UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

Integrating Humans and Artificial Intelligence in Diagnostic Tasks: Automation-Related User Experience & Interaction in Explainable AI

Integration von Mensch und Künstlicher Intelligenz bei diagnostischen Aufgaben: Automatisierungsbezogene User Experience & Interaktion in Erklärbarer KI

Dissertation

zur Erlangung des Dr. rer. nat.
an der Universität zu Lübeck

vorgelegt von:

Tim Philipp Peter Schrills

geboren am 26.11.1994 in Neuss

Betreut durch:

Prof. Dr. Thomas Franke

Lübeck, 24. September 2024

First referee: Prof. Dr. rer. nat. Thomas Franke
Second referee: Prof. Dr. Ing. Nele Rußwinkel
Date of oral examination: 10th December 2024
Approved for printing. Lübeck, 20th March 2025

Kurzfassung

Diese Dissertation untersucht die Integration von Menschen und künstlicher Intelligenz (KI) in diagnostische Aufgaben, wobei der Schwerpunkt auf der Benutzererfahrung und der Interaktion in erklärbaren KI-Systemen (XAI) liegt. Im Mittelpunkt dieser Forschung steht die Entwicklung des Konzepts der subjektiven Informationsverarbeitungswahrnehmung (SIPA), das sich mit der Benutzererfahrung bei der automatisierten Informationsverarbeitung befasst. Die Arbeit befasst sich mit der zunehmenden Abhängigkeit von KI bei der Automatisierung der Informationsverarbeitung in kritischen Bereichen wie dem Gesundheitswesen, wo Transparenz und menschliche Aufsicht durch erklärbare Systeme ermöglicht werden können. Auf der Grundlage von Theorien zur Mensch-Automation-Interaktion entwickelt und validiert diese Forschung ein Modell der integrierten Mensch-KI-Informationsverarbeitung. Vier empirische Studien untersuchen die automatisierungsbezogene Benutzererfahrung in verschiedenen Kontexten: digitale Kontaktverfolgung, automatisierte Insulinabgabe, KI-gestützte Mustererkennung und KI-basierte Diagnostik. Die Ergebnisse heben die psychologischen Auswirkungen von KI-Erklärungen auf Vertrauen, Situationsbewusstsein und Entscheidungsfindung hervor. Auf der Grundlage empirischer Erkenntnisse diskutiert diese Dissertation das Konzept der „Diagnostizität“ als zentrale Messgröße für eine erfolgreiche Mensch-KI-Integration und schlägt einen Rahmen für die Gestaltung von XAI-Systemen vor, die die Benutzererfahrung durch Anpassung an die menschliche Informationsverarbeitung verbessern. Die Dissertation schließt mit praktischen Leitlinien für die Entwicklung menschenzentrierter KI-Systeme, wobei die Bedeutung von SIPA, Benutzerbewusstsein, Systemtransparenz und der Aufrechterhaltung der menschlichen Kontrolle in automatisierten Diagnoseprozessen hervorgehoben wird.

Schlüsselwörter

Erklärbare KI
Mensch-KI-Interaktion
Subjektive Informationsverarbeitungswahrnehmung
Diagnostische Aufgaben
Automatisierungsbezogene Nutzererfahrung
Vertrauen in KI und Entscheidungshilfesysteme
Integrierte Informationsverarbeitung

Abstract

This dissertation investigates the integration of humans and artificial intelligence (AI) in diagnostic tasks, focusing on user experience and interaction in explainable AI (XAI) systems. Central to this research is the development of the Subjective Information Processing Awareness (SIPA) concept, which deal with user experience in automated information processing. The work addresses the increasing reliance on AI for automating information processing in critical domains such as healthcare, where transparency and human oversight may be enabled through explainable systems. Drawing on theories of human-automation interaction, this research develops and validates a model of integrated human-AI information processing. Four empirical studies explore automation-related user experience in different contexts: digital contact tracing, automated insulin delivery, AI-supported pattern recognition, and AI-based diagnosis. The findings highlight the psychological impacts of AI explanations on trust, situation awareness, and decision-making. Based on empirical findings, this dissertation discusses the concept of "diagnosticity" as a central metric for successful human-AI integration and proposes a framework for designing XAI systems that enhance user experience by aligning with human information processing. The dissertation concludes with practical guidelines for developing human-centered AI systems, emphasizing the importance of SIPA, user awareness, system transparency, and maintaining human control in automated diagnostic processes.

Keywords

Explainable AI
Human-AI Interaction
Subjective Information Processing Awareness
Diagnostic Tasks
Automation-Related User Experience
Trust in AI and Decision Support Systems
Integrated Information Processing

Acknowledgements

The work on this dissertation would never have been possible without the great support of many people - while I cannot name them all individually, I would like to highlight a few in particular:

First of all, my special thanks go to my doctoral supervisor Prof. Dr. Thomas Franke for the tremendous trust and opportunities I had. The field of research I chose was very young when I began my work and I have great respect for the task that this entailed for him as my doctoral supervisor. I would also like to thank all the other professors at the Institute for Multimedia and Interactive Systems at the University of Lübeck who have supported my work in many ways. I would also like to thank Prof. Dr. Nele Rußwinkel for her willingness to review this thesis.

Furthermore, I would like to thank my colleagues who have supported me in terms of content, emotionally and in every other conceivable way during the preparation of this thesis. In particular, I would like to thank Marvin Sieger, Mourad Zoubir, Lilian Kojan, Christiane Attig and Marthe Gruner, with whom I worked closely. I would also like to thank Prof. Dr. Mattias Heinrich and Prof. Dr. Christian Herzog, who supported me both inside and outside the joint projects, for their commitment.

I would also like to thank all the student and research assistants I worked with during the time of my dissertation, especially Philipp Bzdok, Mona Bickel-Dabadghao, Susanne Kargl, Paul-Luis Derwort, Jonas Jacobi, Michelle Wrage and Luisa Winzer, and also Anton de Vries, Tom Wetterich and Carl Fedrowitz. It makes me proud to have even produced joint publications with some of them.

My thanks go also to Anja Minzlaff, Carola Mohrmann and Jork Milde, who have kept my back free for many tasks - this support cannot be taken for granted.

Furthermore, I would also like to thank Christine Albrecht, Jan Thomaskamp, Karoline Heiwolt and Yannick Häntzschel, Alena Nag, Laura Hellwege, Maik Glatki and Marvin Sieger (even more as a friend than as a colleague) - I couldn't have done it without your constant support.

Without the support of my family, I would not have been able to complete this work, nor would I have made it this far. When I fell ill with diabetes about 23 years ago, it took my parents, Claudia and Jürgen, an enormous amount of strength to look after me and my brother, who was also ill. I am enormously grateful to them for all their dedication and love during this time. I am all the happier that this work

also revolves around technologies in the field of diabetes research. I would also like to thank my brother Felix, who has always encouraged me.

Finally, I would like to dedicate this work to my husband Marco Baldelli. His life as an artist always shows me that there are other, important perspectives besides science. I thank him for his emotional support, his trust, understanding and his patience.

Contents

1 Introduction **1**

2 Theorizing Human-AI Interaction **5**

2.1 Defining AI & Human-AI Interaction 6

2.1.1 Information Processing in Diagnostic Tasks 8

2.1.2 AI is automated information processing 10

2.2 Perspectives on Human-AI Interaction 12

2.2.1 Levels of Automation 12

2.2.2 Interplay of Human and AI capabilities 16

2.2.3 Explainable AI Systems for Diagnostic Tasks 19

2.2.4 Evaluative AI Systems for Diagnostic Tasks 22

2.2.5 How do HAI perspectives guide HAI design? 23

2.3 Explanations in HAI - The role of XAI in Human-Machine interaction 24

2.3.1 Development of XAI as a research field 24

2.3.2 Characterizing Metrics of XAI in HAI 27

2.3.3 Empirical research on Diagnostic XAI in HAI 34

2.4 Human Awareness of Automated Information Processing 36

2.4.1 From Situation Awareness to Information Processing Awareness 36

2.4.2 Role of SIPA in Trustworthiness & Controllability of AI Systems 40

3 Present Research **42**

4 Study 1: Examining Automation-Related User Experience in Digital Contact Tracing **44**

4.1 Summary of Study 1 44

4.2 Relevance within the dissertation 45

4.3 Contribution to Study 1 45

5 Study 2: Subjective Information Processing Awareness in Automated Insulin Delivery 63
5.1 Summary of Study 2 . . . 63
5.2 Relevance within the dissertation . . . 63
5.3 Contribution to Study 2 . . . 64

6 Study 3: Assessing the influence of Instructions & Trust in AI-supported pattern recognition 99
6.1 Summary of Study 3 . . . 99
6.2 Relevance within the dissertation . . . 99
6.3 Contribution to Study 3 . . . 100

7 Study 4: Examining how to improve integrated & interdependent information processing 138
7.1 Summary of Study 4 . . . 138
7.2 Relevance within the dissertation . . . 138
7.3 Contribution to Study 4 . . . 139

8 General Discussion 158
8.1 Summary of Results . . . 158
8.2 The Underestimated Role of Diagnosticity . . . 160
8.3 Integrated Human-AI Information Processing for Action Regulation . 164
8.3.1 Integration of Automated Information Processing in Diagnosis 166
8.3.2 Input Adequacy . . . 169
8.3.3 Reference Consonance . . . 171
8.3.4 Output Diagnosticity . . . 174
8.3.5 Adequacy, Consonance & Diagnosticity as determinants of automation-related UX . . . 177
8.3.6 XAI for Perceived Input Adequacy . . . 180
8.3.7 XAI for Perceived Reference Consonance . . . 181
8.3.8 XAI for Perceived Output Diagnosticity . . . 183
8.3.9 What is new about AI for engineering psychology? . . . 185
8.3.10 The Need of Human-aware-systems for Integrated Information Processing . . . 187
8.4 Contributions to Research on Human-Centered XAI . . . 188

8.5 Practical Implications for Diagnostic AI in High Risk Areas 192
8.6 Limitations 193
8.7 Future Research 196

9 Conclusion **198**

10 Bibliography **200**

A List of Figures **230**

B List of Tables **231**

1 Introduction

'The question is no longer how to distribute the fruits of labor fairly, but how to make the consequences of not working bearable.' - Anders (1980, p. 80)

As humans, we have the ability to make diagnostic decisions (Baron et al., 1988), that is, formulate and test hypotheses based on available information to arrive at conclusions about the world (Meder & Mayrhofer, 2017) - we wonder, which clothes are the right choice for a day, which treatment will help us to feel better or what podcast we want to listen to. Sometimes, we use technology to support that decision - be it the weather forecast, a symptom-checker application or a recommendation system for podcasts. Our ability to gather information, analyze it and reach a conclusion is the backbone of what we describe as intelligence (Turner, 1992). However, the human ability to process information is not perfect: we are prone to bias (Norman & Eva, 2010), make mistakes (Nendaz & Perrier, 2012), need to develop skills (Bowen, 2006) and are limited by resources like time and attention (see Wickens et al., 2021) to make a successful diagnosis. In addition, the amount of information humans can memorize and process is limited (Norman & Eva, 2010). Yet, the amount of information available for diagnostic processes has increased immensely (e.g., for cancer detection as described by Levine et al., 2019), especially through new technology in the form of sensors, storage and algorithms, generating more data than humans can process. Along with a rising amount of (digital) information, i.e., more information for diagnostic tasks, the need for technology to handle those tasks leads to the revival of Artificial Intelligence (AI), enabled by the availability of data and the performance of modern processors (Sheikh et al., 2023).

AI automates information processing and changes the role of humans in diagnostic tasks. While human decision makers were previously tasked with acquiring informa-

tion (for example, the initial medical history in the frame of diagnosis), automated systems (and as part of them intelligent sensors technology) are now implemented (Naugler & Church, 2019) for data acquisition. Due to the sheer volume of available data and tasks, pre-processing of information by AI is also desirable, e.g., to mark conspicuous values or cases and thus also intervene in the information analysis. Even diagnostic decisions and subsequent actions are increasingly automated, e.g., in medicine (Göndöcs & Dörfler, 2024; Naugler & Church, 2019) when choosing the best available treatment, leading to a high level of automation of medical AI systems. Since AI systems take over information processing, it changes the human role profoundly: we may no longer review all relevant information (and, indeed, we cannot), nor do we assess the relevance of various factors ourselves. Instead, we rely on automated recommendations for diagnosis.. That implies, the design and implementation of AI systems can result in a loss of human ability to control and be responsible for diagnostic processes (see e.g., Ahmad et al., 2023; Shneiderman, 2020b). Accordingly, various governmental institutions have issued regulations and requirements for AI in critical domains as the medical domain (see European Union, 2024; Ho et al., 2019). At the time of writing this dissertation, the regulation with the largest affected market is the European Union Artificial Intelligence Act (AIA), which places a number of requirements on AI systems in the medical field (Gilbert, 2024), including but not limited to requirements such as transparency, human oversight, and traceability.

One approach to meet the ethical and legal requirements of AI applications (e.g., legal requirements of the AIA) is based on in the research and development of explainable AI (XAI, Gunning and Aha, 2019). XAI aims to improve the comprehensibility of AI systems and their results through suitable explanation methods such as salience maps (Borys et al., 2023), counterfactual explanations (Stepin et al., 2021) or Shapley Values (H. Chen et al., 2021). However, the question to what extent explanations must also be adapted to humans in order to limit the potential risks of AI-based automation was brought more into focus with human-centered XAI (HCXAI, Ehsan et al., 2021). HCXAI characterizes the same question that forms the basis of this dissertation: How do we need to integrate AI systems with human users to achieve optimal automation-related user experience and improve interaction between humans and AI systems (Ehsan et al., 2021), e.g., to achieve performance, quality and satisfaction? This

question is - especially in medical systems - central to enabling a human-centered and responsible use of AI systems (see Shneiderman, 2020a). Whether explanations are an effective way to achieve human-AI integration constitutes a substantial research gap.

Specifically, when this work was started, the effect of AI-based explanations on human decision-making in diagnostic tasks was largely unexplored. This was also due to a lack of appropriate operationalizations and measures (Hoffman et al., 2023) for interdisciplinary research of XAI's effect on human users (Weitz, 2022). Therefore, the first aim of this thesis was to embed the psychological effects of XAI in the broader existing literature on human-automation interaction and to examine concepts for the evaluation of XAI systems. For this purpose, the research presented in this work was based mainly on existing research on situation awareness (M. R. Endsley, 1995), information processing (see Parasuraman et al., 2000), and action regulation (Carver & Scheier, 2000) and transferred to the interaction between humans and AI. Based on the construction of the Subjective Information Processing Awareness (SIPA) scale, empirical studies were conducted, where diagnostic tasks (mainly with a medical focus) were examined. Developed measurement methods were evaluated psychometrically and in terms of their usefulness as an indicator of the quality of HCXAI.

This work comprises four first-authored articles in the field of human-technology interaction, submitted to internationally high-ranking journals or conferences. In addition, the results published in three papers with co-authorship (Attig et al., 2024; Calero Valdez et al., 2024; Wester et al., 2024) and a total of three first-authored conference papers (Schrills & Franke, 2020; Schrills, Zoubir, et al., 2021; Schrills et al., 2023), which are thematically related to the research question of this dissertation, are integrated into the synopsis and cited, where appropriate.

In the synopsis I first discuss the definition of human-AI interaction based on an interdisciplinary perspective and how existing research on human-automation interaction can guide research in human-AI interaction (see section 2). Building on this foundation, the development of XAI as an answer to challenges of human-AI interaction (see section 2.3), metrics to evaluate XAI (see section 2.3.2) and an overview of existing XAI research (see section 2.3.3) are described. Four key research

articles (see sections 4-7) are presented to demonstrate the research findings and motivate the design of the presented model of integrated information processing.

The subsequent chapters integrate the results of the studies presented in this dissertation and introduce the concept of diagnosticity as a central metric to achieve successful integration of human and machine information processing (see section 8.2). In addition, a conceptual model depicting central components of integrated information processing of humans and AI systems is derived and discussed (see section 8.3). Finally, the implications of the presented studies and the theoretical conclusions are discussed (see section 8.5).

List of published/submitted papers:

1. Schrills, T., & Franke, T. (2023). How do users experience traceability of AI systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) systems. *ACM Transactions on Interactive Intelligent Systems*, 13(4), 1-34.

2. Schrills, T., Kojan, L., Gruner, M., Calero-Valdez, A., & Franke, T. (2024) Effects of User Experience in Automated Information Processing on Perceived Usefulness of Digital Contact Tracing Applications: Cross-sectional Survey Study. *JMIR Human Factors*

3. Schrills T., Hoesterey., S., Franke, T., & Roesler, E. (2024) Questioning Trust in AI Research? Exploring the Influence of Trust Assessment on Reliance in AI-assisted Decision-Making. *Submitted to Behaviour & Information Technology*

4. Schrills T., van Berkel, N.n & Franke, T. (2024) Information Interdependence in Human-AI Collaboration: Learners' Perception and Performance in Cooperative Ultrasound Diagnosis. *Submitted to Conference on Human Factors in Computing Systems*

2 Theorizing Human-AI Interaction

In a dissertation on AI, the first challenge is to define the term 'artificial intelligence' to reach a scientific working definition. This challenge has already been addressed in textbooks on AI at an introductory level (Russell & Norvig, 2020), and it is subject to ongoing refinement as new technologies emerge and existing ones evolve (e.g., Thirunavukarasu et al., 2023). In essence, a systematic approach to defining AI entails outlining the characteristics of a system that exemplifies attributes of intelligence (like McCarthy, 2007). Possible definitions may encompass aspects such as the application task of the system, the architecture of the system, or the performance of the system, e.g., in learning (see P. Wang, 2019). The first objective of this section is to develop a working definition of AI that is based on the integration of perspectives of multiple disciplines on AI.

In order to engage in a meaningful discourse on the theory of Human-AI Interaction (HAI), it is imperative to draw upon insights from a multitude of disciplines, including computer science, cognitive psychology, human factors and ergonomics, engineering psychology, and information science (for a discussion of the relevance of psychology in design of human-computer interaction, see Carroll, 1997). Each discipline offers a distinctive perspective, methodology, and set of assumptions that can contribute to a human-centered understanding of HAI, thereby facilitating theory-driven research of XAI systems. To illustrate, computer science provides the technical foundations and algorithms that underpin methods of AI, including perturbation methods (see Bosch et al., 2021). Cognitive psychology offers insights into human information processing for action regulation in decision-making (Estes, 2014), for example, into how information is utilized to evaluate a hypothesis. Engineering psychology, in conjunction with human factors and overlapping with cognitive psychology, is concerned with understanding the interaction between humans and automated systems

(see Poulton, 1966; Wickens et al., 2021). It examines how systems may promote biases or complacency and enhance overall indicators such as performance. The convergence of multiple psychological perspectives can facilitate the development of a robust framework for analyzing and improving HAI, as well as defining AI in the context of this thesis.

2.1 Defining AI & Human-AI Interaction

The process of defining AI in the context of HAI inherently involves balancing different domains, as the definition shapes the scope and focus of HAI research. AI can be defined as the capability of machines to perform tasks that typically require human intelligence, such as learning, reasoning, problem-solving, perception, and language understanding (cf. Minsky, 1961). However, this definition is flawed given, that human intelligence is difficult to define or, to be precise, so difficult to define, that there is no single definition of intelligence (Jensen III et al., 2022). Conversely, more technology-focused definitions of AI may prioritize specific implementations, such as machine learning, natural language processing, or robotics (cf. Vashishth et al., 2023). The various definitions direct attention to different facets of HAI. For instance, a definition focused on robotics emphasizes the physical interactions and the embodiment of AI systems (see Chrisley, 2003), as well as the impact of anthropomorphic features on HAI. Conversely, a definition based on machine learning highlights the continuous learning and the influence of system adaptation on HAI.

First, a technological-driven discussion of AI is centered around the technology and its capabilities, including AI's capability to learn, act autonomously and adopt, e.g., trough reinforcement learning. That is, AI is mainly discussed as machine learning, which can be defined as most simply the application of statistical models to data using computers. Machine learning uses a broader set of statistical techniques than those typically used in medicine. Newer techniques such as Deep Learning (DL) are based on models with less assumptions about the underlying data and are therefore able to handle more complex data (Miotto et al., 2018). Defining AI based on the technologies implemented, particularly through the lens of machine learning, leads to a discussion of HAI predominantly revolving around humans as recipients of

algorithmic output and the processes involved in generating these algorithms (see also Hardy and Harvey, 2020). Machine learning that is usually defined as a sub-field of AI (Kühl et al., 2020), involves training algorithms on large datasets to recognize patterns and make predictions or decisions without explicit programming for each task. Accordingly, a technology-driven definition of AI restricts the human role as the entity to carry out or observe model training, validation, and deployment. Defining a more observing and fine-tuning role for humans also underscores the importance of transparency of training data sets, as humans are responsible for adjusting the model's creation process or form, but do not interact with it. Biased training data (Mac Namee et al., 2002) is potential sources of errors and a technological definition might focus on technology or data as sources of error. That is, the role of humans as users, their experience with machine learning systems and how their ability to cooperate with machine learning systems affects HAI is less represented in this definition.

An alternative perspective on AI is that of Human Factors, which identifies the tasks and goals that can be pursued by different entities, whether human or machine (Chignell et al., 2023). For instance, the European Union defines AI as 'a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments' (European Union, 2024, Article (3), Paragraph (1)). From one perspective, this definition characterizes AI as a form of adaptive automation, which reflects the manner in which humans may employ AI systems. Conversely, the definition is task-oriented, delineating the potential applications of the system, rather than process-oriented, describing the methodologies or technologies that may be employed to achieve a goal. This perspective portrays AI as systems that can enhance human capabilities, automate repetitive or complex tasks, and facilitate goal achievement (cf. Raisamo et al., 2019). It is noteworthy that automation can be defined as 'a device or system that accomplishes a function that was previously, or could be, carried out (partially or fully) by a human operator' (Parasuraman et al., 2000, p. 287). In examining AI as a form of automation, human factors research seeks to optimize the allocation of tasks between humans and AI with the aim of

enhancing joint human-system performance and user experience. The theoretical and design concepts that are of particular importance in this field include, for instance, mental workload, situation awareness, and human-in-the-loop systems, which entail the integration of humans and AI processes for the purpose of accomplishing tasks (see e.g., Gil et al., 2019). This integration of human and machine processes is frequently subject to analysis, for example in terms of usability, adaptability, and the effectiveness of AI in supporting human decision-making and actions (e.g., Boreak, 2020; Lin et al., 2019; Reverberi et al., 2022).

2.1.1 Information Processing in Diagnostic Tasks

If we consider AI to be a form of automation, it is necessary to identify the specific processes that are being automated. In general, automation as part of AI refers to the processing of information, as will be explained in the following section. It is evident that embodied, physically acting systems are also capable of performing physical tasks, such as cleaning objects, driving a vehicle, or in the medical domain (for example, Pineau et al., 2003). However, it can be argued that these abilities are also underpinned by their ability to efficiently process information.

The academic study of information processing emerged concurrently with the advent of contemporary computer systems in the 1950s, as exemplified by the contributions of Donald Broadbent (see Broadbent, 1965; Broadbent, 1958 and later Broadbent, 1982). Information processing, as defined by Broadbent, describes the acquisition of information (also referred to as input), as well as the subsequent analysis of this information to generate a decision or action (referred to as output). These terms reflect the close relationship between the development of cognitive psychology and computer engineering between 1950 and 1970. During this period, computer systems were used as metaphors (Simon & Newell, 1964) to describe and study cognitive processes. In HAI, information processing can be employed to achieve a variety of objectives, including diagnostic decision-making, which entails identifying and resolving issues based on data analysis (Jussupow et al., 2021).

In a multitude of fields, including healthcare, engineering, and finance, the ability to make accurate and timely diagnostic decisions is of paramount importance, as the

identification of problems can have significant implications (see e.g., Lyratzopoulos et al., 2015). Decision Support System (DSS) thus represent a subset of AI systems that are designed to recommend or support the selection of a specific diagnosis for human users. In the field of healthcare AI, the role of AI in supporting diagnostic decisions encompasses a range of functions, including data aggregation, pattern recognition, hypothesis generation, and hypothesis evaluation or recommendation of specific decisions or actions Liu et al. (2019). In order for effective HAI to occur in this context, it is necessary for AI systems to be not only accurate and reliable, but also transparent and traceable. This allows human users to understand and rely on the AI's output.

Parasuraman et al., 2000 present a four-stage model of human information processing, which provides a framework for understanding the cognitive activities that occur during task performance. The aforementioned stages of information processing can be exemplified in the context of medical diagnosis, whereby the interactions and processing of information by a physician can be observed and analyzed.

The process of information acquisition includes the sensing and registration of input data, which serves to support human sensory processes. This stage encompasses the positioning and orienting of sensory receptors, sensory processing, or preliminary data pre-processing prior to full perception. In the context of a medical diagnosis, a physician will typically commence the process by acquiring information from a range of sources, including the patient's history, the results of a physical examination, and the results of any diagnostic tests that have been conducted. The physician listens to the patient's symptoms, observes the physical signs, and takes comprehensive notes. This initial data collection serves to establish a comprehensive foundation for subsequent analysis, ensuring the gathering of pertinent information.

The second stage, information analysis, describes perception and manipulation of retrieved information within working memory. At this stage, cognitive operations include but are not limited to, integration and inference - however, cognitive operations of information analysis do not yet involve the formulation of a decision. A physician reviews the collected information and compares it with existing medical knowledge in order to consider a range of potential diagnoses, i.e., hypotheses. The physician integrates the various symptoms and test results in order to gain preliminary insight

into the patient's condition. For example, the presence of a high fever, sore throat, and white patches on the tonsils may prompt the physician to consider the possibility of a bacterial infection.

The third stage, decision and action selection, describes the point at which decisions are reached based on the cognitive processing that has occurred in the previous stage, that is, the analyzed data is used to generate and select hypotheses. This stage includes the selection of a decision alternative. Based on the analysis, the physician formulates a potential diagnosis and selects an appropriate treatment plan or further diagnostic tests to confirm the diagnosis and thereby increase the certainty of the diagnosis. For example, if the symptoms and preliminary test results suggest the possibility of a bacterial infection, the physician may choose to prescribe antibiotics or order a blood test for further confirmation.

