complexiteit van taken, onzekerheid en tijdsdruk). Terwijl veel studies zich hebben gericht op de mensgerelateerde factoren en kenmerken van AI-systemen, is er minder aandacht besteed aan de invloed van contextuele elementen op de besluitvorming tussen mens en AI.

Dit werk draagt bij aan het groeiende corpus van onderzoek naar mens-AI-besluitvorming door de invloed van contextuele factoren op het gedrag en de resultaten van besluitvormers empirisch te onderzoeken. Door middel van een reeks studies tonen we aan dat factoren zoals taakcomplexiteit, taakonzekerheid en groepsdynamiek een significante invloed kunnen hebben op de adoptie van AI-systemen in besluitvormingscontexten. Overdreven vertrouwen op AI-systemen komt vaker voor bij complexe en onzekere taken, wat leidt tot suboptimale uitkomsten en verminderd kritisch denkvermogen. Daarnaast ontdekten we dat de integratie van AI-systemen gunstiger kan zijn voor groepen dan voor individuen, omdat de collectieve intelligentie en verschillende perspectieven binnen een groep het kritisch denken en de besluitvorming kunnen verbeteren. Deze bevindingen kunnen bijdragen aan het ontwerp van AI-systemen en de ontwikkeling van interventies die het juiste gebruik van AI in besluitvorming bevorderen, afgestemd op de specifieke behoeften en kenmerken van de context.

Dit proefschrift levert ook informatie voor het ontwerpen van toekomstige empirische studies die als doel hebben de complexe relatie tussen mensen, AI-systemen en de omringende context beter te begrijpen. Hoewel het misschien niet praktisch is om alle contextuele factoren in real-world omgevingen te controleren, kan een bewustzijn van hun invloed de ontwikkeling van rigoureuze studies leiden die de dynamiek van mens-AI besluitvorming in realistische scenario's kunnen vastleggen. Daarnaast biedt deze proefschrift, door een configureerbaar raamwerk voor te stellen, een methodologische toolkit waarmee toekomstige onderzoekers systematisch de verschillende factoren kunnen onderzoeken die bijdragen aan het succes van de besluitvorming tussen mens en AI.

Acknowledgements

As I reflect on the completion of my doctoral journey, I am filled with a deep sense of gratitude and appreciation for the incredible individuals who have assisted me throughout this endeavor. Earning a PhD is often described as a solitary and challenging experience, but I have been fortunate to have a network of individuals who have provided invaluable guidance and encouragement.

First and foremost, I would like to express my sincere gratitude to my doctoral co-promotor, Ujwal, for his constant support, intellectual guidance, and steadfast belief in my abilities. You have been an exceptional mentor, pushing me to think critically, challenge assumptions, and reach new heights in my academic pursuits. I have never felt alone in this process, as you have been a constant source of patience and kindness. I have always seen you smiling and continuing to cheer me on, even during the most daunting moments. You have been far more than just an advisor to me; you have been a true friend and valued collaborator throughout this journey. I would also like to express my heartfelt appreciation to my promotor, Arie, whose expertise and insights have been invaluable in shaping the direction and quality of my research. You have consistently provided constructive feedback and challenged me to expand my perspectives, which has been crucial to the successful completion of this dissertation. You always addressed my questions and concerns with understanding, providing the true support I needed to navigate the complexities of doctoral research. In addition to my supervisory team, I would like to acknowledge the invaluable contributions of Geert-Jan, the head of the WIS group, who was consistently available to offer advice and wisdom throughout my PhD program. Your leadership has been instrumental in helping me develop both as a researcher and as an individual. I also want to thank Claudia, who provided me with the opportunity to join the WIS group and accompanied me in the initial phase of my doctoral studies. I am also deeply grateful to the members of my dissertation committee, Prof. Dr. Martijn Warnier, Prof. Dr. Abraham Bernstein, Prof. Dr. Katrien Verbert, and Prof. Dr. Simone Stumpf, for reviewing my work, providing insightful feedback and challenging me to refine and strengthen my research.