The final stage, action implementation, describes the implementation of a response or action that is consistent with the decision reached. This stage encompasses the actual execution of the selected action. Subsequently, the physician will prescribe the appropriate medication, schedule follow-up appointments, or perform the necessary procedure based on the diagnosis. In the event of a bacterial infection being diagnosed, the physician will issue a prescription for antibiotics and provide the patient with instructions, ensuring that they are fully aware of the treatment plan and any subsequent follow-up requirements. This four-stage model offers a simplified representation of the complexity of human information processing, while simultaneously providing a useful framework for practical applications, such as the design of automated systems in diagnostic tasks.

2.1.2 AI is automated information processing

In the present work AI is primarily discussed as the automation of information processing, which encompasses a broad range of activities from data collection and analysis to decision support and, eventually, leads to the execution of actions. However, without including a capability-oriented discussion of automated information processing systems (i.e., concerning the complexity of a problem), many systems one would not expect to find within the AI definition, would indeed be defined

as AI (e.g., a calculator). That is, defining AI solely as automated information processing is not sufficient. Wang suggests defining intelligence as 'the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources' (P. Wang, 2019, p. 17), which in turn also distinguishes intelligent from non-intelligent systems: non-AI systems are not able to handle problems, that lack the prerequisites of, e.g., computation (see the initial discussion of computing from Turing, 1936). For a more modern definition, Goodfellow et al. describe computation as 'problems that are easy for people to perform but hard for people to describe formally' (Goodfellow et al., 2016, p. 1). That is, the current description of problems as described by Goodfellow et al. cannot be classified as computable, i.e., there is no sufficient description to deterministically calculate a sufficient result. Describing problems that require AI to be solved in that way does not exclude rule-based systems (as defined, e.g., in Duch et al., 2004), but requires them to handle insufficient input (e.g., when not all symptoms can be described in a diagnostic process), and to be able to still produce effective outputs. Interestingly, when discussing why interpretability is needed in AI system, Doshi-Velez and Kim (2017) also use the term incompleteness to refer to non-exhaustive descriptions of the problems that a system is tasked with solving, which could also describe which tasks can only be addressed by intelligent systems: incompletely described tasks.

In the frame of the present research the following working definition for AI is used: **AI is defined as any machine-based system that carries out automated information processing, operating with insufficient knowledge, data or incomplete description of the task and relevant information**, where insufficient knowledge can usually be related to ambiguous rules to map the input of a system to a definite output. This definition allows for the integration of models of human-automation interaction, where the focus is on how automated systems can assist or replace human involvement in processing information and making decisions, especially when resources or knowledge are scarce. The following section provides an overview of different perspectives on HAI, based on AI as the automation of information processing in contexts characterized by insufficient knowledge or resources.

2.2 Perspectives on Human-AI Interaction

The investigation of automated information processing becomes particularly intriguing when human interaction is involved. The term 'interaction' may be defined as a transaction be- tween two entities, typically an exchange of information, but it can also be an exchange of goods or services (Saffer, 2010), while interaction can be described through more detailed lenses, e.g., interaction as experience (Hornbæk & Oulasvirta, 2017). In accordance with Schmidt (2020, p. 2), the interaction between humans and AI is defined by 'interactive exploration and manipulation in real time' and is 'designed with a clear purpose for human benefit while being transparent about who has control over data and algorithms'. The following sections seek to gradually integrate the psychological concepts associated with Schmidt's definition into existing theoretical frameworks and develop them up to the questions of how explanations work in AI.

2.2.1 Levels of Automation

A fundamental premise of this dissertation is that automation is not a dichotomous concept; a task cannot be accomplished by either a single machine or a single human being. Rather, the execution of a task can be an integrated process involving multiple partners, including both human and machine.

This is reflected in the concept of degrees of automation, which represents a widely used framework for structuring human interaction with automated systems. Originally, this framework was created by assigning various tasks to humans and machines (e.g., Sheridan et al., 1978). The supervision of humans over intelligent controllers as a solution to delays in tele-operative control of submarines was already a topic of discussion in 1967 (see Ferrell and Sheridan, 1967). However, the tasks described by Sheridan were initially limited to the question of how humans enact control over the movement of submarines. Accordingly, the potential outcomes of the interaction were constrained to the movement of those vehicles, rather than, for instance, diagnostic categorization. In a broader discussion of machine-assisted information processing, (Sheridan, 1992) subsequently focused on the human capacity to ascribe diagnostic

value to information and utilize it in the context of diagnostic decisions. It is therefore necessary to introduce the concept of diagnosticity. Diagnosticity is central to this dissertation and describes the ability of a source of data to enable one to distinguish between hypotheses (see 8.2).

The intensified research on human-automation interaction between the 1970s and 1990s resulted in the development of models of human-automation interaction that varied in their level of detail. The increasing integration of human and machine information processing, as well as the interdependence and observation between human and machine in interaction, also constituted a central element in the development of theoretical models. For example, models of human-machine interaction are illustrated by multiple mirror loops representing supervisory control of machines (see Sheridan, 1992). In this context, loops represent iterative interactions, whereby feedback or reactions are exchanged between humans and machines. One frequently cited model is the 'Levels of Automation' LOA, which has been published in various versions. The ten-level framework proposed by (Parasuraman et al., 2000) represents the level of detail in the allocation of tasks and responsibilities described in earlier studies. However, different LOA models with varying numbers of levels exist, for example, the Society of Automotive Engineers defines five stages, accommodating different scenarios of human-automation interaction (as discussed in Hopkins and Schwanen, 2021).

The levels of automation are particularly pertinent in the context of the exercise of control and the distribution of control between humans and machines, particularly in the case of this work which employs AI. Control can be defined as 'the process of applying energy and information according to rules in order to make specified system responses conform as closely as possible to some standard or criterion' (Sheridan & Ferrell, 1974, p. 171), where rules describe a defined set of actions that can be performed. The definition from Sheridan and Ferrell (1974) is more specific than that of the Deutsches Institut für Normung (DIN, German Institute for Standardization), which defines control in the light of AI management as 'any process, policy, device, practice or other conditions and/or actions which maintain and/or modify risk' (see "BS ISO/IEC 42001:2023: Information Technology — Artificial Intelligence — Management System", 2023, p. 5). Without delving into the syntactic debate

surrounding the possibility of control without a goal, it is important to acknowledge that the exercise of control in a shared information processing is inherently defined by the fact that the resources provided by humans bring the system into a state that is preferred by humans. This is why the narrower definition proposed by Sheridan and Ferrell (1974) is more suitable for the present dissertation. The concept of control can be augmented with the concept of autonomy, which can be defined as the absence of external control (Deci & Ryan, 1987). When discussing the controllability of AI systems, it is essential to consider these systems as systems that act with varying degrees of autonomy but can be modified by human influence.

A substantial corpus of research has been conducted based on the LOA frameworks, with the objective of examining the impact of different levels of automation on human attention, trust, and performance, which are considered central variables for the successful integration of automation into human tasks Onnasch et al., 2014; Schaefer et al., 2016; Wickens et al., 2010). For instance, empirical results highlight the necessity to strike a balance between automation and human involvement to prevent issues such as automation complacency, whereby an overreliance on automated systems can result in a reduction in vigilance and degradation of performance when manual intervention is required , e.g. described by Molloy and Parasuraman, 1996. This aspect is of utmost importance in safety-critical environments such as aviation, but also healthcare, where human oversight (i.e., the ability of humans to effectively control) remains essential despite high levels of automation. Prior research on human-automation interaction has also demonstrated that the introduction of automation has a significant impact on psychological variables of individuals using automation, e.g., workload, skill, trust, and situation awareness (Parasuraman et al., 2000). Given that humans interact with automated systems, the appropriate level of automation depends on the task. Therefore, selecting the appropriate level of automation based on the task is necessary to avoid undesirable effects.

In addition to the levels of automation, the stage of information processing at which the automation is integrated can vary between different systems. As described above, in the model from Parasuraman et al. (2000) a distinction is made between four different stages of information processing: Information Acquisition, Information Analysis, Decision Selection and Action Implementation. In the case of a medical

diagnosis, for example, a system may be involved in automatically generating sensor data (see E. M. Miller, 2020) that is used for the diagnostic process, leaving the analysis and diagnosis itself under human control (as described in hybrid systems such as Sherr et al., 2023). However, the collected symptoms could also be analyzed by the system and assigned to clinical diagnoses (McLellan et al., 2023), for example. Previous empirical studies have shown that the extent to which automation intervenes in the diagnostic information processing can have a negative impact on the result (see Bahner et al., 2008) - e.g., because failures by the system at this late stage may have a more disruptive effect than in systems in which automation is used at most up to the information analysis stage. That is, acceptance, trust and performance may decline when too much automation is present in later stages of information processing (Onnasch et al., 2014).

Given the significant impact that automation level can have on the outcome of human-automation interaction, it may be advantageous to refrain from letting the automation level be selected by the system itself. Instead, it may be more beneficial to allow users to determine the optimal automation level based on their specific needs and preferences. This is referred to as adaptable automation, which is in contrast to self-adapting automation. The latter is designed to monitor the user's workload or vigilance and adjust the level of automation accordingly. Consequently, the effects of adaptive automation on human performance, situation awareness, and workload in dynamic control tasks have been extensively studied, for example, by Kaber and Endsley (2004). The findings of their research indicated that intermediate levels of automation are conducive to the maintenance of operator involvement and situation awareness, which are pivotal for dynamic and complex task environments (Kaber & Endsley, 2004). The findings indicated that low-level automation had a positive impact on performance, whereas intermediate levels of automation led to enhanced situation awareness. However, this did not always result in improved performance or reduced workload.

In summary, LOA frameworks and autonomy and control distribution as a perspective on AI provide a comprehensive approach to understanding and designing joint information processing by humans and AI. This perspective aims to optimize system performance while maintaining human oversight and engagement by focusing on the

appropriate balance of autonomy and shared information processing. Concurrently, the implications for not only for the individual but also for society have been subjected to debate: For instance, the potential impact of reduced workload on human work has been examined, as well as the ways in which high levels of automation reliability might affect human self-esteem and the extent to which automation could negatively impact human self-fulfillment (Sutton et al., 2018).

As previously discussed, the LOA framework was initially developed with the objective of assigning distinct roles to human users and machines. However, during its evolution, the framework has primarily focused on the concepts of autonomy and control. Nevertheless, alternative models and research have employed LOA to examine the allocation of diverse competencies and capabilities between human and machine actors.

2.2.2 Interplay of Human and AI capabilities

Based on the idea of automation and shared task, a second perspective on HAI encompasses frameworks that highlight the complementary strengths and collaborative potential of human and machine capabilities. The aim of this approach is to investigate to what extent human and mechanical capabilities are suitable for tasks or sub-tasks and how the respective capabilities can be suitably combined. In doing so, the corresponding models also focus strongly on the mutual dependency that exists when solving tasks, e.g., because machines lack relevant sensor technology or humans do not have the time to perform complex computational tasks or comparisons. Of course, this capability-focused perspective cannot be considered completely independently of LOA - because the reason for a certain level of automation is often based on the individual capabilities of a partner (see Roth et al., 2019). Also, individual autonomy plays a role in the evaluation and subsequent assignment of individual capabilities, as interdependencies (see M. Johnson et al., 2014), for example, have to be taken into account. Nevertheless, work that can be assigned to this perspective focus on how the complementarity of two partners can be established. Accordingly, capability-oriented research tries to identify how partners can exchange knowledge or resources, e.g., information about medical history only a system can access or

impressions of a patient's health only observable by a physician. Thus, the concepts discussed below also frequently form the basis for work in the field of human-AI cooperation.

The here described, capability-focused perspective, has changed considerably over the course of the second half of the 20th century. Although Birmingham and Taylor still asserted that 'the man is best when doing least', the concept of complementarity was already being discussed seven years later (Birmingham & Taylor, 1954, p. 1752). Furthermore, the necessity of allocating functions was elucidated in other works (Chapanis, 1965). A number of subsequent research studies employed lists of functions that prioritized either human or machine entities in the performance of specific tasks (see also Fitts, 1951). However, the potential for transformation of specific processes through automation was not fully appreciated. Rather than merely redistributing sub-tasks, automation has led to the emergence of entirely new processes (see Dekker and Woods, 2002). For instance, users of insulin pumps are no longer required to prepare syringes for injection; instead, they must anticipate when they need to refill the insulin into the pump, as this is not predicted by the majority of insulin pumps. However, the evaluation of human tasks was based on parameters of the previously non-automated task, for example, the frequency with which an insulin injection is forgotten or the accuracy of the medication dose calculation. Consequently, a significant number of tasks were based on technical, non-human-centered concepts, such as the consideration of computational performance, but did not take into account metrics that reflect the context of the task and potentially include the cost of adopting higher degrees of automation. One potential limitation of this performance-centric approach to assessing integrated human-machine information processing is the emphasis on individual capabilities. A more fruitful avenue for investigation would be to focus on identifying the optimal solutions for fostering collaboration between diverse partners, including humans and machines (see Dekker and Woods, 2002).

Specifically important for the present dissertation, the Human-Autonomy System Oversight (HASO) model addresses system properties and interaction paradigms such as LOA, adaptive automation, and granularity (M. R. Endsley, 2017). The system properties 'Transparency, Understandability, and Predictability,' which influence

complexity, the mental model, and situation awareness (see M. R. Endsley, 1995), are particularly relevant to the way automation-related user experience is framed in the present dissertation. Situation awareness generally describes the human state of being able to perceive, understand and predict the development of relevant information in a situation. The development of situation awareness in the context of HAI has already been described: for example, the HASO model illustrates the dependency of situation awareness on system properties (e.g., that situation awareness is influenced by the transparency of the model).

The extent to which (explaining) information in AI systems support situation awareness or the formation of mental models has already been described by J. Y. C. Chen et al. (2018) in their work on situation-awareness-based-transparency (SAT). The aim of their approach is to align the information transfer of automated systems in such a way that users have sufficient situation awareness. This is independent of the technical benefit - but focuses on the human ability to act. For example, an automated insulin pump could indicate that it is not injecting insulin due to exercise, thereby increasing a person's situation awareness. In their framework, J. Y. C. Chen et al. (2018) present a more detailed overview of how the individual levels of situation awareness can be supported by explanations - for example, salience maps can help with understanding a system, but not with projecting results.

Although the described frameworks (HASO, SAT) offer valuable insights into the design of systems that enhance situation awareness, there are still some areas where these models lack precision. For instance, the HASO model does not provide clear guidance on the quality metrics that should be used to assess human-centered information processing. It can be reasonably assumed that enhanced information processing has a beneficial impact on factors such as understandability, which in turn affects situation awareness and situation models. However, the precise nature of this relationship remains unclear. The broad definition provided in the HASO model can be attributed to the inherent difficulty in defining many aspects of the HASO model across different domains. Consequently, the work presented in this dissertation primarily relates to (health) diagnostic processes. Another point of discussion in the context of situation awareness is the discrepancy between human experience and behavior, which has been previously identified in the field of artificial

intelligence, particularly in relation to trust and reliance (see Hoesterey and Onnasch, 2023; Papenmeier et al., 2022; Ueno et al., 2023). When applied to AI, there is a risk that users may perceive that they are in the loop of the system and can detect errors, yet fail to exhibit the requisite behavior. This phenomenon, analogous to the illusion of explanatory depth, represents a potential risk that explanations can exacerbate. The second study of this dissertation addresses this risk and discusses the potential issues that can arise from discrepancies in experience and behavior in relation to situation awareness. In the domain of HAI, systems that alter an individual's experience without enhancing their capacity to act present a significant hazard. In particular, XAI is the subject of intense debate, as explanations may exacerbate erroneous assumptions users have about their information processing awareness. The development of LLM (Wei et al., 2022), which are optimized with the help of human reinforcement training (Ouyang et al., 2022) and in which users must be able to recognize hallucinations (Xu et al., 2024), also demonstrates the necessity to bridge the gap between a person's experience of automated information processing and their actions (e.g., performance in the detection of hallucinated information, i.e., data that seems to fit a task but is not adequate).

In summary, understanding users' awareness of information processing involving AI may change users' ability to exert control over the tasks that are automated. Therefore, situation awareness is important to design AI systems in diagnostic tasks. One strategy to keep users in the loop while increasing the level of automation is to offer them more detailed information about the automation or regarding the task - which leads to XAI.

2.2.3 Explainable AI Systems for Diagnostic Tasks

While I address XAI systems as technological artifacts in more detail later (see 2.3), the following section examines the potential impact of explainable AI on the relationship between humans and AI systems.

The automation of the diagnostic process, or the process of selecting decisions, represents a pivotal point at which the positive effects of automation may be undermined, potentially leading to adverse effects on autonomy (Onnasch et al., 2014).

The advancement of sophisticated, precision-oriented AI systems has, in certain instances, resulted in the creation of systems that are more accurate than human experts, as evidenced by developments in the field of medicine (e.g. Cabral et al., 2024; Tu et al., 2024). The high levels of performance in AI systems may result in the development of AI systems that are designed to be used by humans as a replacement for their own diagnostic processes. Accordingly, a recommendation for a diagnosis is provided, yet no integration of human problem-solving processes occurs. In this case, the performance of HAI is contingent upon the user's ability to assess the recommendation provided by the AI system, rather than the diagnostic abilities of the human user regarding the diagnostic task. The recognition that the assessment of AI results by humans can be a challenge due to the opacity of the recommendations and the processes involved was accompanied by a surge in XAI methods (see Adadi and Berrada, 2018; Barredo Arrieta et al., 2020, see also 2.3). Nevertheless, the incorporation of an explanation alongside a recommendation does not negate the impact that recommendation-based diagnostic systems have had on the role of the human user. In other words, XAI facilitated users' capacity to assess AI recommendations but did not integrate them into the diagnostic process. As a result, there is a risk that interaction with explainable AI systems would reinforce biases in human cognitive processes, such as the confirmation bias (Bertrand et al., 2022; Rosenbacke et al., 2024). A critical reflection of the utilization of explanations can also be found in the second study included in the present dissertation.

As discussed, XAI approaches, which are primarily concerned with providing explanations for AI recommendations, may prove ineffective in engaging users to a sufficient extent. This can result in passive acceptance of outputs produced by AI. Cognitive forcing (see Buçinca et al., 2021 is a design approach that aims to foster deeper engagement with AI recommendations by creating interactions that 'forces' users to consider the recommendations. Cognitive forcing addresses the common issues of over-reliance or dismissal due to a lack of understanding or trust.

One effective strategy is to initially refrain from providing recommendations, thereby requiring users to form their own judgments before seeing the AI's suggestions (see Buçinca et al., 2021). This approach encourages users' cognitive engagement with the diagnostic task and prevents users from becoming overly reliant on the

AI's recommendations. An alternative strategy entails the presentation of both supporting and contradictory evidence for an AI suggestion, thereby prompting users to subject the AI's reasoning to critical scrutiny and identify potential shortcomings or biases in its decision-making process (see T. Miller, 2023). Furthermore, interactive explanations are of considerable importance in the context of cognitive forcing: facilitating interaction with the AI's explanatory components, such as the exploration of diverse scenarios or the posing of 'what-if' queries, enhances user engagement and comprehension (discussed by Bertrand et al., 2023). Furthermore, the incorporation of forcing functions, which necessitate that users justify their decisions or predict the AI's output prior to its revelation, ensures that users engage in the requisite cognitive processes for critical evaluation. However, the increased mental workload demanded by cognitive forcing in comparison to direct recommendations may negatively impact the adoption of cognitive forcing as an effective strategy for HAI (see T. Miller, 2023). In empirical studies, cognitive forcing did achieve a lower number of errors or a reduction in complacency effects (Buçinca et al., 2021; Gajos & Mamykina, 2022). However, these were rated as less satisfying and - as already discussed in a similar way as ironies of automation (see Bainbridge, 1983) - in case of doubt lead to an increase in workload instead of a reduction.

In summary, systems that automate the selection of a decision and pre-select answers are prone to increasing the risk of out-of-the-loop status and complacency (as discussed by Onnasch et al., 2014). As posited by T. Miller (2023), the incorporation of supplementary data or rationale into HAI enables the system to adopt a 'recommend-and-defend' strategy, whereby in addition to the recommendation a rationale is presented in the form of an explanation. As previously discussed, a significant shift in task allocation between human and AI involves the human role becoming more focused on evaluating the performance and reliability of the AI system, rather than participating in the original diagnostic task. Despite the fact that the disclosure of accuracy and robustness in AI systems is provided for in legislative works (e.g., European Union, 2024), it can result in the implementation of sub-optimal strategies: for instance, probability matching, as discussed in the third study included in the present dissertation, may be an undesirable consequence of such disclosure. As described by Bartlett and McCarley (2017) is refers to a sub-optimal (and potentially worsening) strategy of humans when utilizing AI information and own information.

In conclusion, despite the potential for XAI approaches to be developed with the intention of preventing this outcome, XAI approaches can nevertheless result in an out-of-the-loop status and the decoupling of information processing and AI evaluation.

2.2.4 Evaluative AI Systems for Diagnostic Tasks

The final perspective on HAI presented in this dissertation is therefore based on the premise that AI systems should only carefully present recommendations for a decision. The idea of HAI described by T. Miller (2023) as 'evaluative AI' aims to provide users with information for evaluating hypotheses when these are requested. How does evaluative AI relate to earlier perspectives of automation processing? First, in evaluative AI systems, the acquisition of information may be automated (i.e., a glucose value is acquired through a tissue sensor). However, the analysis of acquired information is already dependent on the user's hypothesis, i.e., evaluative AI allows users to evaluate multiple hypotheses. Accordingly, evaluative AI offers lower levels of automation and focuses on human-centered information processing in the stage of information analysis. Finally, the decision selection as well as action implementation may not be automated at all (cf. Lyell et al., 2021, where analysis and decision making are the highest ranking types of automation). Onnasch et al. (2014) demonstrated in their analysis, that automation limited to the information analysis may yield higher advantages for users than the automation of decision, especially in automated systems that possibly generate errors. Hence, I argue that to classify as an evaluative AI, systems require awareness of the current hypotheses of the human user or at least must be directable by the user on the basis of their hypotheses (cf. M. Johnson and Bradshaw, 2021 and M. Johnson et al., 2014 for the concept of directability). In addition, the weighing up of alternatives and the presentation of trade-offs is an integral part of evaluative AI, which can be compared with the development of so-called option awareness (Pfaff et al., 2013) and also be beneficial to reduce the risk of out-of-the-loop status. In short, it can be said that evaluative AI seeks to maximize the user's control and ability to perform a diagnostic task. The resulting perspective on HAI is hypothesis-orientated, i.e., it describes how AI and humans deal with hypotheses (about a diagnosis) in the joint processing of information.

The overview of perspectives on HAI given here cannot claim to be exhaustive - the overlap of concepts and theories is sometimes large - and a uniform description of HAI is perhaps neither possible nor desirable, because it fails to reflect the multiple application areas and different approaches that humans integrate automation into their daily lives and into the diagnostic task they pursue. Accordingly, there are already enough models (with LOA, HASO or the evaluative AI framework). A major contribution of this dissertation is to link existing perspectives of HAI with psychological models, as has already been done in automation. To what extent does this require a different model and new terms or concepts?

2.2.5 How do HAI perspectives guide HAI design?

The conceptualization of HAI plays a crucial role when researchers develop guidelines for the development and design of AI systems. For example, the guidelines published in 2019 by Amershi et al. (2019), focus on the lifecycle of AI-systems and do in fewer instances refer to effects HAI can have on users (e.g., discussion of situation awareness or promotion of cognitive biases in decision making). Many of the guidelines presented by Amershi et al. (2019), however, are strongly connected to all perspectives described in the previos section. For example, Guideline *G4* encourages designers to show contextually relevant information and can support the generation of situation awareness. That is, to bridge psychological research and theory, research in Human-Computer Interaction (HCI) and industrial application, it is important to develop models that 1) can serve for research and development in diagnostic tasks, 2) enable communication and discussion across disciplines and 3) enable empirical research, i.e., provide sufficient detail to derive falsifiable hypotheses (see Popper, 2002). Existing guidelines (e.g., Amershi et al., 2019) may discuss desirable features of HAI but are prone to providing designers with sufficiently concrete examples to be transferred into actual design decisions. Accordingly, one goal of my work was to develop a model of integrated Human-AI information processing, that can be applied in different domains and fields of research and is applicable even with low levels of automation of decision-making (i.e., diagnosis) but focused on previous processes (i.e., information acquisition and analysis).

2.3 Explanations in HAI - The role of XAI in Human-Machine interaction

2.3.1 Development of XAI as a research field

Especially since the publication on XAI presented by DARPA in 2017 (see Gunning and Aha, 2019), a great deal of research has been carried out on defining the term itself, related technology, and the effects of XAI on HAI (e.g., Doran et al., 2017; Holzinger, 2018; Palacio et al., 2021. Between 2016 and 2022 there are 22.100 hits on the term 'explainable artificial intelligence' on Google Scholar - in the period between 2010 and 2016 only 412. Even taking into account that the amount of publications generally rises (Fire & Guestrin, 2019), this numbers indicate great interest of the scientific community. In the context of this dissertation, the term 'explanation' of AI systems will first be narrowed down and then technological approaches in the field of XAI will be defined.

In his influential work on XAI, T. Miller (2019) defines explanation as both a process and a product. An explanation is defined as an assignment of causal responsibility, which involves understanding the causes of an event and conveying this causal information to another party. The process of explanation involves abductive reasoning to identify and select causes, while the product is the actual explanation provided. The social aspect of explanation is crucial, as it involves transferring this causal knowledge in a way that the explainee can understand and use them. Explanation, as discussed by T. Miller (2019), is inherently tied to causality, and understanding it involves, e.g., counterfactual reasoning. This means considering what would happen if certain events did not occur, to infer the causal relationships. Explanation also includes elements of interpretability and justification, where the goal is to make decisions understandable to observers, and sometimes to justify why a decision is good without necessarily detailing the decision-making process itself.

In contrast, according to "DIN SPEC 92001-3" (2023), an explanation of an AI is defined as a 'process of describing and communicating important factors influencing an AI system's behavior to a relevant stakeholder'. In view of the discussed perspectives

on HAI, this definition should be expanded to include the fact that an explanation should be suitable for the respective stakeholder in order to influence their decision selection. Without this addition, a stakeholder-related but still technology-centered definition of the explanation would emerge, as 'important factors' should be dependent on the person receiving an explanation. This perspective resonates with the definition of interpretable AI given by Doshi-Velez and Kim (2017), which refers to the 'ability [of a system] to explain or to present in understandable terms to a human' (p. 2). At the same time, the aforementioned extension of the definition would be suitable for defining the point in time at which the character of an explanation can be determined - exactly when the information processing reaches the stage of decision-making. At the same time, it is important to note that this definition of an explanation of AI systems does not include the direction in which an explanation influences the user's performance. That is, explanations may support users but can also negatively influence them, e.g., in terms of accuracy or confidence (see Bansal et al., 2021). Like Doshi-Velez and Kim (2017), the DIN standard "DIN SPEC 92001-3" (2023) also differentiates the concept of interpretability, in which explanations are assigned meanings (i.e., explanations are defined here in a technology-centered way, while the definition of interpretability is human-centered). In the context of this dissertation, however, the DIN definition is not used because explanations can also have an effective impact on people's behavior without being interpretable (e.g., when there is too much information to process but they convince a user, as discussed in study 2 of the present dissertation) or if they are misinterpreted, e.g., when counterfactual explanations are assumed to be true existing comparisons (see Yacoby et al., 2022). Therefore, in a modification of the DIN standard, any information of an AI system that provides information on the processed data, the underlying model, or the output of the system can be defined as an explanation, as long as it is suitable for influencing the subsequent decisions of a human actor.

In addition to works that aim to define XAI, a number of pubilcations deal with the development of a taxonomy for XAI methods (e.g., Arya et al., 2019; Kochkach et al., 2024; Schwalbe and Finzel, 2023), whereby a technically oriented taxonomy seems pointless due to further technical developments (i.e., new technologies are developed too fast to make a useful technology-based taxonomy of XAI). Speith (2022), for example, focuses on the application of explanations. Explanations can differ in six

central areas of how they are applied in any XAI system: 1) the development stage, i.e., explainability is already given by the technical solution approach (Bayesian Network vs. DL), 2) the scope, i.e., whether individual instances of information processing are explained (also referred to as local, see Amparore et al., 2021) or whether the functionality (e.g., weaknesses of the model) is to be explained in general, which are also referred to as global methods. Recently, increasingly mixed approaches have been developed, sometimes described as glocal methods (Achtibat et al., 2023). 3) The functionality of XAI technology, e.g., perturbation in which AI systems are systematically confronted with different inputs in order to determine sensitivities or structural leveraging in which, for example, gradients can be analyzed to detect the relevance of features. 4) The result of the XAI methodology, i.e., feature relevance, is also an important way to distinguish explanation approaches: according to Speith, counterfactual explanations are another well-known example of the representation of feature relevance (see also T. Miller, 2019). Explanations can also be presented in different 5) output formats, e.g., as text, as a visual representation, or through a numerical representation. Finally, Speith distinguishes 6) the type of task for which the system (and thus the explanation) is intended, e.g., as a classification system or as a temporal predictor.

While the taxonomy described by Speith (2022) as well as similar approaches already present a comprehensive structure to understand and study XAI, an important perspective is lacking. In anticipation of the results of this dissertation, an important level must be added to this taxonomy: the extent of the human-awareness (see Kambhampati, 2020) of an explanation, i.e., the extent to which an explanation is directly related to hypotheses currently evaluated by a human user (see T. Miller, 2023) and whether the value of the explanation for the user's current task can be assessed when explaining the system.

The technological implementations of explanations are just as numerous as the technologies used in the field of AI. Rule-based systems in particular, whose rules are understandable and comprehensible for humans, do not require any additional algorithmic processing of information for interpretability (also known as ante-hoc, Sarkar et al., 2022). In transparent, rule-based systems, the rules applied can also be presented directly. In this case, explanation is a matter of information disclosure (see

study 2 of this dissertation) of rules and input values as implemented and discussed in research objective 2, even though the specific visualization and aggregation of information still pose central challenges for human-centered design of explanations. In algorithms whose information processing does not take place on a symbolic level (DNN are a prominent example here), on the other hand, methods are used that can provide explanations of the information processing after model training. Both systems with ante-hoc explainability (see study 1) and systems where this is not the case (see study 2 & Contribution 4) play a role in the results of this work. One reason for the focus on systems that can only be explained ex-ante and require additional methodologies is the so-called explainability-performance trade-off (Crook et al., 2023). This expresses the fact that systems with high explainability (e.g., Bayesian networks) have significantly lower performance or significantly higher resource requirements (in terms of computing resources) than approaches that are opaque without XAI methodology. However, even without this trade-off, the investigation of different methods for establishing explainability is important, e.g., because there may also be conflicts between explainability and security issues (Zhao et al., 2021) or data protection issues (Shokri et al., 2019). At the beginning of this work and with a focus on the action regulation of a user individual, the focus was also on local methods. The global-level XAI methods existing in 2019, when the work on this dissertation began, were intended, i.e., for model revision by developers, for example. In general, global methods can be just as useful for supporting users and are also discussed (in research objective 3).

2.3.2 Characterizing Metrics of XAI in HAI

Existing approaches to assess the quality of XAI (described as XAI metrics, e.g., in Kadir et al., 2023) include technology-centered benchmarks, e.g., pixel flipping (Bach et al., 2015) or remove-and-debias (Rong et al., 2022). While the range of XAI evaluation methods has increased and continues to increase (Nauta et al., 2023), the present dissertation focuses on approaches that assess the subjective experience of explanations, such as Explanation Satisfaction Scale (ESS) (Hoffman et al., 2023). Another approach is the System Causability Scale (SCS) (Holzinger et al., 2020). The SCS is designed to provide a rapid overview of the impact of explanations, thereby

encompassing a range of dimensions. These include the extent to which users perceive explanations as transferable to others and the degree to which they align with their own knowledge base. While this enables a swift general assessment, it remains unclear whether the SCS can also be applied to address specific, theory-driven inquiries, such as those pertaining to the traceability of particular decisions. Neither of the previously described approaches aligns with the aims of this dissertation, as they focus on the direct evaluation of the explanation itself or are designed to guide the evaluation of a system rather than support the understanding of psychological processes. That is, while evaluating users' perception and impression of an explanation is highly important when designing intelligent systems, this dissertation aims to understand the effects of explanations in human information processing and how XAI affects them - accordingly, from my point of view, metrics of XAI must include measures that assess psychological variables from research on automation. Consequently, the present dissertation aims to investigate the effects of XAI on human diagnostic processes, specifically how XAI methods can enhance situation awareness, alter workload, or mitigate complacency (see also Van De Merwe et al., 2024.

Evaluating explanations based on their impact on human experience and behavior, rather than the explanations themselves, offers two significant advantages: 1) It allows for a comparison between a control condition and a condition with an explanation, which is not feasible with direct measures like the ESS (Hoffman et al., 2023). While evaluating satisfaction serves as an excellent starting point for usability studies in XAI, it does not suffice as a metric for this dissertation, because it does not allow for experimentally understanding the effects of explanation (in contrast to, e.g., no explanation). 2) For the applicability of XAI research outcomes, it is crucial to determine how explanations facilitate human oversight or control over a system. While technical evaluations of explanations can indicate whether prerequisites for oversight are met, only experimental studies with human participants can examine whether people can effectively use explanations to exercise control (see Sterz et al., 2024 on conditions of human oversight). 3) Finally, the aim of the explanations should be considered and integrated into the corresponding study, e.g., whether explanations should contribute to the discovery of errors, to better acceptance of decisions or to a generally higher intention to use, as done, e.g., within research objective 1.

Automation-Related User Experience as Metric for XAI

Opposing the idea of directly evaluating the effect of an explanation, XAI can be conceptualized as part of an automated system, i.e., as the capability of an automated system to enrich its results. Accordingly, established constructs from automation psychology (and the perspectives on HAI presented before) can be used for evaluation. Based on Parasuraman et al. (2008), trust, situation awareness, and mental workload have emerged as central constructs for the evaluation of automation. These will be examined in more detail below and their operationalization in the field of XAI will be critically evaluated. In addition, the explanations generated by XAI systems should help users to better interpret the results of a system and be able to utilize them in their tasks (see Guttmann and Ge, 2024). Therefore, the consideration of usefulness also appears valid (see Davis, 1989), as it is also part of various acceptance models (Adams et al., 1992). As already described by Parasuraman et al. (2008), the constructs listed here do not represent a measurement of performance or the like, but extend the evaluation of XAI by adding dimensions that are important for a long-term successful implementation of XAI systems. In the context of this dissertation, constructs discussed within this section are to be explicitly used as measures of people's own experience, which is why, i.e., study 2 also refers to 'perceived trustworthiness'. To summarize, the constructs of perceived trustworthiness, perceived situation awareness, perceived mental workload and perceived usefulness are referred to as automation-related user experience in the context of this dissertation.

Perceived Trustworthiness in HAI. Trust has been extensively discussed in the context of human-human interactions (see Barney and Hansen, 1994; Blomqvist, 1997; Lewicki and Bunker, 1995). Trust, as defined by Mayer et al. (1995) for instance, is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. Their research identifies three critical aspects of trust: ability, benevolence, and integrity. Ability refers to the skills, competencies, and characteristics that enable a party to have influence within a specific domain. Benevolence is the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive, indicating a specific attachment to the trustor. Integrity involves the

trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable. Definition in the technical domain often build upon this definitions of trust and the facets, as described in the given example.

For human-machine interaction, empirical studies demonstrated that trust is related to reliability of automated systems and can influence how users interact with automated systems (e.g., Dzindolet et al., 2003). Lee and See (2004) defined trust as 'the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability' (p. 54). It is important to note, that trust should be conceptualized as an attitude (see Lee and See, 2004) and not a behavior. That is, behavioral or physiological ways to assess trust are always indirect methods (see Kohn et al., 2021). But how must an agent be designed in order to achieve a trusting attitude on the part of the user? The specific description of trust can differ considerably between different definitions. One example is the approach taken by Lee and See (2004), which distinguishes between the factors performance, purpose and process. Performance describes how reliably or accurately the system works. Purpose describes which goals the system is pursuing or the underlying stakeholders, e.g., designers and developers. Process describes the procedures and methods used by the system. In this model, trust depends, for example, on the three factors performance, purpose and process. This is contrasted, for example, by Madsen and Gregor (2000), who give the following definition of trust: 'the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid' (p. 1). The first thing that stands out here is that it is not an attitude, but rather confidence and willingness that are mentioned. This also complements the operationalization of trust by Madsen and Gregor (2000) in two areas: cognition-based and affection-based trust. The former depends more on the perception of system properties - e.g., reliability. Affective-based trust (Madsen & Gregor, 2000) also refers to personal attachment or faith in the system. Lee and See (2004) describe analytical, analogical and affective processes that underlie trust. In addition to these models, Hoff and Bashir (2015) propose a comprehensive model of trust in automation that emphasizes three primary dimensions: dispositional, situational, and learned trust. Dispositional trust refers to the general tendency of a user to trust automation based on their personality and prior experiences. Situational trust is influenced by the specific context in which the automation is used, including

the task characteristics and environmental factors. Learned trust develops over time through interactions with the system, where positive experiences reinforce trust, and negative experiences diminish it. Hoff and Bashir (2015) present a model that may be particularly valuable as it integrates these dimensions to provide a holistic view of how trust in automation evolves, highlighting the dynamic nature of trust and the importance of user experience over time.

Following the variety of definitions and models, questionnaires addressing trust contain different focuses (study 3), with many empirical studies using measurement instruments that have not been validated (Kohn et al., 2021) or without reporting psychometric quality criteria. There is also not always a link to theories, and in some cases the explicit recording of trust using questionnaires is substituted or described as equal to the use of physiological measures, in some cases a distinction is made between trust and reliance (for my point of view, correctly) and trust is examined as a predictor of behavior (e.g., Merritt et al., 2015). These inconsistent and sometimes contradictory conceptualizations (and, in turn, operationalizations) lead to discussions about the impact of the construct trust (e.g., Bolton, 2022) and the connection between trust values collected by questionnaires and behavior. In addition, when indirectly assessing trust via questionnaires, researches may provoke behavior that participants would not demonstrate without the administered trust assessment - this effect is called question-behavior effect (Sprott et al., 2006) and has been demonstrated for attitudes and intentions in, e.g., donating decisions (Godin et al., 2014). As it poses a challenge for valid research on XAI, the presence of the question-behavior effect is addressed in research objective 3. That is, trust continues to play a central role in the ethical and social debate surrounding the use of AI (Díaz-Rodríguez et al., 2023) and XAI (Langer et al., 2021) and is therefore also the subject of this dissertation.

Perceived Situation Awareness in HAI. Compared to measurement methods that use explicit measures (e.g., correct responses in the Situation Awareness Global Assessment Technique (SAGAT), M. Endsley et al., 2000) or implicit measures (e.g., reaction times) to assess actual situation awareness, the measurement of perceived situation awareness is controversial (M. R. Endsley et al., 1998). While Endsley emphasizes on the one hand that perceived situation awareness is decisive for action

regulation, she also notes that the correlation between perceived (or subjective) situation awareness is often low. The Situation Awareness Rating Technique (R. M. Taylor, 2017) is a method for assessing perceived situation awareness that was published as early as 1990. This simultaneously includes other constructs such as workload (Kaber & Endsley, 2004) or resources provided and therefore differs significantly from Endsley's concept of situation awareness. Edgar et al. (2018) calculate Perceived Situation Awareness on the basis of a confidence rating according to the evaluation of true false statements, whereby no major correlations with behavior can be determined here either. Arguably, equating Perceived Situation Awareness and Confidence can be viewed critically, as Situation Awareness develops over three different stages (M. R. Endsley, 1995), which should also be reflected in the elicitation of Perceived Situation Awareness. Hence, the development of a theory-based and structured procedure for recording perceived situation awareness was necessary at the beginning of the work on this dissertation, whereby Endsley already provided initial approaches in the HASO model, for example by highlighting three central system properties of automation (M. R. Endsley, 2017), Transparency, Understandability and Predictability, which are directly related to situation awareness, as discussed later in 2.4.1.

Perceived Mental Workload in HAI. Mental workload is a central construct in the field of automation (see Vidulich and Tsang, 2012). Mental workload can be defined as the amount of cognitive work required for a person to complete a certain task over time (Longo et al., 2022). Mental workload plays a particularly important role for XAI because explanations lead to an ironic effect: while AI is supposed to lead to more efficient information processing through automation, explanations can increase the mental workload again. It is therefore an important metric for the evaluation of XAI. At the beginning of the first studies for this dissertation, the negative effects of explanations (see e.g., Bansal et al., 2021) and the underlying mechanisms were less researched. In the meantime, the results of several studies (e.g., Sewnath and Crijnen, 2021; Tsai et al., 2021) as well as the experiment on AID systems presented in this dissertation (study 2) show that the use of explanations could lead to information overload. The NASA Task Load Index (NASA-TLX, Hart, 2006) has already been widely used in automation research (Ramkumar et al., 2017) and also serves as a metric in the present dissertation.

Perceived Usefulness in HAI. Perceived usefulness, as defined in the Technology Acceptance Model (TAM, Davis et al., 1989), refers to the degree to which an individual believes that using a particular technology will enhance their performance. In the context of XAI, perceived usefulness is particularly significant because it directly affects whether users will accept, trust, and effectively interact with AI systems (Paleja et al., 2022). Unlike general satisfaction, perceived usefulness specifically relates to the practical benefits users derive from the system, making it a critical factor in the successful deployment of XAI.

The role of perceived usefulness is also encapsulated in international norms on usability ("DIN EN ISO 9241-11", 2018). According to DIN 9241, 'usability is defined by the extent to which a system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use'. Perceived usefulness aligns closely with these dimensions, particularly effectiveness and efficiency, as it reflects the users' perception of how well the XAI system assists them in achieving their tasks. An XAI system that is perceived as useful will likely be seen as more effective and efficient, thereby enhancing its overall usability as defined by the standard.

The work of Hoffman et al., 2023 on scales to assess trust in XAI underscores the significance of perceived usefulness in fostering trust. This scale includes items that measure how users perceive the usefulness of AI systems in their tasks, thereby linking usefulness directly to trust. The research of Hoffman et al., 2023 indicates that trust in XAI encompasses both the cognitive and affective dimensions, with cognitive trust being particularly influenced by the perceived utility of the system (cf. Madsen and Gregor, 2000). This suggests that when users find an XAI system useful, their cognitive trust — based on a rational evaluation of the system's capabilities — may be increased, which can lead to greater reliance and acceptance. However, measurement with established methods (e.g., Davis, 1989) is difficult, as there is a strong overlap with other categories, e.g., trust or workload. In the first study presented in this dissertation, for example, a measurement of perceived usefulness is carried out using the DIN/ISO standard ("DIN EN ISO 9241-11", 2018).

2.3.3 Empirical research on Diagnostic XAI in HAI

In addition to technical studies (see P. Q. Le et al., 2023), a number of empirical studies (e.g., van der Waa et al., 2021) and reviews (e.g., Haque et al., 2023) have already been published on the effects of XAI. In the following section, the results available for diagnostic AI and the underlying scientific contradictions are presented concisely. Firstly, the general need for explanations, then studies with positive effects and studies without effects (or with negative implications) are presented.

Users' XAI Demand in Diagnostic Tasks

The demand for explanations in diagnostic AI systems is well-documented in empirical research and clinical feedback. Tonekaboni et al. (2019) emphasize that for AI to be effectively integrated into clinical practice, explainability is crucial. This entails the AI system justifying its outcomes in a manner that assists clinicians in rationalizing the model's predictions. Their survey of clinicians in acute care settings, such as Intensive Care Units and Emergency Departments, highlighted that clinicians view explainability as essential for trust and practical adoption.

Furthermore, Alam and Mueller (2021) investigated the impact of explanations on user satisfaction and trust within AI diagnostic systems, finding a significant positive correlation between the presence of explanations and user satisfaction. This demonstrates that explanations are a critical component in fostering user trust in AI diagnostics. Additionally, Chanda et al. (2024) showed that AI systems with dermatologist-like explanations significantly enhanced trust and confidence among users diagnosing melanoma, further affirming the necessity of transparent and comprehensible AI explanations in medical diagnostics. Also, Bertrand et al. (2022) conducted a systematic review that underscores the role of cognitive biases in XAI-assisted decision-making. They argue that explanations help mitigate potential biases, thereby improving decision accuracy. Bertrand et al. (2022) suggests that users are more likely to demand explanations when cognitive biases are present, as these explanations clarify AI reasoning and reduce uncertainty in decision-making processes.

Positive Effects of XAI on Automation-related User Experience

Explanations in AI systems positively influence several aspects of the user experience, including trust, confidence, and situation awareness. Bansal et al. (2021) found that explanations significantly enhance the performance of human-AI teams, particularly by increasing users' trust in the AI's recommendations. This trust is crucial for effective human-AI collaboration, as users are more likely to rely on and effectively integrate AI recommendations when they understand the underlying reasoning.

Papenmeier et al. (2022) explored the complex relationship between user trust, model accuracy, and explanations. Their research indicates that while model accuracy is paramount, the addition of explanations can bolster user trust even in less accurate models, suggesting that explanations provide a psychological buffer that enhances user confidence in AI systems. Additionally, Silva et al. (2023) examined both the objective and subjective impacts of XAI on human-agent interaction, revealing that explanations not only improve trust but also enhance users' situation awareness. By providing insights into the AI's decision-making process, explanations help users maintain a clearer understanding of the system's operations and potential outcomes, thereby improving their overall interaction experience (Shin, 2021).

Lack of Effects of XAI on Performance

Despite the positive effects of XAI on user experience, empirical evidence suggests that explanations do not always translate to improved performance. Alufaisan et al. (2021) conducted a study assessing whether explainable AI improves human decision-making. Their findings indicate that while explanations can enhance user understanding and trust, they do not necessarily lead to better decision outcomes. Similarly, X. Wang and Yin (2021) compared the effects of explanations in AI-assisted decision-making and found that explanations did not significantly improve decision accuracy or performance. This suggests that while explanations are beneficial for user trust and satisfaction, they may not directly influence the effectiveness of the decisions made with AI assistance.

Schemmer et al. (2022) conducted a meta-analysis on the utility of XAI in human-AI

decision-making, corroborating the notion that the presence of explanations does not consistently enhance performance. Their analysis revealed a nuanced impact of XAI, where the benefits in trust and understanding do not always equate to measurable improvements in decision quality. These findings underscore the complexity of XAI's impact on performance, highlighting the need for further research to understand the contextual factors that influence the efficacy of explanations in enhancing decision-making and performance.

The results of the studies presented above indicate that explanations must be designed in a human-centered way and that negative effects of explanations can be anticipated (see Chromik et al., 2019; Ehsan and Riedl, 2024). The term HCXAI, coined in 2020, can be seen as an appeal, as can the call for evaluative AI (Ehsan & Riedl, 2020): an explanation must not be seen as legitimization, as sufficient compliance with the duty of transparency or even as sufficient accountability for the delegation of responsibility. This is particularly true if the effects of declarations have not been empirically tested in user studies. The legislative framework (e.g., AI Act, European Union, 2024) can also be seen as a call for the design of human-centered AI and XAI (see Valdez et al., 2024).

2.4 Human Awareness of Automated Information Processing

2.4.1 From Situation Awareness to Information Processing Awareness

To effectively address human-centered research questions in diagnostic XAI systems, it is essential to develop instruments that assess traceability-related facets of user experiences. in congruence with this dissertation's goal, one major contribution of my research was the development of a scale, that supports human-centered research in XAI. Like discussed before, existing measurements for explainable AI XAI include various scales developed to evaluate user experience, such as the Explanation Satisfaction Scale (Hoffman et al., 2023) and the System Causability Scale (Holzinger

et al., 2020). These scales often focus on the explanations provided by the system, but do not target changes in automation-related user experience.

However, at the beginning of this work, there was a need for an instrument that measures the experienced effects of explanations and relates to the experienced traceability of automated systems, rather than directly evaluating the explanations themselves. To address this, the SIPA scale was proposed, derived from Situation Awareness theory (see Schrills, Zoubir, et al., 2021; Schrills et al., 2024 as related works). SIPA describes the user's experience of being enabled by a system to perceive, understand, and predict its information processing (Schrills & Franke, 2023). The SIPA scale aims to evaluate the perception of system characteristics based on user interaction, making it a crucial part of assessing automation-related user experience.

From the beginning on (see Schrills, Zoubir, et al., 2021), the concept of SIPA was strongly related to conception of Situation Awareness (M. R. Endsley, 1995): situation awareness identifies three levels of situational assessment: perception, understanding, and projection. These correspond to the SIPA facets of experienced transparency, understandability, and predictability, respectively (see Fig 2.1). Endsley's research highlights the importance of these facets for maintaining situation awareness in human-automation interactions, emphasizing that transparency, understandability, and predictability are critical for the trustworthiness and effectiveness of automated systems (M. R. Endsley, 2017). It is important to note, that transparency in the SIPA scale refers to the system's property to enable its user to perceive all relevant elements for information processing, rather than the goals of the developer or general information about the model's training.

The SIPA scale's items are grounded in situation awareness conception and were developed based on existing situation awareness scales, theoretical explanations, and expert discussions in engineering psychology as described by Schrills, Zoubir, et al. (2021). Initially composed of 12 items, the scale was refined to 6 items through empirical iteration. Reverse-coded items, which can negatively impact scale comprehensibility, were excluded based on qualitative feedback from users. The resulting scale, as published, e.g., in Schrills and Franke (2023), can be found in Figure 2.1.

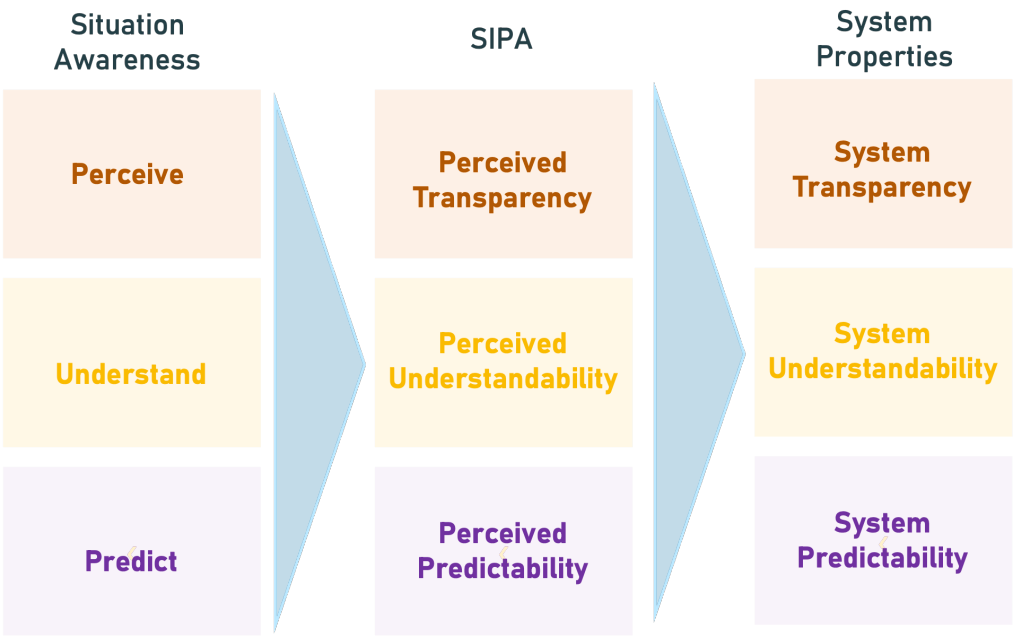


Figure 2.1: Different levels of Situation Awareness, their corresponding levels of SIPA and system properties connected to SIPA levels

Given the close alignment of SIPA with SA, the question arises as to what exactly the difference is between SIPA and SA. In its conception, SIPA (as also discussed in Schrills and Franke, 2023) is a subset of SA. This means that it more specifically describes a certain situation - the information processing by a machine system (in the sense of the definition of the work, therefore, usually by an AI). While situation awareness can also arise in situations in which no information-processing entities other than the person are involved, SIPA is specifically designed to describe situation awareness in connection with the external information processing of a system. In terms of XAI research, this customization of SIPA makes it possible to investigate the effects of AI, which can, for example, act as black boxes (Adadi & Berrada, 2018) to reduce SA. Especially when analysing XAI methods, SIPA and the SIPA scale offer a more specific instrument than SART, for example (R. M. Taylor, 2017). Overall, however, there is a risk that external factors such as mental workload are taken more into account in SART assessments and can therefore be used, for example, in conjunction or in combination with SIPA as an AI and XAI-specific instrument.