In addition, I would like to acknowledge the camaraderie of my fellow PhD colleagues within the WIS group. Our discussions, collaborations, and shared experiences have been invaluable parts of my growth and development as a researcher. Feeling that I was not alone in this journey has been a source of immense strength and motivation, especially during the lock-downs and challenges of the COVID-19 pandemic. Thanks to all my colleagues, especially Alessandro, Asterios, Avishek, Christoph, Rihan, Sole, David, George, Jurek, Sarah, Venktesh, Aditya, Alisa, Andra, Arthur, Christos, Danning, Felipe, George, Gustavo, Kyriakos, Lijun, Manuel, Marcus, Nirmal, Peide, Petros, Robin, Sepideh, Shabnam, Shahin, Tim, Wenbo, Ziyu. I am also very thankful to the Kappa team for the wonderful work and for the fun breaks we have shared throughout this process. Although I joined the team relatively late, you have all welcomed me with open arms and

made me feel like an integral part of the group. I always looked forward to our meetings and discussions, as they offered a much-needed respite from the rigors of research and writing. I remember fondly the laughter, the creative exchanges, and the genuine sense of community that we cultivated together. As the team grew larger, it increasingly felt like a second family to me. I want to thank Jie, Agathe, Anne, Esra, Gaole, Garrett, Lorenzo, Philip, and Shreyan deeply during my time with the Kappa team. Thank you to Nadia and Daphne, who have consistently gone above and beyond to provide invaluable organizational assistance and a warm welcome during my time at the department.

Having had the privilege of doing my PhD as a part of AI for Fintech Research at ING, I express my sincere appreciation for the collaborative, nurturing, and intellectually stimulating environment that the lab has fostered. The various discussions, social events, and opportunities to interact with industry leaders organized by the lab team have been instrumental in broadening my perspectives and enriching my research journey. I am very lucky to have two home-like environments, the WIS group and the AI for Fintech Research lab, which have provided me with a strong sense of belonging and community throughout my doctoral studies. I want to extend a special thanks to Luis, Elvan, Lorena, Eileen, Floris, Patrick, Leonard, Hadi, Jerry, and Marzi for their guidance and genuine interest in my work. I am also deeply grateful to the ING for providing me with generous funding to pursue my doctoral studies.

I am deeply grateful to my family, who have been a steadfast source of love, encouragement, and support throughout this journey. My parents, Maman and Pedar, have been unwavering pillars of strength, always believing in me and providing a sanctuary where I could recharge and find the motivation to push forward. They have instilled in me the values of diligence, perseverance, and compassion, which have been instrumental to my success. I am profoundly indebted to all of you for being the anchor that has grounded me through the most tumultuous phases of my PhD pursuit. You have always understood the challenges I faced and offered reassurance and comfort when I needed it most. Although we may be physically distant, you have remained close to my heart. To my Grandma, Mamani, thank you for your wisdom, your prayers, and your constant encouragement. While we may not have always been able to express it, your presence has been a profound source of strength and inspiration throughout this journey. The brief time you were able to spend with me has been a wellspring of immense joy and comfort. And to my beloved Husband, Sina, thank you for being my constant companion, my rock, and my greatest cheerleader. Your relentless dedication has been invaluable throughout this journey. You have sacrificed your enjoyment to accompany me and consistently reminded me of the bigger picture when I was lost in the details of my research. As my true partner, you have been instrumental in my ability to complete this dissertation. I am deeply grateful for your role as my closest confidant, always there to lend a listening ear, offer a shoulder to lean on, and provide much-needed laughter and love. The celebrations and treats you organized to keep my spirits high during the challenges of this process will not be forgotten. You have encouraged me to step back, take breaks, and savor the small joys along the way. The time and effort you have invested in reading my work, providing insights, and refining my ideas have been invaluable. I am eternally thankful for your pure love and support, which have helped me become a better researcher and a better person. I am overjoyed that we were able to get married and begin our life together during this pivotal time, making my PhD journey all the more meaningful and special. To my dear friends, near and far, Parmiss, Azin, Marjaneh, Zahra, and Fatemeh, who have been steadfast companions since our childhood. Thank you for the cherished memories we've shared and for always reminding me that there is a world beyond the PhD. Your presence in my life has been a constant source of strength and rejuvenation throughout this journey.