Equating SIPA with the correctness of the user's mental model can be seen as a

Table 2.1: Items of the Subjective Information Processing Awareness (SIPA) Scale and the Corresponding Instruction

The following questionnaire deals with your **experience in the interaction with the system. Information** refers to all data that the system can work with. **Result** refers to the output of the system, which is presented at the end of the system's information processing

Please indicate the degree to which you agree/disagree with the following statements		completely disagree	largely disagree	slightly disagree	slightly agree	largely agree	completely agree
SIPA-01	It was transparent to me which information was collected by the system.						
SIPA-02	The information that the system could acquire was observable for me.						
SIPA-03	It was understandable to me how the collected information led to the result.						
SIPA-04	The system's information processing was comprehensible to me.						
SIPA-05	With the information accessible for me, the results was foreseeable for me.						
SIPA-06	The system's information processing was predictable for me.						

contradiction to the necessity of SIPA. It is therefore necessary to discuss both concepts in contrast: SIPA is defined as the perception of system properties and not as self-perception. In the study published by Schrills et al. (2024), it was shown that a SIPA version in which self-related statements (i.e., beginning with 'I am able...' instead of 'The information are...', that is, actively referring to the person versus to the system and its information processing) occur has a correlation with personality measures such as the ATI scale (Franke et al., 2019), other than the one presented here. In principle, however, the ability to correctly predict the behavior of a system, for example, is closely related to the correctness of one's own mental model (see Colin et al., 2022), which is why there is an overlap here. In addition, SIPA also depends on a specific situation and can differ - without learning taking place - between situations with the same system. For example, if the system displays relevant information in one case and the users are aware of it and in another case they are not aware of it. In this case, the perception of the system as transparent may change for a specific situation without there being a change in the mental model.

In conclusion, the SIPA scale developed in this dissertation provides a vital tool for

assessing the subjective effects of explanations in XAI, bridging a critical gap in user experience research in automated information processing and especially in XAI. By aligning with Situation Awareness theory, SIPA offers a nuanced understanding of how systems enable users to perceive, comprehend, and predict system behavior, enhancing the evaluation of XAI systems. It was applied for research objectives 1-3.

2.4.2 Role of SIPA in Trustworthiness & Controllability of AI Systems

Two constructs of the automation related user experience beyond situation awareness are considered in particular in this dissertation: trustworthiness, as already discussed in Chapter X, and controllability, which has already been defined and briefly introduced in Chapter X.

First trustworthiness - according to Schlicker et al. (2022), perceived trustworthiness depends on trustworthiness cues, which must be perceived and utilized by the trustor. These can overlap with information that can also improve the SIPA - e.g., the output of the system, the confidence displayed or information about the input of the system. The importance of trustworthiness cues that are related to a system's information processing demonstrates a connection between SIPA and trustworthiness, as the same cues might demonstrate the capability of a system to handle a task but also support users' understanding of how the system processes information. The concept described by Madsen and Gregor (2000) as cognitive trust also has a close connection with SIPA, e.g., because perceived understandability is an explicit part of the model and belongs to cognitive trust. this is also obvious because cognitive-based trust is defined as 'user's intellectual perceptions of the system's characteristics' (see Madsen and Gregor, 2000, p. 2). The cognitive elements of experienced trustworthiness therefore strengthen a close connection with SIPA. However, compared to the items of Madsen and Gregor, for example, SIPA can again be described as a structured approach, as the levels of situation awareness are taken into account.

In multiple trust models in the field of HCI, affective components are described in addition to the cognitive components (see Bae et al., 2023; Madsen and Gregor, 2000; Ueno et al., 2023), which may arise through 'emotional reactions' of the system.

Schlicker et al. (2022) also describe cues that are not related to the characteristics of the information processing of the system, e.g., the reputation of the manufacturer or social cues such as usage behavior of people in one's own environment. Affective-based components of trust or trustworthiness constitute a clear differences to SIPA, in which emotional reactions do not play a role. Therefore, an overall medium relationship between SIPA and the experience of trustworthiness can be expected (as found in all studies of the present dissertation).

The relationship between controllability and SIPA can be described less clearly. In principle, the predictability of an AI system is important in order to be able extert control over it effectively (Cavalcante Siebert et al., 2023). Similar to SA, the predictability of a system also depends on how understandable and transparent it is (see M. R. Endsley, 1995). In this respect, correct information processing awareness Information Processing Awareness (IPA) is also the basis for human supervision - but what role does SIPA play as an construct of user experience (similar to SART compared to SAGAT)? Human action regulation is guided by one's own perceptions and beliefs (cf. Carver and Scheier, 2000). This means that a user does not make the use of a system dependent on the IPA, but on the SIPA. A particularly high risk for the controllability of a system therefore arises when SIPA and IPA diverge. A related phenomenon has already been described in the XAI literature in particular (e.g., Bansal et al., 2021). Chromik et al. (2021) use the term 'illusion of explanatory depth', which describes that people assume to have understood something better than they actually do (see also Rozenblit and Keil, 2002).

The aim of the research in this dissertation was therefore to understand how SIPA is related to the control of a system (in particular predictability as a proxy for control) and how situations generating an illusion of controllability can be avoided.

3 Present Research

The overall objective of the present dissertation was to examine how the information displayed by an AI system in diagnostic tasks that integrate human and machine information processing affects automation-related user experience and behavior. To this end, empirical studies were conducted to 1) assess the structure of automation-related user experience and its influence on HAI, 2) assess the effects of information disclosure on automation-related user experience and users' prediction performance, 3) assess whether a question-behavior effect affects trust-related AI research and, 4) how novices benefit from instructions regarding interdependence in information processing.

In more detail, the following research objectives were derived:

RO1: the first research objective was to structurally assess the relationship between automation-related user experience (SIPA, trust, usefulness) and use intention as a key indicator of HAI in the context of automated processing of medical data (i.e., digital contact tracing in pandemics).

RO2: The second research objective was to assess the effects of varying levels of information disclosure as form of XAI on users' automation-related experience of the AI system and to examine the relation between UX with prediction performance in the context of an automated drug dosage system (i.e., automated insulin delivery).

RO3: The third research objective was to investigate the presence of the question-behavior effect when eliciting automation-related trust assessment and the (experimental) bias on behavioral data in the context of an abstract AI-assisted pattern recognition task (i.e., identification of geometric rules)

RO4: The fourth research objective was to examine how instructions in an ultrasound-based and AI-supported diagnostic process can support novices to understand informational interdependence in information processing and thus affect their automation-related UX as well as diagnostic quality

RO5: The fifth research objective was to integrate the results of related research, findings of RO1 - RO3, and conceptual models from automation research and action regulation into a conceptual model of integrated information processing of humans and AI, that can be used in the design and evaluation HCXAI systems.

Hence, the present dissertation aims to improve the understanding of automation-related user experience in diagnostic AI systems and contributes to the development of human-centered information processing.

4 Study 1: Examining Automation-Related User Experience in Digital Contact Tracing

4.1 Summary of Study 1

This study examines the relationship between automation-related user experience and perceived usefulness and adoption of digital contact tracing applications during the COVID-19 pandemic. The study focuses on variables such as perceived trustworthiness, traceability, and usefulness in the context of digital contact tracing systems and how these factors influence users' intention to adopt such applications. Through a survey of 317 users of a German digital contact tracing app, the research applies a partial least squares structural equation model to analyze these relationships. The results emphasize that the perceived diagnosticity of the app, i.e. how effectively it supports user decision making in pandemic situations, is crucial for fostering trust and, ultimately, adoption intention. The paper concludes that designing digital contact tracing apps with high diagnosticity, while balancing privacy concerns, is key to improving their adoption and effectiveness.

4.2 Relevance within the dissertation

This study is integral to the dissertation, as it demonstrates how framing contact tracing apps as a form of automation can help examine automation-related user experience and promote intention to use in the event of a public health crisis. The study advances theoretical models by linking automation-related variables (such as usefulness and trust) to established health behavior models, such as the Health Belief Model. This cross-disciplinary integration supports the thesis' broader exploration of public health automation and human-centered AI. It also lays the groundwork for the dissertation's exploration of the experience of AI-based automation.

4.3 Contribution to Study 1

I was primarily responsible for the design and execution of this study. I developed the research framework, specifically framing digital contact tracing as a form of automation. In addition, I co-designed the PLS-SEM model, created the measurement instruments, and oversaw the entire data collection process. My contributions extended to both quantitative and qualitative analyses, where I provided critical insights and conducted all analyses in addition to the PLS-SEM model. I also authored all sections where I synthesized theoretical perspectives and contextualized the findings within broader research on automation and user experience.

Original Paper

Effects of User Experience in Automated Information Processing on Perceived Usefulness of Digital Contact-Tracing Apps: Cross-Sectional Survey Study

Tim Schrills, BSc, MSc; Lilian Kojan, BSc, MSc; Marthe Gruner, BSc, MSc; André Calero Valdez, Prof Dr; Thomas Franke, Prof Dr

Institute for Multimedia and Interactive Systems, Universität zu Lübeck, Lübeck, Germany

Corresponding Author:
Tim Schrills, BSc, MSc
Institute for Multimedia and Interactive Systems
Universität zu Lübeck
Ratzeburger Allee 160
Lübeck, 23560
Germany
Phone: 49 451 3101 ext 5135
Fax: 49 451 3101 5104
Email: Tim.schrills@uni-luebeck.de

Abstract

Background: In pandemic situations, digital contact tracing (DCT) can be an effective way to assess one's risk of infection and inform others in case of infection. DCT apps can support the information gathering and analysis processes of users aiming to trace contacts. However, users' use intention and use of DCT information may depend on the perceived benefits of contact tracing. While existing research has examined acceptance in DCT, automation-related user experience factors have been overlooked.

Objective: We pursued three goals: (1) to analyze how automation-related user experience (ie, perceived trustworthiness, traceability, and usefulness) relates to user behavior toward a DCT app, (2) to contextualize these effects with health behavior factors (ie, threat appraisal and moral obligation), and (3) to collect qualitative data on user demands for improved DCT communication.

Methods: Survey data were collected from 317 users of a nationwide-distributed DCT app during the COVID-19 pandemic after it had been in app stores for >1 year using a web-based convenience sample. We assessed automation-related user experience. In addition, we assessed threat appraisal and moral obligation regarding DCT use to estimate a partial least squares structural equation model predicting use intention. To provide practical steps to improve the user experience, we surveyed users' needs for improved communication of information via the app and analyzed their responses using thematic analysis.

Results: Data validity and perceived usefulness showed a significant correlation of r=0.38 (P<.001), goal congruity and perceived usefulness correlated at r=0.47 (P<.001), and result diagnosticity and perceived usefulness had a strong correlation of r=0.56 (P<.001). In addition, a correlation of r=0.35 (P<.001) was observed between Subjective Information Processing Awareness and perceived usefulness, suggesting that automation-related changes might influence the perceived utility of DCT. Finally, a moderate positive correlation of r=0.47 (P<.001) was found between perceived usefulness and use intention, highlighting the connection between user experience variables and use intention. Partial least squares structural equation modeling explained 55.6% of the variance in use intention, with the strongest direct predictor being perceived trustworthiness (β=.54; P<.001) followed by moral obligation (β=.22; P<.001). Based on the qualitative data, users mainly demanded more detailed information about contacts (eg, place and time of contact). They also wanted to share information (eg, whether they wore a mask) to improve the accuracy and diagnosticity of risk calculation.

Conclusions: The perceived result diagnosticity of DCT apps is crucial for perceived trustworthiness and use intention. By designing for high diagnosticity for the user, DCT apps could improve their support in the action regulation of users, resulting in higher perceived trustworthiness and use in pandemic situations. In general, automation-related user experience has greater importance for use intention than general health behavior or experience.

(JMIR Hum Factors 2024;11:e53940) doi: 10.2196/53940

KEYWORDS

COVID-19; contact tracing; user experience; trust; health information processing

Introduction

Background

During pandemic situations, efficiently acquiring, storing, and evaluating information on physical contacts can be crucial for both individuals and public health agencies aiming to curb infection dynamics [1]. Manual tracing of such contacts is practically impossible, leading to a growing development and research of digital tools supporting such efforts, commonly referred to as digital contact tracing (DCT) apps [2]. By allowing for automation, DCT tools effectively allow for contact tracing. They aim to allow individual users to assess their own risk status with minimal effort and offer support in daily action regulation, such as in decision situations, regarding isolation or notification of previous contacts [3]. If used correctly, DCT can aid in breaking chains of infection and thereby support curbing pandemic spread. For example, in Germany, a DCT called *Corona-Warn-App* (CWA) [4] was developed on behalf of the Federal Ministry of Health, and it was downloaded >40 million times [5].

However, the extent to which individuals use DCT can vary vastly [6]. Previous research has shown that it is crucial whether users perceive a DCT app as beneficial to guide them in pandemic contexts [7]. This core factor is in line with existing models of health behavior (eg, the influential Health Belief Model [HBM] [8]). Within the HBM, perceived benefit is outlined as a central determinant for the implementation of health behavior [7]. When investigating health-related technology, the HBM is frequently connected with models of technology acceptance [9]. As part of these models, the perceived usefulness or performance of technology is similarly postulated as a central variable for use intention. In this paper, we refer to the term *usefulness* as it is better suited than *benefits* to describe the effects of a specific technology. Thereby, we refer to usefulness as "the degree to which a person believes that using a particular system would enhance their [...] performance" [10].

Examining psychological processes revolving around the perception of DCT usefulness is a crucial research topic to understand the adoption and efficient implementation of DCT. Extensive research has shown the importance of the perceived usefulness of DCT for different applications and in different countries [11-15]. All in all, extending existing theoretical approaches such as the HBM by focusing on user experience variables in DCT allows for clear guidelines on improving DCT design and uptake.

The usefulness that a user can experience from DCT results from the automation it provides. DCT takes over tasks that would otherwise need to be done manually (eg, recording contacts, estimating distance and exposure to contacts, and calculating risk based on the vaccination status of contacts). Therefore, it can be defined as an automated system. In general, automation can be defined as a system's ability to "offload, assist, or replace human performance at corresponding stages of human information processing" [16]. The human action that DCT seeks to automate is the continuous recording and analysis of contact data to monitor an individual's risk of infection. While there is a large body of research on automation, its adverse biases, and its impact on human performance [17-19], less research focuses on the psychological processes involved when users evaluate the usefulness of automated contact tracing.

Parasuraman et al [20] define 4 evaluation criteria on how automation can affect human performance: situation awareness [21], trust (cf complacency and trust [22]), skill degradation [23], and workload [24]. When users want to make situation-adequate decisions, they benefit from improved situation awareness. Situation awareness, in turn, can be improved by DCT. As long as the information or recommendations provided by DCT apps are perceived as trustworthy, users may use them to determine the right course of action. Accordingly, a DCT's ability to support situation awareness as well as trust formation (refer to the study by Hoff and Bashir [25]) may lead to perceived usefulness. On the other hand, in the context of DCT apps, one cannot assume that users are potentially losing a previously existing skill through automation; DCT app users are not able to stop sick individuals or themselves. Along the same line, DCT app users profit from automation as it reduces manual work in contact tracing. Therefore, we propose to examine users' experience of situation awareness and trustworthiness when using DCT apps.

While research has demonstrated that usefulness strongly impacts use intention [26], factors unrelated to the specific DCT app might affect whether people intend to use the system. The HBM positions threat appraisal as another factor directly influencing use intention [7]. While using a DCT app changes neither the susceptibility nor the severity (in comparison, refer to the study by Costa [27]) related to an infection, it is still plausible that users with higher threat appraisal are more interested in their own risk status and, therefore, more likely to use a DCT app (eg, to be able to detect and react to an infection as early as possible). Therefore, threat appraisal may influence use intention independent of the specified design of DCT apps. In addition, recent research has also shown that the theoretical framework of the HBM does profit from incorporating prosocial aspects of decisions [28,29] (ie, using a DCT app may provide a sense of moral obligation to others). Even though individuals with immunity may perceive a lower personal threat, they may feel a personal obligation to track and inform contacts. Overall, to fully investigate the influence of the perceived usefulness of a DCT system on the use intention, a comparison with system-nonspecific factors (ie, threat appraisal) and personal moral obligation should be made. To the best of our knowledge, no previous study has focused on examining the perception of automation-related usefulness while addressing threat appraisal and moral obligation as system-independent factors influencing use intention.

Research Objective

The objective of this research was to examine how automation-related user experience affects the perceived usefulness of contact tracing as well as use intention of DCT apps and how user experience could be improved. To do so, our approach consisted of multiple methods. The first was quantitatively assessing and analyzing the impact of automation-related user experience (ie, experienced system traceability and perceived trustworthiness) as well as system knowledge on the intention of using a DCT app. The second was contextualizing the effects of automation-related user experience measures with factors related to health protection behavior (ie, threat appraisal and moral obligation). The third was a qualitative analysis of user demands for improved information communication between users and the DCT app. Therefore, the key contribution of this research is a better understanding of how system characteristics lead to perceived usefulness of DCT and how optimal DCT apps can increase use intention through automation-related user experience. Thus, this research supports the human-centered design of DCT apps.

To address these research objectives, 317 users of the CWA DCT system were surveyed about their experience with the app through a web-based questionnaire. A partial least squares structural equation model (PLS-SEM) was used to quantitatively describe the relationships among psychological factors regarding DCT use. This approach was supplemented by a thematic analysis of qualitative user requests on desired communication of information between users and the system.

Related Research

Use Intention of DCT

DCT describes software applications that support documenting information of physical contact or proximity between people (cf [30]). This includes both the (partially) automated acquisition of contact information and the analysis of this information (eg, to determine an individual's risk of infection [31]). In pandemic situations, users might have the goal to avoid contributing to the further spread of the pandemic disease and, thus, face a control task. This means that users need to constantly self-regulate their actions in relation to their environment (eg, how many people around them are infected). While users strive to achieve this goal, they are constantly facing a changing environment (ie, exposure to infected persons). To maintain control, they need to constantly acquire and analyze information and decide, for example, whether they want to isolate themselves. Such actions taken by users have a profound impact on the trajectory of their individual situation—they potentially curtail further contacts and, thereby, change the future information acquisition process. In this process, DCT constitutes a crucial tool for behavioral control as the information provided functions both as feedback for previous behavior and as an indicator for future behavior.

Although DCT applications, especially on mobile devices, first generated high interest during the COVID-19 pandemic [32], they had already been used previously (refer to, eg, the study by Sacks et al [33]). Due to their wide applicability and potential role in public health systems during the COVID-19 pandemic, research on user behavior toward DCT has increased. Here, diverging acceptance models (such as the Unified Theory of Acceptance and Use of Technology and the technology acceptance model) have been evaluated to understand DCT use intention (eg, the study by Velicia-Martin et al [34]).

As indicated at the outset, previous research on DCT app use has leveraged not only acceptance models but also more general models of health behavior such as the HBM or the Theory of Planned Behavior [35]. Such models have been successfully used in research on the uptake and maintenance of other pandemic protective behaviors. In that context, there is consistent evidence of the importance of factors related to the behavior itself, such as perceived usefulness; factors related to perceived risk, such as threat appraisal; and social and normative factors [11,36]. However, in the DCT context, results are mixed. While there is broad support for the importance of factors such as use intention [35] and perceived usefulness [7], evidence of the role of the other factors is less consistent. For example, Tomczyk et al [35] found evidence of the role of both subjective norms and threat appraisal. In contrast, Walrave et al [7] did not include normative factors in their study and found no significant relationship between threat appraisal and DCT adoption. In a different approach to conceptualizing norms, Zabel et al [37] found a strong association between DCT adoption and moral intensity, a construct that derives the perceived obligation for DCT adoption from a range of beliefs, including beliefs about both usefulness and risk. This not only mirrors findings on the association between moral obligation and other pandemic protective behaviors, but as the community benefit of DCT might outweigh the individual benefit, it also appears to be a promising avenue for exploring the relationship between norms and DCT use. Accordingly, it remains an important task of DCT research to understand the relative influence and interplay of both factors such as perceived usefulness, and factors such as threat appraisal or moral obligation on use intention.

One reason for the ambiguity of existing results can be the variability of operationalizations—trust, for example, is highlighted in multiple studies as decisive for DCT use intention [7,35,37]. However, the conceptualization of trust can be challenging and context-dependent [38]. In DCT, for example, trust could influence one's belief regarding how effectively DCT can support the individual in avoiding an infection. On the other hand, trust can be related to the data security of private information (refer to, eg, the study by Altmann et al [39]). Therefore, a context-sensitive and theory-based conceptualization of trust is necessary to operationalize it adequately.

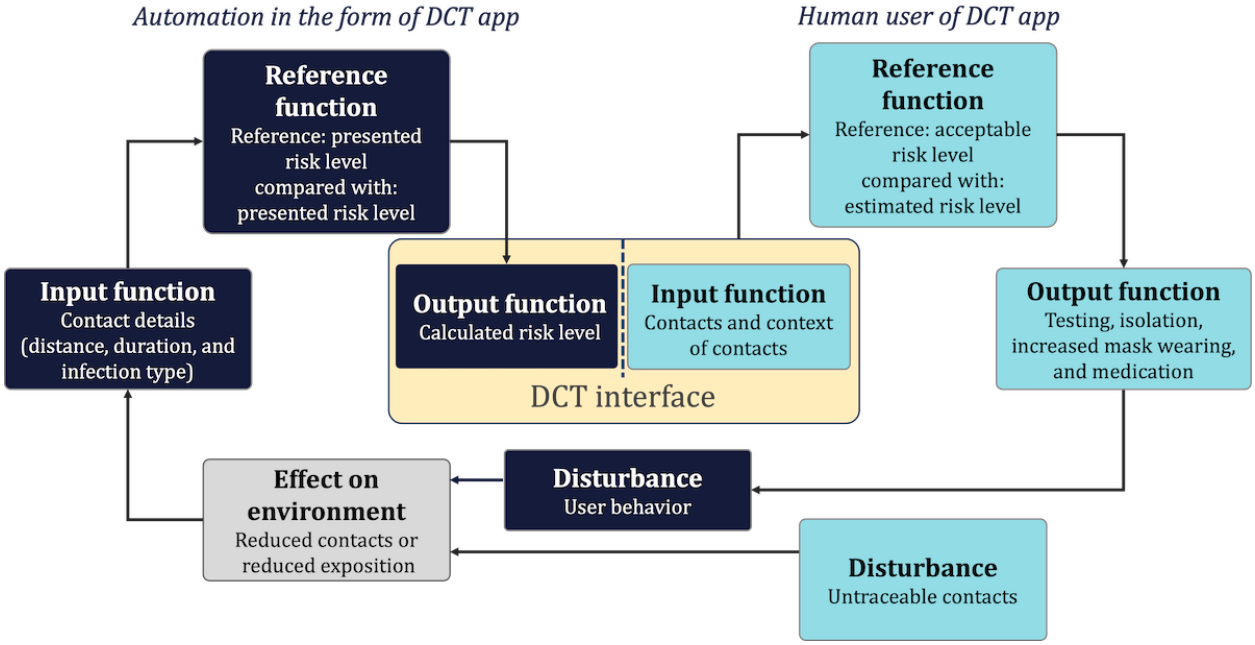
Breaking Down Automation-Related User Experience in DCT

In a pandemic context, the goal of users can be characterized as behavior that avoids both becoming infected and spreading infection to others. Still, they may desire to meet other people or use public transport and, therefore, are continuously adapting their behavior based on how they perceive the risk situation (ie, for simplification, a perceived risk level; refer to the study by Wilde [40]). This risk level refers to the probability of being

infected by, for example, a virus. Acquisition of information on the current risk level is supported by DCT and becomes critical information for comparison, prompting actions to reduce risk.

Contact tracing involves data gathering but also decision-making processes that influence individual and collective health outcomes. It integrates continuous information processing and, therefore, can be viewed through the theoretical lens of control-theoretical conceptions of human-machine systems. The control loop model of action regulation in contact tracing can be extended to accommodate for DCT as automation (ie, a system) that takes over tasks in the acquisition, analysis, and decision selection of contact information [20]. However, maintaining an acceptable risk level [40] is not a singular, finite process but a continuous one. Accordingly, we propose to model information acquisition, analysis, and decision selection as parts of an action regulation consisting of an input function, a reference function, and an output function. As depicted in Figure 1, both human and machine information processing can be modeled within a conceptual control loop to reflect continuous information processing. The conceptual control loop model (Figure 1) illustrates the integration of human and automation activities into a joint action regulation.

Figure 1. Conceptual control loop model of joint human-machine action regulation in digital contact tracing (DCT). The assessment of the machine processing steps (input, reference, and output) is central to the perceived trustworthiness (perceived data validity, perceived goal congruity, and perceived result diagnosticity) of the system.



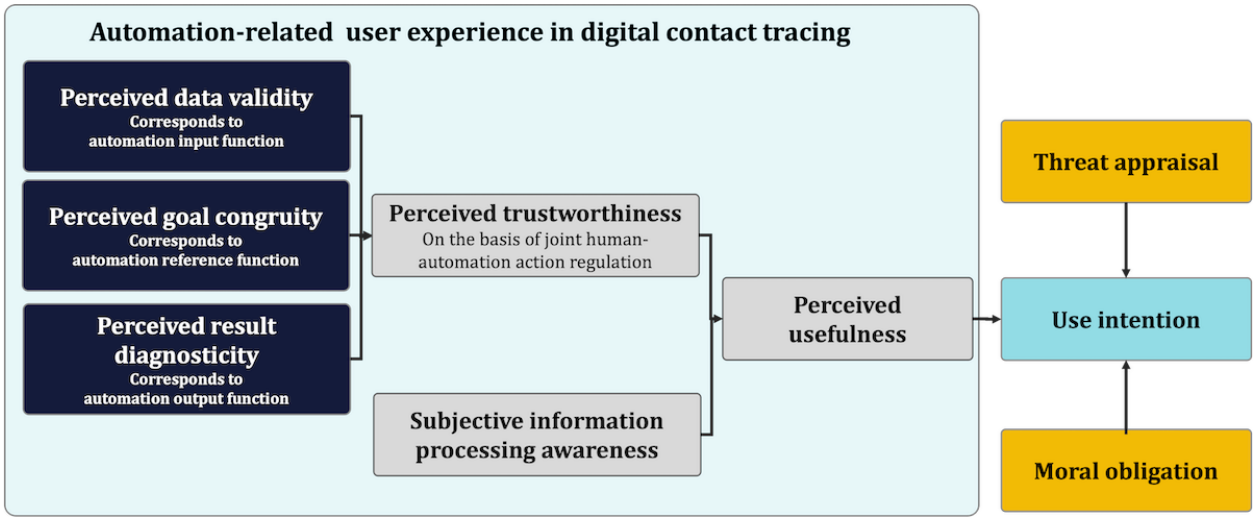
Based on the model presented in Figure 1, we assumed that users' interaction with DCT apps is based on their evaluation of automated input, reference, and output functions. They assess the correctness of the data that the DCT system uses (*input function*), the data's congruence with the users' goals (*reference function*), and the utility of the data's communicated results (*output function*). Any lack of transparency in their joint action regulation can diminish perceived trustworthiness as well as hamper situation awareness. For instance, if the system fails to capture necessary data accurately or align with personal goals such as identifying the source of infection versus alerting those potentially infected, perceived trustworthiness may decline. Accordingly, parallel to similar phenomena in other automation contexts that do not reveal which information is used as part of the input function, an out-of-the-loop unfamiliarity might cause decreasing situation awareness [20]. Furthermore, the user experience may suffer if the system's output, such as an imprecise infection risk description, is insufficient for users to decide the next course of action, therefore impeding the perceived usefulness.

In addition, users' perception of the system is dependent on their expectations of information processing (cf [41]; ie, how the DCT system processes contact-related data). For example, whether a DCT app processes others' vaccination status will only matter to users who are interested in that information, and disclosing that the app processes vaccination information will only impact the system perception of those users. As such, to understand the formation of perceived usefulness, users' subjective situation awareness is more important than their factual situation awareness. However, as introduced by Schrills and Franke [42], subjective evaluation of a user's ability to "perceive, understand and predict a system's information processing," described as subjective information processing awareness, can serve as a construct to assess users' perception of an automation's effect on situation awareness. However, users' perception of their information processing awareness might not be reflected in the accuracy of their knowledge about the system's information processing.

The previous concepts of perceived data validity, goal congruity, result diagnosticity, trustworthiness, subjective information processing awareness, and perceived usefulness can be subsumed as automation-related user experience. Automation-related user experience, following the 9241 standard from the International Organization for Standardization, can be

defined as *the perception and response of a person resulting from using or anticipating the use of automated systems*. On the basis of our proposed conception of automation-related user experience, we conceptualized a model of factors of use intention in DCT centered on perceived usefulness of automation as depicted in Figure 2. In addition, threat appraisal and moral obligation as factors independent of DCT use are integrated as measures to evaluate the influence of automation-related user experience on use intention comparatively. Threat appraisal and moral obligation are not connected with properties of the DCT app; that is, they influence whether a user wants to demonstrate behavior to trace contacts but not how useful a specific app is perceived to be.

Figure 2. Research model on automation-related user experience and the effect on use intention of digital contact-tracing apps.



This Study

On the basis of the presented research model, the objective of this study was to investigate how automation-related user experience affects the perceived usefulness of contact tracing as well as the use intention of DCT and how user experience could be improved. We aimed to contribute to research on DCT adoption and use by examining possible pathways to enhance use intention via user experience. On the basis of the proposed research model, we analyzed the following hypotheses: (1) perceived trustworthiness correlates positively with perceived usefulness (hypothesis 1), (2) subjective information processing awareness correlates positively with perceived usefulness (hypothesis 2), and (3) perceived usefulness correlates positively with use intention (hypothesis 3).

In addition, we examined the relationship among all the aforementioned variables in a structural equation modeling (SEM), where we tested automation-related variables as well as variables not related to the specific DCT system: (1) threat appraisal is positively related to use intention (hypothesis 4) and (2) moral obligation is positively related to use intention (hypothesis 5).

Accordingly, the research model depicted in Figure 2 serves as a basis for an SEM analysis that integrated both automation-related user experience and automation-independent variables (threat appraisal and moral obligation).

We supplemented our quantitative findings with qualitative data on the requirements for improved information processing, providing a deeper insight into users' interactions with the app. This mixed methods approach allowed us to uncover underlying patterns and themes that cannot be identified through quantitative data alone, providing a more comprehensive understanding of the user experience.

Methods

Participants

Participants were recruited via social networks (Twitter [subsequently rebranded X] and Facebook), where an image and a link to the study were shared showing a picture of the CWA and asking for participation (ie, our sample was self-selected). The recruitment strategy specifically targeted individuals who had experience using the CWA. Eligibility for the study required participants to be aged ≥18 years and have at least fluent German skills. The study was conducted on the web, with data collection taking place via a web-based questionnaire between June 1, 2022, and July 31, 2022, using LimeSurvey (LimeSurvey GmbH) [43]. We decided not to inquire further about demographic variables to maintain high levels of privacy due to the context of the study (tracking apps).

A total of 317 participants were included in the study (refer to the Data Exclusion section for further details). As user diversity can have a significant impact on the individual user experience and the perceived trustworthiness, we assessed the affinity for technology interaction (ATI) [44]. ATI describes the individual tendency to actively engage in intensive technology interaction. The ATI was measured using a scale validated in various large samples. Our sample ranged from 1 to 6, with an average value of 4.19 (SD 1.26) which was somewhat higher than the value of 3.5 that Franke et al [44] assumed for the general population based on quota sampling. This corresponds with the self-selection of the sample; we can assume that users who installed the CWA may have, in general, a higher level of ATI than the general population.

Ethical Considerations

This study was registered (under 2022-413) at the Ethics Committee of the University of Lübeck. Before participating in the study, individuals received detailed information about the study and provided written consent to partake. For anonymity, no additional demographic data of the users were queried. No financial remuneration was provided for participation.

Scales and Procedure

Overview

To capture the psychological concepts described previously, multiple scales were developed and presented to participants after they provided informed consent. Except for those for experienced system traceability [42], all items were generated by the researchers based on theoretical considerations and discussed within a team of 3 experts in human-machine interaction.

All items used a 6-point Likert response scale (*completely disagree*=1, *largely disagree*=2, *slightly disagree*=3, *slightly agree*=4, *largely agree*=5, and *completely agree*=6), with the only exception being the semantic differential used for perceived usefulness. For all variables except knowledge, a mean score of all items of the scale was calculated and used for further analysis. All the original items were in German and are presented in this manuscript in English.

Use Intention

Use intention was captured using a 3-item scale focusing on participants' intention and future commitment to use the CWA during the pandemic (Multimedia Appendix 1).

Threat Appraisal

A 4-item scale was used aiming to comprehend the participants' perceived risk and concerns related to a possible infection (Multimedia Appendix 1).

Experienced System Traceability

Experienced system traceability was assessed using the 6-item Subjective Information Processing Awareness scale [42] measuring the perceived transparency, understandability, and predictability of information collection and processing by the system (Multimedia Appendix 1).

Moral Obligation

Moral obligation was evaluated using a 3-item scale capturing the participants' sense of responsibility and ethical obligation toward using the CWA (Multimedia Appendix 1).

Perceived Trustworthiness

Perceived trustworthiness was measured across 3 subscales, each addressing the trustworthiness of input, reference, and output in the cybernetic control loop (Multimedia Appendix 1).

Perceived Usefulness

Perceived usefulness was assessed using a semantic differential scale with labels indicative of the perceived efficiency, precision, safety, complexity, and reliability of the system when cooperating with it (for instructions and labels, refer to Multimedia Appendix 1).

Statistical Analysis

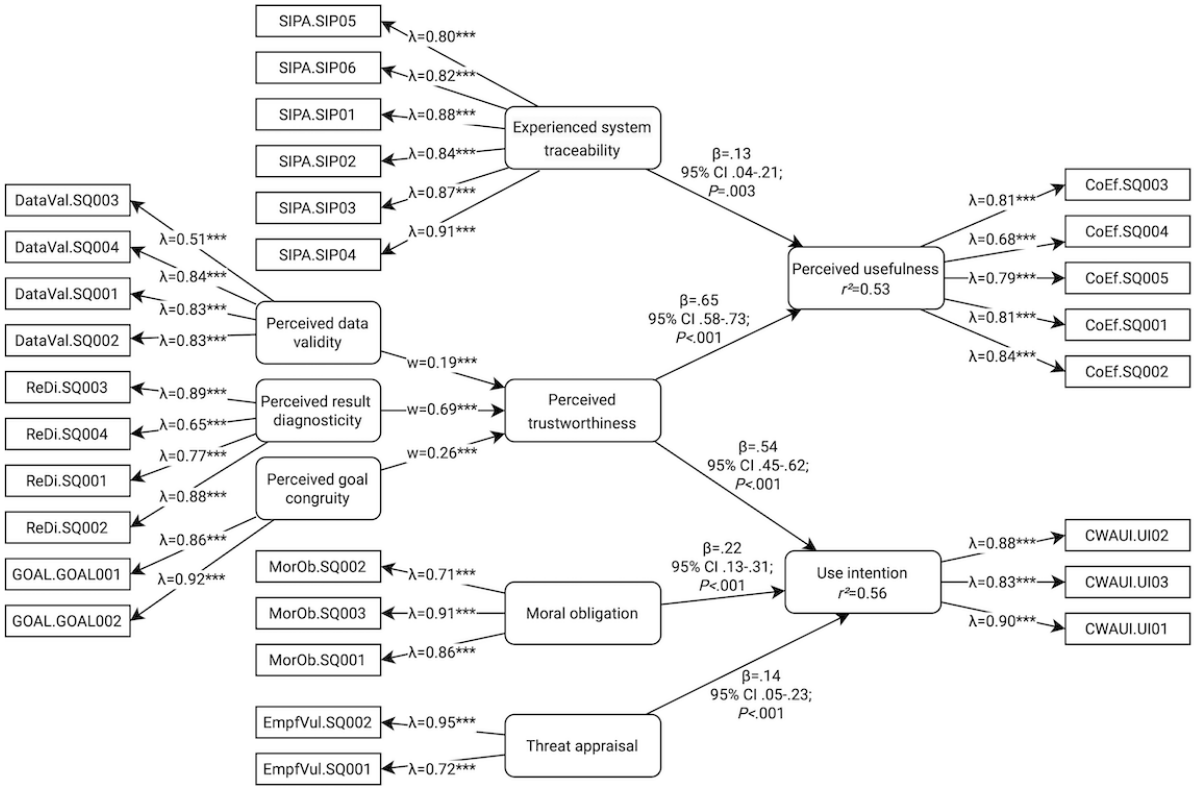
Overview

The data collected in this study were analyzed using R (version 4.31; R Foundation for Statistical Computing) [45]. Initially, the normal distribution of the data was tested to ensure that assumptions of normality were met. Given that the data did not follow a normal distribution, nonparametric tests such as the Welch 2-tailed *t* test were applied to determine statistical significance. In addition, considering the multiple comparisons performed in the calculation of correlations, a Bonferroni correction was used to control for the risk of type I error. Corrected *P* values are reported. The analysis was based on the preregistration, which can be found under <omitted for blinded review>.

PLS-SEM is a statistical modeling method combining aspects of regression and factor analysis. It allows for the simultaneous estimation of the relationship between indicators (ie, manifest variables) and constructs (ie, the latent variables formed from the manifest variables) and the relationship between the constructs themselves. These parts of the models are called the measurement model and structural model [46]. PLS-SEM is robust to nonparametric data, can work with small samples, and is especially suited for exploratory research [47], making it a great fit for this study. We followed the extensive iterative process of model assessment described in the work by Hair [46]. Our iterative approach is documented in Multimedia Appendix 2.

The hypothesized PLS-SEM contains all paths depicted in Figure 3. In addition, we tested whether the paths from perceived trustworthiness, system knowledge, and experienced system traceability to use intention were all mediated by perceived usefulness or whether there were also direct effects.

Figure 3. Partial least squares structural equation model after multiple iterations for the proposed research model. Rounded corners indicate constructs based on our research model; rectangular shapes denote indicators that were measured directly in the survey.



All constructs except perceived trustworthiness were specified as mode-A constructs. The respective indicators are described in the Scales and Procedure section. Perceived trustworthiness was specified as a mode-B higher-order construct consisting of perceived data validity, perceived result diagnosticity, and perceived goal congruity. We report explained variance using R^2, path coefficients using β with *P* values and 95% CIs, and effect sizes using the Cohen f^2.

Power

For the PLS-SEM, a retrospective power analysis using the inverse square root method revealed that, given our sample size (N=317), the smallest path coefficient, and a 5% significance level, we achieved a statistical power of 72% [48].

Data Exclusion

Before the statistical analysis, the data set with 370 responses was carefully reviewed for any inconsistencies, missing data, and outliers. Cases with incomplete or implausible responses (53/370, 14.3% in total) were identified and excluded from the analysis to maintain the integrity of the data set.

Qualitative Data Analysis

To obtain a deeper insight into users' demand for information provision and preservation in the interaction with the CWA, qualitative data were collected via open-ended questions (ie, *what information would you like to get from the system?* [Automation to human; question 1] and *What information would you like to feed to the system?* [Human to automation; question 2]).

As a widely used tool, thematic analysis aims to support the systematic identification, analysis, and reporting of patterns (ie, themes) in qualitative reporting data. Both inductive and deductive approaches were applied using theoretical assumptions as the basis for creating the themes, which were then adapted based on the data collected [49]. The data were coded using MAXQDA (version 20; VERBI GmbH [50]). For a structured and reliable analysis approach, a coding scheme with clear definitions of codes and example coding was developed in multiple iterations (Multimedia Appendix 3). For the evaluation, two perspectives of information needs between humans and automation should be covered: (1) human to automation and (2) automation to human. In total, 2 coders coded the data based on the developed scheme. An intercoder reliability of κ=0.90 (for automation-to-human information demands) and κ=0.87 (for human-to-automation information demands) was achieved. Hence, the level of agreement was strong in both cases [51].

Coded themes for information needs in both automation to human and human to automation included contact or risk information, pandemic-related information, app-related information, and assumptions for perceived information processing. Subcodes were created to enhance coding accuracy (Multimedia Appendix 3) but were not analyzed in detail as the focus remained on the top-level codes. Codes that could not be assigned to one of the themes were assigned to the category *others*. As several participants commented, for example, on the suspected reasons for the limitation of information processing, another category was added (ie, assumed reasons for perceived information processing) to avoid losing these data. Both the categories *others* and *assumed reasons for perceived information processing* were not evaluated for this study.

Missing answers to the questions asked and specific statements that there was no demand for information were assigned the code *none*. This code was assigned only once per person and statement. Thus, in the end, it was possible to clearly distinguish how many of the 317 respondents indicated information needs and how many did not. Ultimately, automation-to-human information demand statements from 45.4% (144/317) of the participants and human-to-automation information demand statements from 27.1% (86/317) of the participants were analyzed (Table 1).

Table 1. Number of respondents that indicated information demands versus no information demands.

Variable	Response distribution, n (%)	
	Respondents (n=317)	Responses (n=377)
Demands		
Information demand (A2H[a])	144 (45.4)	257 (68.2)
Information demand (H2A[b])	86 (27.1)	120 (31.8)
No demands		
Information demand (A2H)	173 (54.5)	120 (31.8)
Information demand (H2A)	231 (72.9)	257 (68.2)

[a]A2H: automation to human.

[b]H2A: human to automation.

Results

Overview

For hypothesis 1, the analysis revealed moderate positive correlations for all factors of perceived trustworthiness. The correlation between data validity and perceived usefulness was significant, with a coefficient of r=0.38 and P<.001. The correlation between goal congruity and perceived usefulness showed a coefficient of r=0.47 and P<.001, indicating a moderate positive linear relationship. Result diagnosticity and perceived usefulness exhibited a strong positive correlation, with a coefficient of r=0.56 and P<.001. In general, all measures of perceived trustworthiness and perceived usefulness exhibited a positive relationship, supporting hypothesis 1.

For hypothesis 2, a correlation coefficient of r=0.35 (P<.001) was observed, suggesting a moderate positive linear relationship between subjective information processing awareness and perceived usefulness; a positive relationship between SIPA and perceived usefulness (hypothesis 2) was supported by the data. This indicates that automation-related phenomena such as changes in situation awareness might influence the perceived usefulness of DCT.

For hypothesis 3, the correlation coefficient between perceived usefulness and use intention was r=0.47 and P<.001, indicating a moderate positive correlation. Hence, our results support the hypothesis (hypothesis 3) that perceived usefulness is positively related to use intention (hypothesis 3). In combination with our previous results, this indicates strong relationships between user experience variables and use intention.

In summary, all variables showed statistically significant correlations with perceived usefulness. These correlations ranged from moderate to strong positive relationships. These results strengthen our assumption that perceived usefulness of DCT is strongly related to automation-related user experience.

SEM Approach

The final PLS-SEM is depicted in Figure 3. The explained variance for use intention was R^2=0.56. It was directly predicted by perceived trustworthiness (β=.54, 95% CI .45-.62; P<.001; f^2=0.44), moral obligation (β=.22, 95% CI .13-.31; P<.001; f^2=0.07), and threat appraisal (β=.14, 95% CI .05-.23; P<.001; f^2=0.04). Thus, there was a large effect for perceived trustworthiness and a small effect for the other constructs. Still, hypotheses 4 and 5 were supported.

Within the perceived trustworthiness higher-order construct, the highest weight was assigned to perceived result diagnosticity (w=0.69; P<.001), implying that this subconstruct contributes most to perceived trustworthiness, followed by perceived goal congruity (w=0.26; P<.001) and perceived data validity (w=0.19; P<.001).

We did not find evidence for a mediating effect of perceived usefulness on the paths from perceived trustworthiness, system knowledge, and experienced system traceability to use intention. However, we did find direct effects of perceived trustworthiness (β=.65, 95% CI .58-.73; P<.001; f^2=0.65) and experienced system traceability (β=.13, .04-.21; P=.003; f^2=0.02) on perceived usefulness (R^2=0.53).

Qualitative Analysis

Overview

Two directions of information flow were analyzed to assess the information demands of CWA users: (1) human to automation—information that users want to provide to the system and (2) automation to human—information that users want to receive from the system. In total, 3 overarching themes were explored and analyzed in more detail (Textbox 1).

Textbox 1. Analyzed themes and description of each theme. The detailed coding scheme can be found in Multimedia Appendix 3.

Contact- or risk-related information

- Time-related information: information regarding the period of the contact, the duration of the contact, the time passed since the contact, and the period during which contact tracing was possible
- Location-related information: information related to the place of contact, direct or indirect contact, and indoor or outdoor contact
- Exposition-related information: information about the masking status in the contact situation and the distance between the persons in contact
- Action-related information: information on possible and suggested courses of action after contact
- Information related to the warning person: information concerning the time when the warning person tested positive, the time when the warning person became infected, the warning person's first symptoms, the warning person's vaccination status, and the infected person's virus variant

Pandemic-related information

- Statistics: information related to statistical content on the pandemic in terms of the number of defects or infections

App-related information

- Number of users: information about the number of users of the Corona-Warn-App
- General calculation-related information: information on reasons for changing risk calculation and the system parameters used for calculations
- Certainty about the result: information related to the certainty of the results calculated by the system
- Integration of tests (self- and externally administered): information about the possibility to enter or delete test results on the app
- Linking with private data: information on the possibility of linking app functions with private data

Descriptive Data

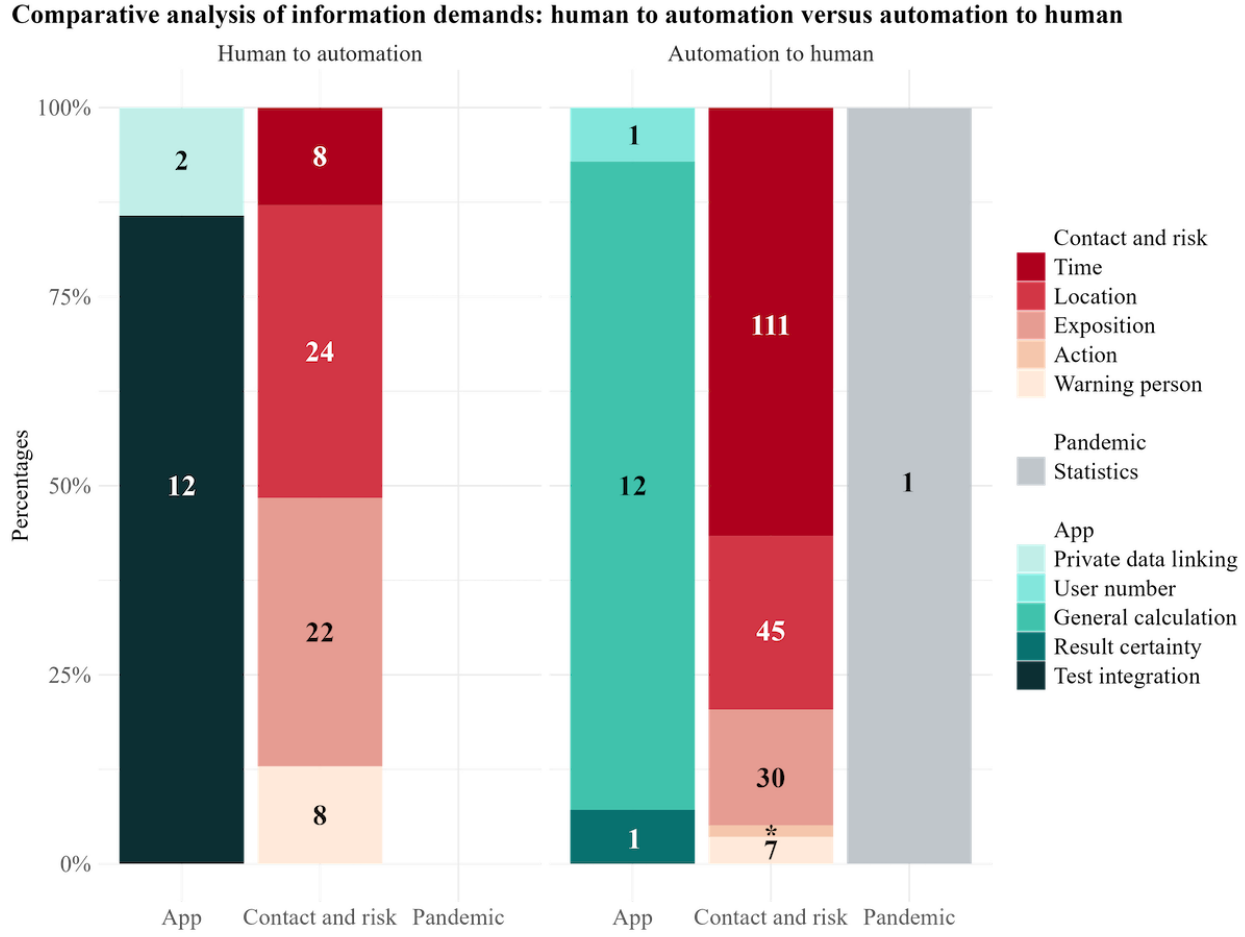
Overview

The overall number of statements amounted to 211 in automation to human and 76 in human to automation. Within these 2 categories, the themes were distributed unevenly. Information regarding contact and risk accounted for most statements in both categories (automation to human: 196/211, 92.9% of statements; human to automation: 62/76, 82% of statements). The remaining statements were (almost) exclusively distributed among app-related information (automation to human: 14/211, 6.6% of statements; human to automation: 14/76, 18% of statements) as barely any needs were stated for pandemic-related information (automation to human: 1/211, 0.5% of statements; human to automation: 0 statements).

Regarding the subcodes, the distribution also varied between both themes (Figure 4). For contact- and risk-related information, the information related to time, location, and exposure accounted for the largest proportion of demands within this theme in both categories. However, the distribution of statement proportions differed clearly between automation to human and human to automation. Time-related information was demanded most in automation to human (111/196, 56.6% of statements) but least in human to automation (8/62, 13% of statements). Demands for location-related information did not differ greatly between automation to human (45/196, 23% of statements) and human to automation (24/62, 39% of statements), nor did exposition-related information (automation to human: 30/196, 15.3% of statements; human to automation: 22/62, 35% of statements).

In terms of *app-related information*, the demands for information about the system's general calculation (automation to human: 12/14, 86% of statements; human to automation: 0% of statements) and the integration of tests (automation to human: 0% of statements; human to automation: 12/14, 86% of statements) differed in particular between the categories. The remaining subcodes hardly received any consideration. In both categories (automation to human and human to automation), almost no statements regarding *pandemic-related information* were made.

Figure 4. Relative demands regarding information from automation to human (left) and from human to automation (right). The numbers in the column sections indicate the number of statements under each code.



Human to Automation

In human to automation, certain claims emerged with particular frequency in the demand for contact- and risk-related and app-related information. Information demands on contact and risk mainly focused on time- and exposure-related information. For example, the interest in informing the app of one's location and whether one was in an enclosed space or outdoors was present:

> *Tell the app something about the specific location (enclosed space, fresh air).*

Exposition-related information demands mainly focused on informing the app when one wore or had worn a mask:

> *The wearing of a mouth-nose covering should be entered and thus taken into account in the risk calculation.*

Regarding the demand for the integration of app-related information, the participants predominantly highlighted the integration of self-administered or externally administered tests:

> *That I am Corona positive without having done a Polymera-Chain Reaction (PCR) test. (Perhaps with indication that the result is not PCR verified).*

Automation to Human

In the automation to human category, contact- and risk-related and app-related information were queried with similar frequency. The contact- and risk-related information in this category most often referred to time-related information with a request for the time of the risk encounter. However, the desired preciseness of the temporal data differed (exact time vs more approximate time: "When was the encounter? (At least as a time frame, e.g., between 8-12 o'clock)" vs "The specific time [...] of a risk encounter would be helpful"). The location of the risk encounter was another type of information that participants commonly solicited. Most asked for information about a rather specific location ("At which location did a contact take place?"); few seemed to be interested in the characteristics of the location ("Indoors or outdoors?").

Exposure-related information demanded from the system included the number of devices or persons present at the time of exposure ("[...] with how many devices was the contact?"), the distance to the warning person ("At what distance was the encounter?"), and the masking status. In particular, masking status included the person's own status of having worn a mask or whether the other person was wearing a mask at the time of the risk encounter ("Was I wearing a mask? Was the other person wearing a mask?").

App-related information demands mainly focused on the parameters of the calculation ("What factors led to this result?") and reasons for a status change ("How exactly the risk determination works, i.e., how distance and time to a positively tested person actually have to be, in order for me to receive a notification and for the status to be changed").

Discussion

Principal Findings

The objective of this study was to understand automation-related user experience, its connection to perceived usefulness, and the use intention of DCT. Our data showed that perceived trustworthiness is a critical factor in understanding use intention as well as the perceived usefulness of DCT apps. Interestingly, users' experience of a system as supportive in their action regulation affects their use intention more strongly than external factors such as threat appraisal or moral obligation. In addition, our qualitative analysis revealed that users mainly want to communicate with the system about information that is relevant to their decision-making. For instance, providing more precise information about masking status when in contact with other people could assist a user in making an immediate decision regarding isolation. Overall, our findings suggest a strong relationship between the diagnosticity of automated information processing and use intention.