Curriculum Vitæ

Sara Salimzadeh

20-09-1995	Born in Tehran, Iran.

Education

2019–2024	Doctor of Philosophy (PhD), Computer Science Delft University of Technology, the Netherlands
2017–2019	Master of Science (MSc), Computer Science University of Amsterdam, the Netherlands Vrije Universiteit Amsterdam, the Netherlands
2013–2017	Bachelor of Science (BSc), Information Technology Engineering University of Tehran, Iran

Awards

2017	Recognition as Exceptionally Talented Student, University of Tehran
2019	Graduation with Distinction (Cum Laude), University of Amsterdam
2021	Honorable Mention for Best Student Paper, ACM ICTIR Conference

List of Publications

1 **Sara Salimzadeh** and Ujwal Gadiraju. 2024. When in Doubt! Understanding the Role of Task Characteristics on Peer Decision-Making with AI Assistance. In Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 01–04, 2024, Cagliari, Italy. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3627043.3659567

2 **Sara Salimzadeh** and Ujwal Gadiraju. 2024. "DecisionTime": A Configurable Framework for Reproducible Human-AI Decision-Making Studies. In Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24), July 01–04, 2024, Cagliari, Italy. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3631700.3664885

3 **Sara Salimzadeh,** Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision-Making. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3641905.

4 **Sara Salimzadeh**, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), June 26– 29, 2023, Limassol, Cyprus. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3565472.3592959

5 **Sara Salimzadeh,** Ujwal Gadiraju, Claudia Hauff, and Arie van Deursen. 2022. Exploring the Feasibility of Crowd-Powered Decomposition of Complex User Questions in Text-to-SQL Tasks. In Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT '22), June 28-July 1, 2022, Barcelona, Spain. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3511095.3531282.

6 **Sara Salimzadeh**, David Maxwell, Claudia Hauff. 2021. On the Impact of Entity Cards on Learning-Oriented Search Tasks. In Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3471158.3472255. **Honorable Mention for Best Student Paper.**

SIKS Dissertation Series

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UvA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web

19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UvA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UvA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UvA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TU/e), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UvA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VUA), Logics for causal inference under uncertainty

23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VUA), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VUA), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TU/e), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TU/e), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech

30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data

06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VUA), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming

36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OU), Learning visually grounded and multilingual representations

---

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TU/e), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OU), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation-Methods for Long-Tail Entity Recognition Models

12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst

31 Gongjin Lan (VUA), Learning better – From Baby to Better

32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

---

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

22 Sihang Qiu (TUD), Conversational Crowdsourcing

23 Hugo Manuel Proença (UL), Robust rules for prediction and description

24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing

25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications

26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables

27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You

28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games

02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey

03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare

04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework

05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization

06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding

07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics

08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design

09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach

10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines

11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring

12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases

13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge

14 Michiel Overeem (UU), Evolution of Low-Code Platforms

15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining

16 Pieter Gijsbers (TU/e), Systems for AutoML Research

17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation

19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media

21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations

23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents

24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values

25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences

27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences

28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems

29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality

30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays

31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management

32 Cezara Pastrav (UU), Social simulation for socio-ecological systems

33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management

34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change

35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions

02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts

03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations

04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval

05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications

06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment

07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning

08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning

09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing

11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications

12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries

13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation

14 Selma Čaušević (TUD), Energy resilience through self-organization

15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models

16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters

17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight

18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation

19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals

20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning

21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain

22 Alireza Shojaifar (UU), Volitional Cybersecurity

23 Theo Theunissen (UU), Documentation in Continuous Software Development

24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning

25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs

26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour

27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions

28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts

29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education

02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems

03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis

04 Mike Huisman (UL), Understanding Deep Meta-Learning

05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair

06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems

08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation

09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks

10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science

11 withdrawn

12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning

13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence

14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling

15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health

16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems

17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design

18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations

19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior

20 Ritsart Anne Plantenga (UL), Omgang met Regels

21 Federica Vinella (UU), Crowdsourcing User-Centered Teams

22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts

23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution

24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour

25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics

26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUDelft), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making