Practical Implications

As a first major implication, the high effect of result diagnosticity on perceived trustworthiness demonstrates the importance of human-centered information processing in (partially) automated health applications. Within the interconnected human-machine information processing loops (Figure 1), the machine provides information as part of the human input function. As discussed by Miller [52], intelligent systems such as DCT should aim to improve users' ability to access and use (processed) information rather than to present and justify a particular outcome. In DCT app design, the integration of DCT information into a joint human-automation action regulation should be prioritized. Accordingly, when developing evaluative systems [52] that support the evaluation of alternatives rather than suggesting specific actions, it is important to consider what evaluative process a user needs to undertake. While previous research has already identified the need for actionable information [53], the information presented by DCT apps needs to be understood in the context of human action regulation and the influence of automated systems in human action regulation. A possible solution to support diagnosticity in DCT is so-called proactive contact tracing [54], which integrates more information sources and can potentially enrich DCT results.

Second, the results indicate a strong user need for information to be provided in sufficient detail. An interface optimized for communicating information could enable users to make their own assessment of the situation. In many DCT apps, users request the ability to retrieve information about possible contacts, such as time, location, or even the person involved [55]. Our study showed similar results (eg, a high demand for detailed information about the [exact] time of detected contacts). Again, the demand for more detailed information relates to the diagnosticity of the information provided by the system. If users are only given information about their current risk of infection, they cannot evaluate the validity of this information, potentially leading them to ignore it. They would require additional context-related information about potential contacts, such as whether the individuals were wearing masks or were located in an enclosed room, to make informed decisions about their behavior. Our results demonstrate that use intention is strongly connected to the perceived diagnosticity of the DCT app. On the basis of our qualitative findings, we can assume that the diagnosticity of DCT users depends on the level of detail they receive about possible contacts. Accordingly, the provision of details that support users' information processing is even more important for their use intention than threat appraisal or moral obligation. In accordance with psychological research on motivation [56], supporting users' intrinsic motivation for diagnostic information could lead to better adherence regarding DCT apps than, for instance, exposing them to extrinsic motivators that increase threat appraisal (eg, describing the consequences of infection [57]).

Third, in contradiction to users' demand for detailed information on contacts, a major concern in DCT is privacy [55]. While it is often argued that too much detail conflicts with privacy, it is important to find ways to improve the diagnosticity of information as this determines the use intention. Possible solutions include differential privacy, which allows for sufficient detail for increased diagnosticity while keeping personal data confidential. In addition, many users requested features that do not compromise the privacy of others, such as the ability to inform the system about masking status. Thus, allowing users to refine the input received by the DCT app may increase the perceived diagnosticity of the results. The integration of masking status can be seen as a measure to improve the accuracy of the apps in determining risk levels, ultimately increasing the use intention.

Overall, our results suggest that focusing on the diagnosticity of the information presented in DCT apps could result in improvement in users' health behavior. During the COVID-19 pandemic, users reported that they were unsure about the correct or best action to take to contain the pandemic or could not correctly assess the risk of certain situations [58]. However, this certainty is particularly important when it comes to health decisions. With sufficient diagnostic accuracy, DCT apps may be able to better reduce this uncertainty and, thus, become a crucial component in the management of pandemics in the long term, also positively affecting users' willingness to provide data on a social level. It is also crucial that DCT apps do not follow the *recommend and defend* principle [52], which could lead to a long-term reduction in motivation, but instead provide information that supports individual decisions. If compliance with effective pandemic control measures can be increased as a result, it will be possible to respond more effectively to future pandemics.

Theoretical and Methodological Implications

In our data, the perceived trustworthiness of a DCT app had a greater influence on use intention than threat appraisal or moral

obligation. Furthermore, while previous studies [26] have relied on perceived usefulness, our findings in the PLS-SEM do not suggest that it mediates the relationship between perceived trustworthiness and use intention. However, usefulness can be seen as an ambiguous concept without a specific connection to the design of DCT apps. In this way, focusing on perceived usefulness could hinder approaches to improve DCT by adopting DCT app design and functionality. In contrast, a lack of perceived result diagnosticity indicates to developers that the information provided by a DCT app needs to be adapted to have an impact on joint action regulation. Our research suggests that designers of automated systems should specify the potential actions that users can take and identify decision points at which users may require diagnostic information, such as whether to proceed with a specific action. In addition, highlighting the role of diagnosticity indicates how models of technology in medical systems should be developed. Existing models (such as the technology acceptance model) do not specify to what extent a system's usefulness depends on perceived diagnosticity. Our research demonstrates that behavioral models focusing on information-based decisions are needed to address automated technology in health, for example, DCT.

However, one can argue that the difference between perceived result diagnosticity and perceived usefulness is arbitrary; in a joint human-automation action regulation, the diagnosticity of information seems to be equal to perceived usefulness. However, by directly addressing perceived result diagnosticity as a central variable of automation-related user experience, empirical research can identify paths to improve action regulation support of DCT without previously defining what is useful about a system or not. When a DCT app can deliver information that users can use to regulate their actions, users report a higher intention to use it. Therefore, applying result diagnosticity as a variable in human-automation research is a methodological contribution supporting future research in intelligent automation.

On the basis of our findings, future research on DCT needs to determine how to improve the diagnosticity of DCT apps. This paper introduced a conceptual control loop model of joint human-machine action regulation, which can support research approaches in optimizing perceived diagnosticity as a central variable for automation-related user experience. Addressing the joint action regulation in DCT and health behavior is crucial to understand how the information provided by DCT apps can be integrated into human information processing and how DCT apps influence the human output function. Information that improves the evaluation of individual contacts, such as contact location, masking status, or vaccination level, could improve perceived trustworthiness and use intention of DCT apps. By demonstrating how information processing between human users and DCT apps is integrated, our research supports a shift from viewing human users as receivers of machine results to viewing them as actors using DCT information.

All in all, our findings regarding the significance of diagnosticity have implications for the design of automated information processing in a broader context. Users did not primarily prioritize data validity or goal congruence; instead, their focus lay in determining whether they could trust the system to provide information that would assist their own decision-making process. This may be a general trend in automated information processing.

Limitations and Further Research

All participants of this study were users of the CWA. However, as Walrave et al [59] describe, many citizens in Germany did not use DCT apps, for example, because they did not want to share their data or did not think they were effective. Thus, the findings presented on the impact of perceived diagnosticity may not be applicable to citizens who did not use the app at all. These individuals may have chosen not to use the app for reasons beyond those discussed in this paper. The perceived diagnosticity of a DCT app is only relevant for use intention when potential users are interested in determining their individual risk level or making decisions based on their estimated risk level. That is, our sample may bias the results and underestimate factors relevant to nonusers. For example, nonusers might reject the app because they do not trust the provider of the system. Accordingly, the results of our study may support improving DCT for existing users but not convincing nonusers to use DCT. Further studies need to address nonusers and examine how automation-related user experience affects their decision not to use DCT.

In addition, users may have misconceptions about the factors contributing to the risk of infection and may expect the system to provide irrelevant information that does not aid in making an informed decision. Accordingly, they might report a low perceived diagnosticity while the information provided in the app offers sufficient diagnosticity. The accuracy of one's mental model [60] may influence the perception of actual diagnostic information as nondiagnostic (for a discussion of diagnosticity, refer to the study by Garcia-Marques et al [61]). To tackle false models of diagnosticity, DCT apps should support users in correcting their mental model, for example, by explaining how they can use the provided information. This could be done by simulating decision situations with and without DCT information, offering users the experience of diagnosticity.

Improving the perceived diagnosticity could be beneficial for use intention but could negatively affect perceived data privacy [55]. For example, a function that allows users to communicate when they are wearing a mask could be abused to track specific contacts, therefore revealing potential infections of other users. Data privacy is a critical concern in DCT use [59]. Therefore, current DCT apps are designed to protect the data of other users at the cost of the diagnosticity of information. This research did aim to understand the effect of user experience in automated DCT but did not include how users evaluate potential risks of data privacy violations or approaches to address them (cf [62]). Future research should identify how to balance the desired level of perceived result diagnosticity and data privacy concerns. For example, in direct communication, users who reveal information about their web-based status can see the web-based status of others, allowing them to choose which balance between diagnosticity and data protection they desire. The same function could be implemented in DCT apps to support automation-related user experience. Allowing users to choose their level of diagnosticity themselves allows them also to

control how DCT apps influence their decision-making, thus strengthening user autonomy.

Finally, this study had a cross-sectional design that did not assess how automation-related user experience and use intention regarding a DCT app may change over time. Previous research has demonstrated that automation-related user experience can change over time (eg, because users adapt to the system or they improve how they use the system). Future research on automation-related user experience in DCT apps needs to include a longitudinal study design to capture effects of behavior change and users' perception.

Conclusions

In conclusion, this research highlights the relevance of automation-related user experience in DCT and its role in enabling the effective action regulation of DCT users. Here, providing detailed and diagnostic information is crucial for users to make informed assessments of their situation and actions. The presented quantitative results echo the qualitatively assessed user demand for more detailed information about potential contacts, such as time, location, and context (eg, mask use and indoor or outdoor setting).

Interestingly, our data suggest that other factors not directly related to the app, such as moral obligation and threat appraisal, are less relevant compared to automation-related user experience, especially to the perceived diagnosticity of the information provided by DCT apps. The presented results are also more specific than those of previous studies that relied on perceived usefulness. Our research model did not suggest that perceived usefulness mediates the relationship between perceived trustworthiness and use intention. Instead, we propose that DCT designers should focus on providing diagnostic information at critical decision points.

However, privacy remains a major concern in DCT. While it is often argued that too much detail conflicts with privacy, it is crucial to find ways to improve the diagnosticity of information without compromising privacy. Solutions could include differential privacy or features that do not compromise the privacy of others, such as the ability to inform the system about masking status.

The main impact of our results on the design of DCT apps and health policy is that DCT apps need to provide sufficient diagnosticity to be perceived as useful. This means that (1) the possible actions of users need to be understood before the design of the DCT algorithm and apps and (2) the presented information needs to support them in choosing the correct action. Focusing on the diagnosticity of the information presented in DCT apps could, in turn, also influence user performance. During the COVID-19 pandemic, a significant percentage of users reported uncertainty about the best actions to take or could not correctly assess the risk of certain decisions. Therefore, improving diagnostics could contribute to better and safer decisions.

In summary, our study underscores the importance of balancing detailed and diagnostic information with privacy concerns in DCT apps. As we move forward in this digital age, it is crucial to continue exploring ways to optimize DCT while respecting user privacy.

Acknowledgments

This research was funded by the Federal Ministry of Education and Research of Germany within the framework of the Cooperative and Communicating AI project (project 01GP1908).

Authors' Contributions

TS contributed to conceptualization, methodology, investigation, resources, data curation, data analysis, writing—original draft, visualization, and funding acquisition. LK contributed to data curation, data analysis, and writing—review and editing. MG contributed to investigation, data curation, writing—review and editing, and visualization. ACV contributed to writing—review and editing and supervision. TF contributed to conceptualization, writing—review and editing, and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Scales for the cross-sectional survey study on user experience.
[PDF File (Adobe PDF File), 68 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Documentation of the iterative process of partial least squares structural equation modeling.
[PDF File (Adobe PDF File), 418 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Coding scheme.
[PDF File (Adobe PDF File), 125 KB-Multimedia Appendix 3]

References

1. Barrat A, Cattuto C, Kivelä M, Lehmann S, Saramäki J. Effect of manual and digital contact tracing on COVID-19 outbreaks: a study on empirical contact data. J R Soc Interface. May 05, 2021;18(178):20201000. [FREE Full text] [doi: 10.1098/rsif.2020.1000] [Medline: 33947224]
2. Shahroz M, Ahmad F, Younis MS, Ahmad N, Kamel Boulos MN, Vinuesa R, et al. COVID-19 digital contact tracing applications and techniques: a review post initial deployments. Transp Eng. Sep 2021;5:100072. [doi: 10.1016/j.treng.2021.100072]
3. O'Connell J, Abbas M, Beecham S, Buckley J, Chochlov M, Fitzgerald B, et al. Best practice guidance for digital contact tracing apps: a cross-disciplinary review of the literature. JMIR Mhealth Uhealth. Jun 07, 2021;9(6):e27753. [FREE Full text] [doi: 10.2196/27753] [Medline: 34003764]
4. Corona-Warn-App. GitHub. 2020. URL: https://github.com/corona-warn-app [accessed 2023-10-16]
5. Corona-Warn-App: 45 millionen downloads. Presse- und Informationsamt der Bundesregierung. 2022. URL: https://www.bundesregierung.de/breg-de/themen/coronavirus/cwa-45-mio-downloads-1994916 [accessed 2023-10-23]
6. Li T, Cobb C, Yang JJ, Baviskar S, Agarwal Y, Li B, et al. What makes people install a COVID-19 contact-tracing app? Understanding the influence of app design and individual difference on contact-tracing app adoption intention. Pervasive Mob Comput. Aug 2021;75:101439. [FREE Full text] [doi: 10.1016/j.pmcj.2021.101439] [Medline: 36569467]
7. Walrave M, Waeterloos C, Ponnet K. Adoption of a contact tracing app for containing COVID-19: a health belief model approach. JMIR Public Health Surveill. Sep 01, 2020;6(3):e20572. [FREE Full text] [doi: 10.2196/20572] [Medline: 32755882]
8. Janz NK, Becker MH. The health belief model: a decade later. Health Educ Q. Jan 01, 1984;11(1):1-47. [doi: 10.1177/109019818401100101] [Medline: 6392204]
9. Ahadzadeh AS, Pahlevan Sharif S, Ong FS, Khong KW. Integrating health belief model and technology acceptance model: an investigation of health-related internet use. J Med Internet Res. Feb 19, 2015;17(2):e45. [FREE Full text] [doi: 10.2196/jmir.3564] [Medline: 25700481]
10. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. Sep 1989;13(3):319. [doi: 10.2307/249008]
11. Huang Z, Guo H, Lim HY, Chow A. Determinants of the acceptance and adoption of a digital contact tracing tool during the COVID-19 pandemic in Singapore. Epidemiol Infect. Mar 02, 2022;150:e54. [doi: 10.1017/s0950268822000401]
12. Oyibo K, Morita PP. Factors influencing the willingness to download contact tracing apps among the American population. In: Proceedings of the Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization. 2023. Presented at: UMAP '23; June 26-29, 2023:147-156; Limassol, Cyprus. URL: https://dl.acm.org/doi/abs/10.1145/3563359.3596983 [doi: 10.1145/3563359.3596983]
13. Oyibo K, Sahu KS, Oetomo A, Morita PP. Factors influencing the adoption of contact tracing applications: systematic review and recommendations. Front Digit Health. 2022;4:862466. [FREE Full text] [doi: 10.3389/fdgth.2022.862466] [Medline: 35592459]
14. Vogt F, Haire B, Selvey L, Katelaris AL, Kaldor J. Effectiveness evaluation of digital contact tracing for COVID-19 in New South Wales, Australia. Lancet Public Health. Mar 2022;7(3):e250-e258. [FREE Full text] [doi: 10.1016/S2468-2667(22)00010-X] [Medline: 35131045]
15. Prakash AV, Das S, Pillai KR. Understanding digital contact tracing app continuance: insights from India. Health Policy Technol. Dec 2021;10(4):100573. [doi: 10.1016/j.hlpt.2021.100573]
16. Onnasch L, Wickens CD, Li H, Manzey D. Human performance consequences of stages and levels of automation: an integrated meta-analysis. Hum Factors. May 20, 2014;56(3):476-488. [doi: 10.1177/0018720813501549] [Medline: 24930170]
17. Bowden V, Griffiths N, Strickland L, Loft S. Detecting a single automation failure: the impact of expected (but not experienced) automation reliability. Hum Factors. Jun 2023;65(4):533-545. [doi: 10.1177/00187208211037188] [Medline: 34375538]
18. Clegg BA, Vieane AZ, Wickens CD, Gutzwiller RS, Sebok AL. The effects of automation-induced complacency on fault diagnosis and management performance in process control. Proc Hum Factors Ergon Soc Annu Meet. Oct 17, 2014;58(1):844-848. [doi: 10.1177/1541931214581178]
19. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. Jun 20, 2010;52(3):381-410. [doi: 10.1177/0018720810376055] [Medline: 21077562]
20. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A Syst Hum. May 2000;30(3):286-297. [doi: 10.1109/3468.844354] [Medline: 11760769]
21. Endsley MR. Automation and situation awareness. In: Parasuraman R, Mouloua M, editors. Automation and Human Performance: Theory and Applications. Boca Raton, FL. CRC Press; 1996:163-181.
22. Bahner JE, Hüper AD, Manzey D. Misuse of automated decision aids: complacency, automation bias and the impact of training experience. Int J Hum Comput Stud. Sep 2008;66(9):688-699. [doi: 10.1016/j.ijhcs.2008.06.001]

23. Volz K, Yang E, Dudley R, Lynch E, Dropps M, Dorneich MC. An evaluation of cognitive skill degradation in information automation. Proc Hum Factors Ergon Soc Annu Meet. Sep 15, 2016;60(1):191-195. [doi: 10.1177/1541931213601043]
24. Röttger S, Bali K, Manzey D. Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. Ergonomics. May 28, 2009;52(5):512-523. [doi: 10.1080/00140130802379129] [Medline: 19296323]
25. Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors. May 2015;57(3):407-434. [doi: 10.1177/0018720814547570] [Medline: 25875432]
26. Krüger N, Behne A, Beinke JH, Stibe A, Teuteberg F. Exploring user acceptance determinants of COVID-19-tracing apps to manage the pandemic. Int J Technol Hum Interact. 2022;18(1):27. [doi: 10.4018/ijthi.293197]
27. Costa MF. Health belief model for coronavirus infection risk determinants. Rev Saude Publica. May 07, 2020;54:47. [FREE Full text] [doi: 10.11606/s1518-8787.2020054002494] [Medline: 32491096]
28. Kojan L, Burbach L, Ziefle M, Calero Valdez A. Perceptions of behaviour efficacy, not perceptions of threat, are drivers of COVID-19 protective behaviour in Germany. Humanit Soc Sci Commun. Mar 24, 2022;9(1):97. [doi: 10.1057/S41599-022-01098-4]
29. Chu H, Liu S. Integrating health behavior theories to predict American's intention to receive a COVID-19 vaccine. Patient Educ Couns. Aug 2021;104(8):1878-1886. [FREE Full text] [doi: 10.1016/j.pec.2021.02.031] [Medline: 33632632]
30. Kleinman RA, Merkel C. Digital contact tracing for COVID-19. CMAJ. Jun 15, 2020;192(24):E653-E656. [FREE Full text] [doi: 10.1503/cmaj.200922] [Medline: 32461324]
31. Maccari L, Cagno V. Do we need a contact tracing app? Comput Commun. Jan 15, 2021;166:9-18. [FREE Full text] [doi: 10.1016/j.comcom.2020.11.007] [Medline: 33235399]
32. Ahmed N, Michelin RA, Xue W, Ruj S, Malaney R, Kanhere SS, et al. A survey of COVID-19 contact tracing apps. IEEE Access. 2020;8:134577-134601. [doi: 10.1109/access.2020.3010226]
33. Sacks JA, Zehe E, Redick C, Bah A, Cowger K, Camara M, et al. Introduction of mobile health tools to support Ebola surveillance and contact tracing in Guinea. Glob Health Sci Pract. Nov 12, 2015;3(4):646-659. [FREE Full text] [doi: 10.9745/GHSP-D-15-00207] [Medline: 26681710]
34. Velicia-Martin F, Cabrera-Sanchez JP, Gil-Cordero E, Palos-Sanchez PR. Researching COVID-19 tracing app acceptance: incorporating theory from the technological acceptance model. PeerJ Comput Sci. 2021;7:e316. [FREE Full text] [doi: 10.7717/peerj-cs.316] [Medline: 33816983]
35. Tomczyk S, Barth S, Schmidt S, Muehlan H. Utilizing health behavior change and technology acceptance models to predict the adoption of COVID-19 contact tracing apps: cross-sectional survey study. J Med Internet Res. May 19, 2021;23(5):e25447. [FREE Full text] [doi: 10.2196/25447] [Medline: 33882016]
36. Shushtari ZJ, Salimi Y, Ahmadi S, Rajabi-Gilan N, Shirazikhah M, Biglarian A, et al. Social determinants of adherence to COVID-19 preventive guidelines: a comprehensive review. Osong Public Health Res Perspect. Dec 2021;12(6):346-360. [FREE Full text] [doi: 10.24171/j.phrp.2021.0180] [Medline: 34965686]
37. Zabel S, Schlaile MP, Otto S. Breaking the chain with individual gain? Investigating the moral intensity of COVID-19 digital contact tracing. Comput Human Behav. Jun 2023;143:107699. [FREE Full text] [doi: 10.1016/j.chb.2023.107699] [Medline: 36818428]
38. Johnson M, Bradshaw JM. The role of interdependence in trust. In: Nam CS, Lyons JB, editors. Trust in Human-Robot Interaction. Cambridge, MA. Academic Press; 2021:379-403.
39. Altmann S, Milsom L, Zillessen H, Blasone R, Gerdon F, Bach R, et al. Acceptability of app-based contact tracing for COVID-19: cross-country survey study. JMIR Mhealth Uhealth. Aug 28, 2020;8(8):e19857. [FREE Full text] [doi: 10.2196/19857] [Medline: 32759102]
40. Wilde GJ. Risk homeostasis theory: an overview. Inj Prev. Jun 01, 1998;4(2):89-91. [FREE Full text] [doi: 10.1136/ip.4.2.89] [Medline: 9666358]
41. Roscoe RD, Wilson J, Johnson AC, Mayra CR. Presentation, expectations, and experience: sources of student perceptions of automated writing evaluation. Comput Human Behav. May 2017;70:207-221. [doi: 10.1016/j.chb.2016.12.076]
42. Schrills T, Franke T. How do users experience traceability of AI systems? Examining subjective information processing awareness in automated insulin delivery (AID) systems. ACM Trans Interact Intell Syst. 2023;13(4):1-34. [FREE Full text] [doi: 10.31234/osf.io/3v9b8]
43. LimeSurvey: an open source survey tool. LimeSurvey GmbH. URL: http://www.limesurvey.org [accessed 2024-04-29]
44. Franke T, Attig C, Wessel D. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. Int J Hum Comput Interact. Mar 30, 2018;35(6):456-467. [doi: 10.1080/10447318.2018.1456150]
45. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2021. URL: https://www.R-project.org/ [accessed 2024-04-29]
46. Hair JF, Hult GT, Ringle CM, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). 2nd edition. Thousand Oaks, CA. Sage Publications; 2017.
47. Hair JF, Risher JJ, Sarstedt M, Ringle CM. When to use and how to report the results of PLS-SEM. Eur Bus Rev. Jan 14, 2019;31(1):2-24. [doi: 10.1108/ebr-11-2018-0203]

48. Kock N, Hadaya P. Minimum sample size estimation in PLS ‐ SEM: the inverse square root and gamma ‐ exponential methods. Inf Syst J. 2018;28(1):227-261. [doi: 10.1111/isj.12131]
49. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. Jan 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
50. All-in-one tool for qualitative data analysis & mixed methods. MAXQDA - Distribution by VERBI GmbH. URL: https://www.maxqda.com [accessed 2022-08-17]
51. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282. [FREE Full text] [Medline: 23092060]
52. Miller T. Explainable AI is dead, long live explainable AI!: hypothesis-driven decision support using evaluative AI. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023. Presented at: FAccT '23; June 12-15, 2023:333-342; Chicago, IL. URL: https://dl.acm.org/doi/10.1145/3593013.3594001 [doi: 10.1145/3593013.3594001]
53. Caballero A, Leath K, Watson J. COVID-19 consumer health information needs improvement to be readable and actionable by high-risk populations. Front Commun. Aug 7, 2020;5. [FREE Full text] [doi: 10.3389/fcomm.2020.00056]
54. Gupta P, Maharaj T, Weiss M, Rahaman N, Alsdurf H, Minoyan N, et al. Proactive contact tracing. PLOS Digit Health. Mar 13, 2023;2(3):e0000199. [FREE Full text] [doi: 10.1371/journal.pdig.0000199] [Medline: 36913342]
55. Hsu J. The dilemma of contact-tracing apps: can this crucial technology be both effective and private? IEEE Spectr. Oct 2020;57(10):56-59. [doi: 10.1109/mspec.2020.9205550]
56. Martela F, Hankonen N, Ryan RM, Vansteenkiste M. Motivating voluntary compliance to behavioural restrictions: self-determination theory–based checklist of principles for COVID-19 and other emergency communications. Eur J Soc Psychol. Jan 18, 2021;32(2):305-347. [doi: 10.1080/10463283.2020.1857082]
57. Morbée S, Vermote B, Waterschoot J, Dieleman L, Soenens B, Van den Bergh O, et al. Adherence to COVID-19 measures: the critical role of autonomous motivation on a short- and long-term basis. Motiv Sci. Dec 2021;7(4):487-496. [doi: 10.1037/mot0000250]
58. Kučera D. Our words in a state of emergency: psychological-linguistic analysis of utterances on the COVID-19 situation in the Czech Republic. Psychol Stud (Mysore). Jul 13, 2021;66(3):239-258. [FREE Full text] [doi: 10.1007/s12646-021-00613-y] [Medline: 34276073]
59. Walrave M, Waeterloos C, Ponnet K. Reasons for nonuse, discontinuation of use, and acceptance of additional functionalities of a COVID-19 contact tracing app: cross-sectional survey study. JMIR Public Health Surveill. Jan 14, 2022;8(1):e22113. [FREE Full text] [doi: 10.2196/22113] [Medline: 34794117]
60. Villius Zetterholm M, Lin Y, Jokela P. Digital contact tracing applications during COVID-19: a scoping review about public acceptance. Informatics. Jul 22, 2021;8(3):48. [doi: 10.3390/informatics8030048]
61. Garcia-Marques L, Sherman SJ, Palma-Oliveira JM. Hypothesis testing and the perception of diagnosticity. J Exp Soc Psychol. May 2001;37(3):183-200. [doi: 10.1006/jesp.2000.1441]
62. Hang CN, Tsai YZ, Yu PD, Chen J, Tan CW. Privacy-enhancing digital contact tracing with machine learning for pandemic response: a comprehensive review. Big Data Cogn Comput. Jun 01, 2023;7(2):108. [doi: 10.3390/bdcc7020108]

Abbreviations

ATI: affinity for technology interaction
CWA: Corona-Warn-App
DCT: digital contact tracing
HBM: Health Belief Model
PLS-SEM: partial least squares structural equation model
SEM: structural equation modeling

Edited by A Kushniruk; submitted 25.10.23; peer-reviewed by H Kang, CN Hang; comments to author 09.01.24; revised version received 12.03.24; accepted 07.04.24; published 25.06.24

Please cite as:
Schrills T, Kojan L, Gruner M, Calero Valdez A, Franke T
Effects of User Experience in Automated Information Processing on Perceived Usefulness of Digital Contact-Tracing Apps: Cross-Sectional Survey Study
JMIR Hum Factors 2024;11:e53940
URL: https://humanfactors.jmir.org/2024/1/e53940
doi: 10.2196/53940
PMID: 38916941

5 Study 2: Subjective Information Processing Awareness in Automated Insulin Delivery

5.1 Summary of Study 2

This study investigates how users perceive the transparency and explainability of automated insulin delivery systems for type 1 diabetes management. The research introduces and examines the construct of Subjective Information Processing Awareness, which is a key measure of how well users can track and understand automated decision processes. Through an experimental study with 80 participants, the paper examines how different levels of information disclosure affect SIPA and related factors such as confidence and prediction accuracy. The results indicate that repeated exposure is necessary to manifest differences in SIPA. In addition, higher levels of information disclosure can sometimes lead to a mismatch between perceived understanding and actual performance, highlighting the challenges of designing user-friendly XAI systems.

5.2 Relevance within the dissertation

This research is essential to the dissertation because it extends the study of automation-related user experience to the medical domain, focusing on critical applications such as diabetes management. The research links human-AI interaction with concepts of

transparency and explainability, emphasizing the balance between providing enough information for effective decision making while avoiding information overload. It contributes to the broader thesis by illustrating how user experience and trust in automated systems may not be calibrated after XAI interfaces are presented.

5.3 Contribution to Study 2

I was responsible for the primary development of the research idea, conceptualized the framework of SIPA, and designed the experiment, including the levels of information disclosure and the methodology used to measure SIPA. I also conducted the data collection, performed the quantitative analysis, including conducting the contrast analysis, and contributed to the training of the model underlying the interface used in the study. In addition, I wrote the entire manuscript, including the introduction, methodology, results, and discussion sections, where I integrated theoretical perspectives on human-centered AI and XAI with practical implications for medical technologies.

How Do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems

TIM SCHRILLS and THOMAS FRANKE, Universität zu Lübeck, Germany

When interacting with artificial intelligence (AI) in the medical domain, users frequently face automated information processing, which can remain opaque to them. For example, users with diabetes may interact daily with automated insulin delivery (AID). However, effective AID therapy requires traceability of automated decisions for diverse users. Grounded in research on human-automation interaction, we study Subjective Information Processing Awareness (SIPA) as a key construct to research users' experience of explainable AI. The objective of the present research was to examine how users experience differing levels of traceability of an AI algorithm. We developed a basic AID simulation to create realistic scenarios for an experiment with $N = 80$, where we examined the effect of three levels of information disclosure on SIPA and performance. Attributes serving as the basis for insulin needs calculation were shown to users, who predicted the AID system's calculation after over 60 observations. Results showed a difference in SIPA after repeated observations, associated with a general decline of SIPA ratings over time. Supporting scale validity, SIPA was strongly correlated with trust and satisfaction with explanations. The present research indicates that the effect of different levels of information disclosure may need several repetitions before it manifests. Additionally, high levels of information disclosure may lead to a miscalibration between SIPA and performance in predicting the system's results. The results indicate that for a responsible design of XAI, system designers could utilize prediction tasks in order to calibrate experienced traceability.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **User studies**;

Additional Key Words and Phrases: Explainability, trust, human-centered AI, human AI cooperation

ACM Reference format:
Tim Schrills and Thomas Franke. 2023. How Do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 25 (December 2023), 34 pages.
https://doi.org/10.1145/3588594

1 INTRODUCTION

The availability of intelligent technology for **type 1 diabetes mellitus (DMT1)** therapy [33] has increased, reflecting the general development of personalized medicine based on **artificial intelligence (AI)**. In DMT1, self-adapting learning algorithms are used for personalized calculation

This research has been funded by the Federal Ministry of Education and Research of Germany in the framework of the project CoCoAI (Cooperative and Communicating AI, project number 01GP1908).
Authors' address: T. Schrills (corresponding author) and T. Franke, Universität zu Lübeck, Ratzeburger Allee 160, Lübeck, Germany; emails: {schrills, franke}@imis.uni-luebeck.de.

2160-6455/2023/12-ART25
https://doi.org/10.1145/3588594

of insulin needs, e.g., at different times of the day, at different stages of the female period, or depending on physical activity. The goal of these systems, also known as **automated insulin delivery (AID)** systems, is to improve therapy while reducing the workload for people with DMT1. The incidence of DMT1 has increased in recent years and was 15 per 100,000 cases in 2020 [83]. In order to improve therapy conditions and effectiveness, AID systems can provide fully or partially automated diabetes therapy, for example, through integrating advanced wearable glucose sensors and intelligent insulin pumps [116]. All in all, the core of AID technology is the automated processing of information, especially to regulate current blood glucose levels in relation to therapy goals while dealing with high temporal dynamics, latency, and complexity of human physiology.

The first empirical studies suggest that people with DMT1 can benefit significantly from AID systems [3, 18, 64]. Both long-term metrics (e.g., the **"time in range" (TIR)** referring to desired glucose level) and the frequency of acute life-critical blood glucose levels can be reduced [6]. However, the positive effect of AID systems seems to depend on, for example, the previous quality of therapy [15, 80]. That is, individuals who had problematic long-term metrics before starting AID therapy are more likely to discontinue AID-based therapy. Paradoxically, they would profit the most from AID systems. Thus, more inclusive methods that enable a wide diversity of users to continue AID therapy are needed. Parallel to findings on the beneficial therapeutic effects of AID therapy, several recent studies [4, 40, 80] explicate the need for human-centered development of AID systems, referring to problems well known in human-automation interaction: positive effects of AID can, e.g., be hindered by a high number of alarms [14] and the associated alarm fatigue [106]. While reducing the burden of treatment [113] is one of the main goals of AID systems, the continuous efforts while using AID systems as well as initial familiarization with this form of therapy are considered important discontinuation criteria for therapies with AID systems [80]. Human-centered improvement of the interaction between intelligent, highly adaptive AID systems and people with DMT1 is therefore a key scientific challenge to improve treatment options for individuals with different levels of experience and competence in using technology. At the same time, AID systems also provide an excellent context to examine the dynamics of human-XAI interaction in a situation where high risks and high benefits for users are juxtaposed.

Problematic expectations and experiences with AID systems play a decisive role in the current acceptance of these systems [72]. For instance, if users have an incorrect understanding (e.g., in the sense of an inaccurate mental model, c.f. [59]), this can lead to incorrect predictions of the results and capability of the system [9]. Such false mental models could result from people being uncertain about how system adaptability affects information processing in AID systems, e.g., whether they are able to change therapy goals or not [67]. In addition, AID systems often work differently than users did when they manually regulated their glucose levels: for example, information is processed by AID systems every 5 minutes [12], while in other forms of therapy (e.g., before using an AID system) the blood glucose level is sometimes only checked, e.g., four times a day with fingerstick glucose measurements [120]. Therefore, AID systems as a case for examining the real-time cooperation of humans with intelligent algorithms potentially lead to an advanced understanding of cooperative disease management between humans and AI. The performance of many AID systems regularly relies on information from the user [20, 116], so correct communication between both partners may lead to increased performance. On the other side, an incorrect understanding of the AID system could also have a critical impact on the success of the therapy [21]. While regulatory technical briefing is mandatory, the extent to which the functions and capabilities of such a system are understood is not tested prior to its use. If users have an incorrect mental model, the ability to correctly predict the information processing of the system may decrease. However, the self-assessment of how well one understands the information processing of a system may differ

from the actual correctness. Explanations could help individuals to recognize errors in their mental model, leading to a better fit between experienced traceability and performance. However, they could also erroneously increase the confidence in an incorrect mental model and thus worsen the calibration [34], which results in wrong expectations about system behavior and potentially confuses users, ultimately leading to a reduction of trust [110]. Explanations can have an ambiguous effect on the calibration between the experienced traceability of a system and the user's ability to correctly predict information processing. To address inaccurate calibration, metrics for both experience and performance need to be measured at the same time. All in all, AID systems represent a prototypical example of interactive systems where human-machine cooperation is centrally influenced by user experience and where incorrect mental models or disparity between experienced traceability and performance may lead to unexpected issues in therapy quality.

The goal-oriented communication of information and the correct predictability of, e.g., an insulin calculation are two central characteristics of human-machine cooperation [62]. In the field of **explainable AI (XAI)**, various approaches exist that are intended to help users cooperate with AI systems by addressing the challenge of opacity (such as [25, 84, 96]. As demonstrated in examples outside of AID therapy, the calculation of results can be presented transparently by revealing weights of relevant factors [100]. Furthermore, the elements that particularly favored certain results can be highlighted [70], or alternatives close to the given result can be presented [23]. In addition to improving predictability, explanations in AID systems could also help improve users' opportunities to exert directability (see [58] and [28]). In DMT1, a loss of "sense of control" is a typical problem users experience [105]. Thus, when using intelligent AID systems, increasing directability could play an important role and influence acceptance. Ultimately, "common ground" is an important prerequisite for cooperation [62]. In the case of AID systems, a common ground could consist of (1) current information on blood glucose levels, physical activity, or food intake; (2) reference values for therapy, i.e., goals; or (3) personalized parameters like insulin sensitivity. Therefore, it is important to disclose relevant elements or information that users can process themselves and use to manually adjust the therapy [95]; see also [111]. However, in order to reduce the workload, many AID systems process information automatically and do not actively share it with the users. These barriers have already led to user-initiated projects enabling access to their data (cf. [97]). Yet, in relation to the clinical relevance and the opportunities for human factors research, empirical studies on how and when to present detailed information on the AID's information processing are still in an early stage of development. Comprehensive and empirical work with a high ecological validity to derive guidelines on how AID systems can be improved to enable cooperation is needed and constitutes an important next step in human-centered diabetes technology.

The objective of the present research was to examine the effects of explanations that vary in the amount of disclosed information as well as repeated interaction on users' subjective perception of trust and traceability in AID systems. To this end, we trained a basic yet prototypical AID algorithm based on artificial yet plausible data and designed a minimalistic AID simulation to create stimuli for an online experiment, where people with DMT1 repeatedly interacted with AID calculations and also predicted AID results. The information available to the algorithm was disclosed to participants to a different extent, in order to create three different experimental conditions. It was investigated whether a greater amount of information leads to higher experienced traceability and trust while task completion time and perceived workload increase. Furthermore, it was analyzed to what extent repeated viewing of explanatory information can lead to an increase in experienced traceability. Similarly, the relationship between experienced traceability and the ability to make correct productions was assessed to allow evaluation of the calibration of the mental model with the system's information processing.

2 RELATED WORK

2.1 Automation in Diabetes Mellitus Type 1

The continuous therapy of DMT1 sometimes can represent a great burden in everyday life for those affected [112]. Many therefore expect the digitalization of diabetes therapy to improve the quality of treatment while at the same time reducing the burden of treatment for patients [68]. This goal is also being pursued by the development of an "artificial pancreas," which allows complete automation of diabetes management [116]. For now, full automation is only possible to a limited extent due to various factors or may be associated with reduced precision of the therapy (c.f. [20]).

AID systems in the form of so-called hybrid closed-loop systems acknowledge those limits while still offering relief for patients. These systems are not fully automated, since a system-dependent level of information or decisions by the user is required. [89] provides a suitable framework that distinguishes four stages of information processing (1) information acquisition, (2) information analysis, (3) decision making, and (4) action implementation) and therefore allows a characterization of AID systems' level of automation. For example, there are already differences between existing systems in **information acquisition** (1): the system described by [12] only requires information on physical activity and food intake, while [45] already no longer requires information on physical activity. In **information analysis** (2), AID systems show a high degree of automation, as this is supposed to be a crucial element of relief for the users. Here, learning systems such as [12] can be distinguished from static systems such as [19]; the latter requires users to manually adjust parameters and thereby increase the quality of information analysis, whereas this is not necessary for self-learning systems. Thus, self-learning AID systems promise continuous improvement in therapy with greater automation, yet may be more complex to understand and to predict for users. The (3) **decision making** of, e.g., administration of insulin can be illustrated very well by the levels of automation presented by [89] and at the same time represents an important feature for interaction design in AID systems. For example, after input, a single suggestion for the administration of insulin can be made (level 4, cf. [92]) or an automatic administration of insulin occurs where the user can intervene but is not informed in any case (level 8). **Action implementation** (4) is performed automatically by many systems in the event of identified insulin needs. However, systems currently available do not offer the injection of, e.g., glucose in case of hypoglycemia, so action implementation for low glucose levels is not automated. All in all, AID systems in their various forms represent not only a broad field of automation in medical systems but also systems that are highly dependent on cooperation between humans and technology.

However, various studies also show the challenges of automation: for example, people fear an error-proneness of digital systems in the field of DMT1, with simultaneous fears to be faced with high complexity [80]. But also, for example, too high expectations of performance or degree of system autonomy, especially of AID systems without a high degree of automation, pose substantial challenges [61, 93]. Furthermore, it remains to be seen to what extent a more technologized therapy could further exacerbate the already existing inequality between individuals from different socioeconomic strata or educational levels. In addition to accessibility (c.f. [69]), the design of systems may also improve unequal opportunities for empowered and autonomous diabetes therapy [74, 87]. These challenges can be addressed with the human-centered development of interactive and cooperative yet traceable AID systems, which could make a decisive contribution to the empowerment of people with DMT1, regardless of their diverse backgrounds, e.g., in terms of affinity to technological interaction or educational level.

2.2 Explanation and Cooperation in AID Systems

Explanations and higher levels of transparency may improve cooperation between humans and intelligent systems [118]. They may support the temporally adequate exchange of information

between humans and the system, which is of central importance for both partners to fulfill their respective functions [47]. In AID systems, for example, the human must signal the intake of carbohydrates in a timely manner, while the system must communicate a deviation in blood glucose levels to the user, for example, so that the human can take action. Mutual anticipation of information demands can be a central criterion of cooperation in the sense of collegiality (cf. [28]). Especially with higher degrees of automation, the human's task can also be to monitor or check results. For this task, the information used by the machine can be a central function for cooperation, as this allows the inputs for the machine calculation to be traced. The extent to which the information processing of a system is accessible for the users and thus also provides the basis for cooperative actions can be described as traceability (unlike the definition of [66], where traceability refers to the creation process of the system and not of an individual calculation). An empirical investigation of the disclosure of information in the context of a decision-making process can therefore make an important contribution to the design of human-centered AID systems. To the best of our knowledge, no results on how different quantities of information contributed to the calculation of insulin needs affect user experience have been published.

However, communication—if it does not take place at the right time—can have negative effects on cooperation or the performance of other functions by a partner [32]. Accordingly, previous research does not show a clear impact of explanations on perceived workload [2]. In the case of AID systems, the existing workload, contrary to their initial purpose, is partly a major problem that could motivate dropouts. In addition, unreliable integration of sensor technology still contributes to the frequent negative perceived interaction with the system based on alarms [80]. Therefore, when developing explanations or other approaches to increase the traceability of results of intelligent systems, the objective and subjective workload should be controlled.

Additionally, information or explanations can influence trust in intelligent systems [9, 107, 126]. In order for trust to be relevant, risk needs to be present [56]. The incorrect dosing of insulin by an AID system can result in significant health consequences, which is why trust can not only be investigated in the present use case but is also addressed as a prerequisite and challenge for AID use [65]. In this context, clinical reviews, as required from professionals in studies regarding medical AI systems [48], are one way to provide evidence of trustworthiness and thus increase "extrinsic trust" [56]. However, clinical evidence does not affect the traceability of systems. Experienced traceability allows for "intrinsic trust" and, as discussed, the possibility of cooperation. Therefore, human factors research calls for studies on trust in AID systems in dependence on explanations as a suitable means to support intrinsic trust.

Findings in the literature on the beneficial effects of explanations are still inconclusive; i.e., different studies observe that the use of explanations did not lead to an objective change in observed behavior. For example, [7] could not find better predictions of AI outcomes even though additional explanations were offered. Similarly, [10] showed that explanations did not significantly increase the joint performance of AI and humans in judging texts. Aggravation of this problem is shown by [29] and [36], where explanations are positioned as "placebic explanations" or even as "dark pattern explanations": these explanations do not contain any information to increase transparency but induce a better experience of the interaction, e.g., in terms of perceived trustworthiness, adversely leading to "unwarranted trust." This could result in overconfidence and thus an unjustifiably high reliance on, e.g., the AID system. Thus, rather than empowering users, explanations could give them a false sense of security. Especially in the automated delivery of drugs such as insulin, interactions must be designed to prevent the development of overconfidence. Accordingly, the study of objective and subjective measures together in experiments is crucial in the human-centered development of AID systems.

2.3 From Situation Awareness to Subjective Information Processing Awareness

To adequately address human-centered research questions in AID systems, instruments to assess traceability-related facets of user experiences of a system's results are necessary. In recent years, different scales to evaluate XAI have been proposed. [51] gave an overview of user experience metrics for XAI, introducing the **Explanation Satisfaction Scale (ESS)**. The ESS was developed to measure the subjective quality of explanations provided by an intelligent system. Being based on multiple existing methods from the field of trust in automation (such as [57]), it incorporates both affective and cognitive implications of explanations (see [76]). The ESS is meant for experts constructing and developing AI systems or experienced users, as they need to rate, e.g., the usefulness of results. In iterative development, also quick interaction with systems needs to provide sufficient data to guide further development. An additional scale allowing inexperienced users, e.g., first-time customers and end-users, to participate is crucial for XAI research because usage of AI-based systems is not limited to experts. Another scale addressing system traceability specifically designed for the medical domain is the **System Causability Scale (SCS)** from [54]. The SCS focuses on a quick overview of the impact of explanations and thus also captures different dimensions, e.g., to what extent users see explanations as transferable to others or whether the explanations fit their own knowledge base. While this allows for a quick general assessment, it is not yet clear to what extent the SCS can also be used for specific, theory-driven questions, e.g., about the traceability of certain decisions. As [127] elaborates in its review, the usability of measurement methods for evaluating explanations depends on the user group, the experimental design, and the specific properties of the explanation. All in all, existing instruments of XAI research for surveying the subjective effects of XAI often refer directly to the added interaction elements, i.e., explanations given by the system [51, 54].

While these instruments could be used in the selection of appropriate explanations, especially at the beginning of the design process or in formative evaluations, a direct comparison, e.g., to a baseline without explanations may be difficult. To address experimental designs with, e.g., a control group, an instrument that aims to measure the subjective effects of explanations and relates to experienced traceability of automated systems rather than directly evaluate explanations themselves would be advantageous. For this purpose we derive *Subjective Information Processing Awareness***(SIPA)** [102] from **Situation Awareness (SA)** theory. SIPA describes "the experience of being enabled by a system to perceive, understand and predict its information processing" [102]. When users act within a dynamic system, they make situation assessments [38], which result in a user state that has been established as SA. SA theory postulates three levels within this assessment: (1) perception, where the state of environmental information in the current situation is perceived; (2) understanding, where comprehension of the current situation is formed; and (3) projection, where future states of the situation are predicted. Previous work on automation demonstrates how SA may play an important role in XAI research: for example, low SA could be the reason for missing anticipation when information needs to be communicated in order to ensure cooperation [109]. SA loss is a known problem in existing research in human-automation interaction [89]. Hence, understanding the effects of automation on SA is important and applicable to XAI. However, current methods to survey SA have often focused on the interaction's context. On the other hand, SIPA focuses on the transparency of relevant elements, understandability, and predictability of information processing as it is relevant for the trustworthiness and traceability of AID systems.

While Situation Awareness focuses on processes within the person, the goal of the SIPA scale is to describe the experience of system properties that lead to SIPA. These can be built up analogously to Situation Awareness. Instead of Perception, the first facet of the SIPA scale is experienced transparency, which describes the extent to which the system interaction allows the user to perceive all relevant elements for information processing. Hence, "Understanding" and "Prediction"

can analogously be positioned as "experienced understandability" and "experienced predictability." The facets adopted in the SIPA scale are thus grounded in the levels described in SA theory and can be clearly placed within the broad discussion of the definition of, e.g., transparency [26]. Thus, transparency, as defined in the SIPA scale, does not refer to, e.g., the goals of the developer or global information on, e.g., training of the model, but to the person's experienced accessibility to information to which the system has access.

To ground the specific items of the SIPA scale in SA theory, we examined different SA scales assessing subjective (c.f. [115]) as well as objective SA (cf. [37]). The items of the scale were developed on the basis of these questionnaires as well as theoretical explanations of situation awareness (e.g., [39, 125]) and discussed by various experts from the field of engineering psychology. The scale, initially developed with 12 items [102], was shortened by multiple, empirically supported iterations to 6 items. Two of the items are assigned to each of the facets of SIPA. While reverse-coded items were sparingly integrated with the original generation of items, these showed the negative effects discussed in [121]. After weighing the comprehensibility of the scale against the potential negative effects of uniformly one-sided items, no reverse-coded item was included in the 6-item scale—also on the basis of qualitative comments from users.

3 PRESENT RESEARCH

Based on the research issues presented above, hypotheses were derived for the present study. For hypotheses H1 to H3 the level of information disclosure is the independent variable, while SIPA, the time-on-task, and the subjective workload are the dependent variables.

- **H1**: SIPA increases when there is an increase in relevant explaining information disclosed by an intelligent system.
- **H2**: Time-on-task increases when there is an increase in relevant explaining information provided by an intelligent system.
- **H3**: Subjective workload increases when there is an increase in relevant explaining information provided by an intelligent system.

Further, we assume that the dependent variable SIPA increases over time, regardless of the condition, as individuals are given repeated opportunities to make assumptions about the system and correct their mental model.

- **H4**: SIPA increases with increasing observations.

As mentioned above, we expect a close relationship between SIPA and trust, since, for example, the experienced predictability of a system as depicted via SIPA is a crucial influencing variable for trust. Furthermore, we expect a strong correlation with ESS due to the similarity of the underlying constructs.

- **H5a**: SIPA and trust correlate moderately to strongly.
- **H5b**: SIPA and explanation satisfaction correlate moderately to strongly.

Hypotheses H6 to H10 relate to participants' performance on the prediction task or the effects of the prediction task. Here, the prediction of insulin needs calculated by the AID system represents a measurement dependent on the correctness of the participant's mental model. Based on previously discussed theories in the area of cooperation, we hypothesize in H6 to H9 that higher availability of information leads to better SIPA and to better prediction. Additionally, we expected the SIPA value to rise in the performance block.

- **H6**: Higher SIPA ratings before the performance block correlate with better performance in the prediction task.

- **H7**: Higher levels of information disclosure lead to better performance in the prediction task.
- **H8**: SIPA increases over the course of the performance block.

The influence of intra-individual differences (such as attitude toward AI or duration of diabetes) could affect the user experience of an AID system. To assess the inclusiveness of explanations, we formulate the following research question for exploratory analysis:

- **EQ**: How are intra-individual differences related to SIPA ratings and performance in the prediction task?

4 METHOD

We conducted an AID simulation experiment among people living with DMT1. Specifically, we examined how different levels of information disclosure affected the participants' experience of an algorithm calculating insulin needs after repeated interaction with varying levels of information disclosure of the system. The study was pre-registered under https://doi.org/10.17605/OSF.IO/NUJTE at OSF [42]. Changes in the planned and performed analyses are described under Results.

4.1 Participants

Eighty participants with DMT1 completed the experiment. Ethics approval for this study was granted by the Ethics Committee of the University of Lübeck before the start of the experiment (Tracking number: 21-438). Participants volunteered to participate in the study, and informed consent was required. The experiment was implemented using the Labvanced online experiment platform [41]. Participants were instructed to conduct the study only with appropriate screen sizes, i.e., on desktop computers, laptops, or tablets. We recruited DMT1 patients via mailing lists and social media channels (Twitter, Facebook, Instagram) applying convenience sampling. Participants were compensated €10 for their time in the study due to the approximated duration of 60 minutes. In addition, the three best-performing participants could win €80 each. This additional price was applied in order to put an additional incentive for motivation into performance tasks on top of the general compensation.

To safeguard data quality, we defined two exclusion criteria before the experiment and applied these after study completion: (1) participants with over-long completion times (>2 *SD*, $N = 2$ with 412 and 319 minutes in comparison to $M = 63$ of final sample) were excluded because participants were instructed to complete the experiment in one single continuous session, and (2) participants with very low knowledge of DMT1 management were excluded because the experiment required the most correct understanding of the relationships between the factors influencing blood glucose. To screen for diabetes knowledge, we developed 10 items (see Appendix C). To be able to assume sufficient uniform knowledge of diabetes management, we defined six correct responses (60% to reach a reliable differentiation from chance) as a cutoff criterion for exclusion prior to the experiment ($n = 1$ excluded with knowledge score = 4, final sample with $M = 7.89$ and $SD = 0.78$). In addition to these pre-defined criteria, we observed in the first data inspection that some users reported the same rating for all items in the observation blocks and excluded them to avoid invalid data being part of the analysis. Furthermore, in the prediction task, we observed users to only respond with "0" or positive values in the prediction class, which caused biased results for the prediction. Overall, seven participants were removed based on those additional criteria.

The final sample consisted of 70 participants ranging from 18 to 61 years ($M = 28.9$, $SD = 10.5$). Forty-nine participants identified themselves as female (70.0% of the sample), 20 as male (28.6% of the sample), and one person as neither. To better classify the sample in relation to the

general population with regard to at least one fundamental facet of user diversity (i.e., diversity in human-technology interaction), the Affinity for Technology Interaction scale [43] was assessed. Our sample had a wide range (from 1.22 to 5.67) with an average value of 4.11 being well in the medium range (possible ATI score range = 1–6) yet somewhat higher than reported for the general population (3.5 as described in [43]). Yet, it has to be noted that the average ATI score in the population of AID users is not known (e.g., there is a chance that low-ATI patients are more reluctant to adopt an AID therapy or treatment). The average duration of diabetes was 14 years (SD = 10.1, $Range$ = 1–44), which is similar to distributions of recent clinical studies for AID systems, such as, for example, [12]. Only $n = 9$ participants stated to have previous knowledge of AID systems. These were evenly distributed across the groups and showed no correlation with performance in the prediction task (all $p > .050$).

4.2 Experimental Environment

To create an experimental environment we developed an AID simulation system that was designed to meet three criteria: (1) high ecological validity for good transferability of the results to the practical application of systems, (2) information that structurally resembles real dynamics in DMT1 treatment with AID systems, and (3) high experimental control, which allows the systematic manipulation of independent variables and thus enables the research questions to be addressed. Further, the application had to be sufficiently distinct from existing systems, which could otherwise have led to potential confounding based on existing experience and prior knowledge. The AID simulation was created in three steps described in the following sub-sections: (1) the manual creation of valid training data, (2) the training of a basic machine learning model for use in the context of a runtime-capable AID simulation, and (3) the generation of static scenarios for a controlled experiment.

4.2.1 Development of Artificial Training Data for AID Simulation. An artificial dataset of information relevant to AID systems was developed to be independent of individual medical data and the complications that come with it in terms of using personal health data. Each instance consisted of 12 different attributes and the insulin requirement. The individual datasets represent different individuals and therefore contain individualized factors as attributes, such as the amount of correction for excessive glucose levels. All attributes and their meaning are found in Appendix A. Negative insulin needs refer to the need to take in carbohydrates when, e.g., too much insulin is in the body. The different attributes are based on data that is already used in various clinically tested AID systems [12, 81]. After creation, the dataset was reviewed by two independent diabetologists. Both independently rated the dataset as plausible. In total, over 480 instances were created, with 400 to train and test a model.

The attributes have been divided into three different groups, following the approach discussed in Related Work: (1) information provided to the system by the user depending on the situation or automatically determined by the system and **representing physiological variables** influencing the amount of insulin, (2) information representing general or dynamic therapy **goals or preferences of the user**, and (3) **information learned by the algorithm, which provides information about the calculated insulin sensitivity** and thus factors influencing the outcome of the AID system. The information of the first group is oriented to give one (1) common ground about information that both humans and machines absolutely need for cooperative action. The information of the second group shows which possibilities the system has for (2) implementing user preferences and can thus give users information about the extent of directability. While all information increases the predictability of the system, the information from the third group represents influencing factors for the concrete (3) computation of the system.

Table 1. Hyperparameters of Applied Random Forest Model

Mean Absolute Error (MAE)	3.02
Mean Squared Error (MSE)	13.69
Root Mean Squared Error (RMSE)	3.70
Mean Absolute Percentage Error (MAPE)	1.52
Explained Variance Score	0.39
Max Error	7.97
Median Absolute Error	2.20
R^2	0.38

4.2.2 Training of Random Forest Model for AID Simulation. Subsequently, a model was trained based on the data. To predict insulin needs based on the dedicated attributes as input parameters, a random forest regressor was implemented [104]; see also [85]. A train-test-split where 25% of the data was reserved for testing was used, resulting in four datasets: X_train, X_test, y_train, y_test. The X datasets include the input parameters for the regressor, while the y datasets only contain the corresponding target values (results).

Through a grid-search cross-validation algorithm, an (on average) best set of hyperparameters for the random forest were found to be 80 estimators and 10 max depth. These parameters are used for the construction of the random forest and control the number of trees in the forest and the max depth of those trees. A lower number of trees would have resulted in an underfitted model, while a higher number of trees (> 100) would not have increased performance further. The maximal tree depth of 10 shows a good performance for the dataset at hand, while deeper trees are more prone to noise in the data.

The random forest was then fitted to the training datasets (X,y) with the hyperparameters. The regression model exhibits metrics when comparing predicted values with real result values (y_pred, y_test) as shown in Table 1.

4.2.3 Generation of Scenarios for a Simulation-based Experiment. The AID simulation was used to generate scenarios for an experiment. The interactive input of individual data was excluded for this experiment in order to (1) have uniform scenarios for each participant and thus avoid biases due to different inputs, (2) focus on scenarios close to the application, and (3) reduce the risk of technical problems in the ongoing experiment in the context of the experiment conducted online.

To create scenarios, calculated insulin needs were removed from the 80 remaining instances of the previously described dataset and used as inputs for the AID simulation. The outputs were saved as screenshots, with all 80 scenarios saved in three different formats and used in the experiment as conditions: (1) **low information disclosure (LowID)**, (2) **medium information disclosure (MedID)**, or (3) **high information disclosure (HighID)**. The allocation of information is based on the groups described above and is presented in Table 2.

The resulting interfaces can be seen in Figure 1. Participants consistently saw only one of these conditions throughout the experiment, in both the observation and performance blocks. Because of feedback in pre-tests, the concept of correction strength was explained to all participants from MedID and HighID before each block of stimuli.

4.3 Measures

4.3.1 SIPA Scale. The SIPA scale as a measure to assess users' experience while interacting with intelligent systems was used to examine the effects of different levels of information disclosure. The goal for the development of the SIPA scale was to construct a highly economical scale closely linked

Table 2. Overview of Attributes Used in the Simulation

Condition	Attributes
Low Information Disclosure (LowID)	Current Tissue Glucose Current Insulin in Body Current Carbohydrates in Body Current Activity
Medium Information Disclosure (MedID)	Tissue Glucose Target Avoid Hypoglycemia Duration of Insulin Effect Correction Intensity
High Information Disclosure (HighID)	Risk of Hypoglycemia in Next Hour Blood Glucose Lowering per 1 Unit Insulin Insulin Units per 10 Grams Carbohydrates Predicted Exercise

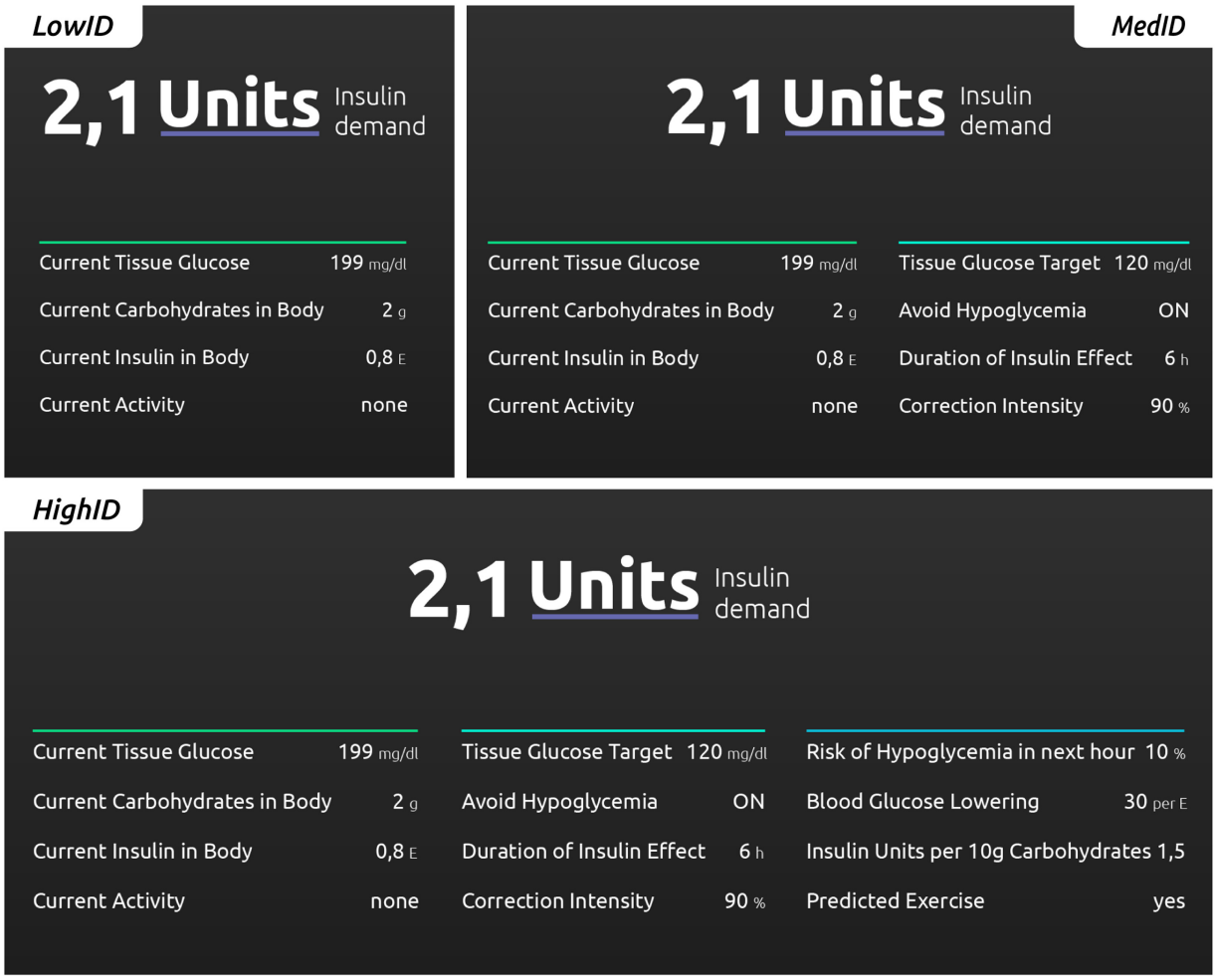


Fig. 1. Stimuli from the study as they were shown to participants for the three conditions: LowID, MedID, and HighID.

Table 3. All Items of the Subjective Information Processing Awareness (SIPA) Scale and the Corresponding Instruction

The following questionnaire deals with your **experience in the interaction with the system.** **Information** refers to all data that the system can work with. **Result** refers to the output of the system, which is presented at the end of the system's information processing.							
Please indicate the degree to which you agree/disagree with the following statements.	completely disagree	largely disagree	slightly disagree	slightly agree	largely agree	completely agree
01 It was transparent to me which information was collected by the system.						
02 The information that the system could acquire was observable for me.						
03 It was understandable to me how the collected information led to the result.						
04 The system's information processing was comprehensible to me.						
05 With the information accessible for me, the results were foreseeable for me.						
06 The system's information processing was predictable for me.						

to SA but focused on an application in intelligent automation, respectively XAI. Additionally, the scale is specifically designed to assess the three facets of SIPA as described above (see Related Work) with two items for each facet (one and two for transparency, three and four for understandability, and five and six for predictability). All items are shown in Table 3.

The 6-item SIPA scale uses a 6-point Likert response scale from completely disagree = 1, largely disagree = 2, slightly disagree = 3, slightly agree = 4, largely agree = 5, to completely agree = 6. The SIPA scale introduced in the present article was additionally tested over all points of measurement of SIPA with a three-factor structure to examine if a separate evaluation of the three individual facets of SIPA was supported. Here, the approach to analyze three facets received support based on a confirmatory factor analysis demonstrating a good fit with χ^2 (6) = 7.49, p = .278, *CFI* = .997, *TLI* = .992, *RMSE* = .06 (90% CI: .00, .17). The correlation between transparency and understandability was significant (r_S = .64, p < .001), which was also true for the correlation between transparency and predictability (r_S = .53, p < .001) as well as for the correlation between understandability and predictability (r_S = .79, p < .001).

4.3.2 User Diversity Variables. User diversity can have a significant impact on the individual user experience and, for example, influence initial trust in a system [8]. To examine the role of user diversity on the experience of interaction with an AID system, two additional variables were collected: (1) **affinity for technology interaction (ATI)** [43], which is based on the personality trait need for cognition [24] and describes the individual tendency to actively engage in intensive technology interaction. ATI was measured with a scale validated in various large samples [43], and the present sample was assessed as rather affine to interact with technology (see Section Participants above). Furthermore, the (2) individual attitude toward artificial intelligence was surveyed. To this end, a brief definition of artificial intelligence was first given. Based on this, six statements from the Internet Attitude Scale [60] were adapted, with "Internet" as the subject being replaced by "Artificial Intelligence" in all used questions (see Appendix). A mean value was calculated to evaluate the **Artificial Intelligence Attitude (AIA)**. In addition, questions on prior

diabetes knowledge were used (see Appendix). This included 10 different statements about the treatment of diabetes to ensure that the results of the study were not affected by significant differences in prior knowledge about the treatment of diabetes. Everyday examples of the treatment of type 1 diabetes or questions about how insulin works were used. Finally, the duration of diabetes in years was requested.

4.3.3 Subjective Measures for Trust, Satisfaction, and Workload. In addition to the SIPA scale, subjective variables were collected with economical scales. The **Facets of System Trustworthiness (FOST)** Scale [117] was used to measure trust. With 5 items, this can be used much more economically in a repeated-measures experiment compared to, for example, the more widely used scale of [57]. As for trust, the mean value of the FOST items was calculated for each point of measurement.

The perceived workload was collected through the **NASA Task-Load-Index (NASA-TLX)** [49]. However, due to the experimental conditions, not all dimensions of the NASA-TLX were used, but the question about perceived physical workload was excluded. Furthermore, the results for effort, mental demand, and time demand were summed to a mean value. Experienced frustration was evaluated independently of other values. The estimation of own performance was only used as a confidence measure after the subjects themselves made a prediction of the algorithm's results. Additionally to SIPA and trust, the **Explanation Satisfaction Scale (ESS)** was measured to allow a comparison to another scale examining the quality of explanations [51]. The ESS was developed to measure the subjective quality of explanations provided by an intelligent system.

4.3.4 Objectives Measures for Performance and Time-on-task. In the present experiment, **time-on-task (TOT)** and a performance indicator were assessed as objective variables. For TOT, the time that the users spent in the different task blocks was measured in seconds. For the analysis, the sum of the time in seconds was calculated. For the assessment of the performance, 20 of the 80 stimuli created with the AID simulation environment were changed in such a way that no prediction of the algorithm was displayed, but the different levels of information disclosure were (depending on the condition). Participants were prompted to estimate the output of the algorithm (this could be negative or positive with one decimal place, or the "0"). The deviation of each estimate was determined per person and a mean value was calculated, which was used as an indicator of performance.

4.4 Procedure

The study was conducted in German. In the beginning, the participants were instructed to watch a video where an instructor of the study explained the purpose of the study as well as the tasks. The spoken text was displayed later in written form and could be read again if needed. Afterward, informed consent was obtained from all participants. The experiment was conducted in multiple segments as depicted in Figure 2: first, demographic data was collected (1); then, knowledge questions about diabetes were asked to minimize the effects of divergent prior knowledge (2). Subsequently, all participants were randomly assigned to one of three conditions: low, medium, or high level of information disclosure. Depending on this, 15 stimuli were shown in random order in an (3) Observation Block, after which SIPA, FOST, and the NASA-TLX were queried. Three additional observation blocks with other stimuli followed by SIPA, FOST, and NASA-TLX followed (blocks 4–6). Subsequently, the ESS was surveyed (7). Finally, in a performance block (8), 20 stimuli were presented in which participants had to estimate for themselves the insulin needs calculated by the algorithm. The stimuli again differed in the level of information disclosure and were stimuli the participants did not see before. However, the same instances were shown to all participants in a randomized order (i.e., each participant saw the same tasks, but with different information being

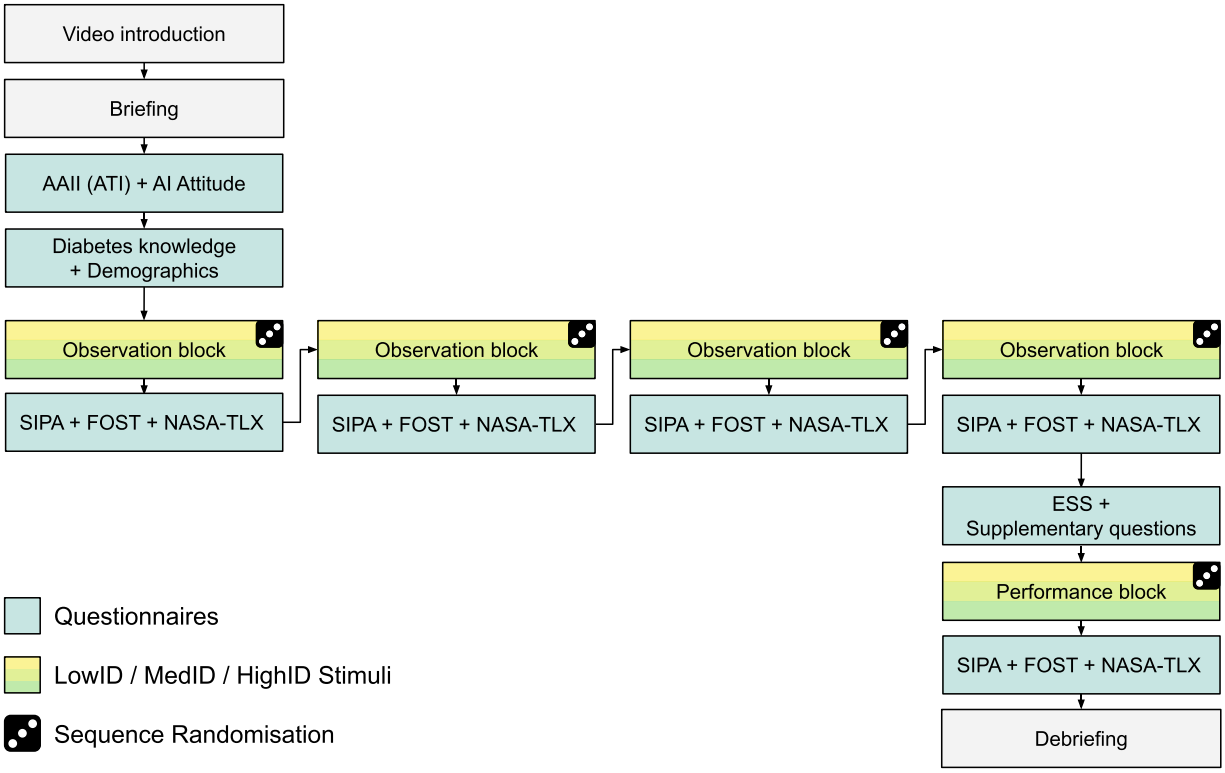


Fig. 2. Overview of course of the experiment.

presented and in different sequences depending on the condition they were assigned to). Then, SIPA, FOST, and NASA-TLX were collected again. Furthermore, the time for each observation block as well as for the performance block was collected. Depending on the individual deviation from the correct calculated insulin needs, a code was created and displayed to the participants in the last frame of the study. To ensure the anonymity of all subjects, the code only corresponded with the deviation and didn't give any indication of personal information.

5 RESULTS

As a direct test of our hypotheses, we applied contrast analysis, which allows for more precise testing of hypotheses [22, 123]. Note that this approach was different from our pre-registration, where we only described an ANOVA. Yet, as an omnibus F-test, ANOVAs are not optimal in order to test the directed hypotheses within the present research. Hence, in order to fit the statistical method used with the specificity of our hypotheses, contrast analysis was chosen. Note that contrast analysis results in t- rather than F-values, also for comparisons of more than two groups [123]. The core hypotheses H1 to H5 related to the development of user experience in repeated observations were part of the pre-registration. Additional hypotheses H6 to H10 relate to performance or self-assessment of performance and were not pre-registered. One-tailed t-tests were conducted to assess the hypotheses. All p-values were corrected for family-wise error [13] for each hypothesis and variable using the Bonferroni-Holm correction [53]. Despite random assignment, not all groups are exactly equally distributed ($n = 24$ for LowID, $n = 22$ for MedID, and $n = 24$ for HighID). Since multiple variables studied were not normally distributed (or no linearity could be assumed), Spearman's Rho was calculated for all correlations and interpreted accordingly depicted as r_S. Effect sizes for r and r_S were interpreted based on [44, 98]; effect sizes for d were analyzed according to [30] with respect to [44]. Cohen's d was reported for contrast analysis of dependent measures instead of Hedge's g because both are almost equal in sample sizes greater than 20 [63].

Table 4. H1: Contrast Analysis for Each SIPA Facet Comparing Ratings between Conditions (LowID, MedID, and HighID) for All Blocks

Block	SIPA Transparency			SIPA Understandability			SIPA Predictability		
	t	p	$r_{(effect\ size)}$	t	p	$r_{(effect\ size)}$	t	p	$r_{(effect\ size)}$
Observation Block 1	0.32	.375	.04	−1.03	.612	−.13	1.14	.258	.14
Observation Block 2	1.89	.063	.23	−0.39	.349	−.05	1.35	.363	.16
Observation Block 3	2.37	.031*	.29	0.64	.786	.08	0.36	.360	.04
Observation Block 4	2.47	.032*	.30	0.56	.578	.07	1.16	.375	.15
Performance Block	2.46	.040*	.29	1.56	.309	.19	2.08	.104	.25

Note: $df = 67$ for all analyses.
*$p < .050$. **$p < .010$. ***$p < .001$.

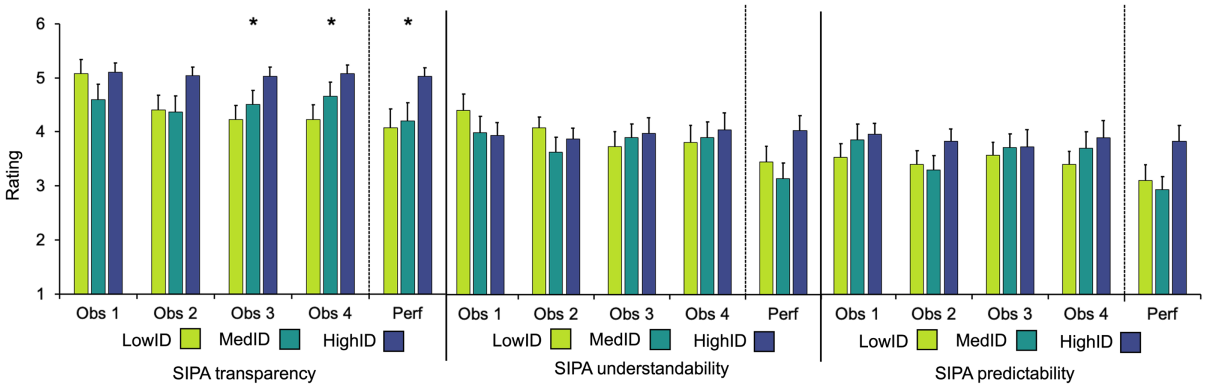


Fig. 3. H1 and H4: Ratings of the SIPA scale for all points of measurement. Bars depict M and SE for all SIPA facets at each time measured. * indicate $p < .050$ for contrast analysis, as shown in Table 4.

5.1 H1: SIPA Increases When There Is an Increase in Relevant Explaining Information Disclosed by an Intelligent System

H1 was examined using multiple contrast analyses [22, 123], one for each SIPA facet (transparency, understandability, and predictability) and for each point of measurement. The different amounts of information disclosed to each group and the corresponding relationship between attributes were used to determine the weights (i.e., lambda values). It is assumed that each attribute (i.e., a total of LowID: 4, MedID: 8, or HighID: 12) can be related to each other attribute seen in one condition. The number of relations between attributes is given by the binomial coefficient (i.e., the number of attributes over two). Thus, the number of relations between attributes is for LowID = 6, for MedID = 28, and for HighID = 66. Following [22], to calculate the weights, the following lambda values for the contrast analysis were defined: $\lambda_{\text{LowID}} = -2.5$, $\lambda_{\text{MedID}} = -0.5$, $\lambda_{\text{HighID}} = 3$. Table 4 shows the t-statistics, the corrected p-value, and $r_{(effect\ size)}$. M and SE are depicted in Figure 3. All descriptive data can be found in Appendix B. Results regarding the SIPA facet of transparency supported H1 for observation blocks 3 to 4 and the performance block, while the first observation blocks 1 to 2 did not show significant effects supporting H1 (see Table 4). The other two SIPA facets, understandability and predictability, showed weak effects in the expected direction, which were all non-significant (except ratings for SIPA understandability after Observation Block 1 and Observation Block 2, which were small but contrary to the hypothesis). Hence, H1 was supported for experienced transparency after considerable experience with the system, yet not directly after the first interaction and not for the properties of the more complex system measured by SIPA (i.e., understandability and predictability).

Table 5. H2: Contrast Analysis Comparing Time-on-task between Conditions (LowID, MedID, and HighID) for All Blocks

Block	Time on Task		
	t	p	$r_{(effect\ size)}$
Observation Block 1	1.83	.107	.22
Observation Block 2	1.03	.153	.13
Observation Block 3	1.51	.136	.19
Observation Block 4	1.82	.146	.22
Performance Block	4.20	<.001***	.47

Note: $^{*}p < .050$. $^{**}p < .010$. $^{***}p < .001$.

Table 6. H3: Contrast Analysis Comparing Subjective Workload between Conditions (LowID, MedID, and HighID) for All Blocks

Block	NASA-TLX		
	t	p	$r_{(effect\ size)}$
Observation Block 1	−0.92	.540	.03
Observation Block 2	−1.45	.304	.10
Observation Block 3	−0.69	.492	.04
Observation Block 4	−0.18	.429	.03
Performance Block	−1.64	.264	.12

5.2 H2: Time-on-task Increases When There Is an Increase in Relevant Explaining Information Provided by an Intelligent System

To test H2, multiple contrast analyses were used. The corresponding results can be found in Table 5. Contrary to the hypothesis, there was no significant difference between the groups for all blocks, apart from one exception (performance block). Interestingly, a medium effect aligned with the hypothesis was present in the performance block. Thus, the performance block stands out and supports the hypothesis, while the data of the observation blocks do not.

5.3 H3: Subjective Workload Increases When There Is an Increase in Relevant Explaining Information Provided by an Intelligent System

To test H3, multiple contrast analyses were used. The corresponding results can be found in Table 6. Contrary to the hypothesis, in all blocks, workload ratings were not significantly higher in conditions with more information. Indeed, negative signs in t-statistics at all points of measurement indicate that the effect was actually in the other direction (i.e., more information disclosure decreases workload). In fact, an exploratory re-calculation of the contrast with inverted weights (i.e., $\lambda_{\text{LowID}} = 3$, $\lambda_{\text{MedID}} = -0.5$, $\lambda_{\text{HighID}} = -2.5$) of the effect would support an oppositely formulated hypothesis, e.g., with $p < .001$ and $r_{(effect\ size)} = .36$ for Observation Block 1.

5.4 H4: SIPA Increases with Increasing Observations

To test H4, multiple contrast analyses were conducted for each SIPA facet (transparency, understandability, and predictability) but followed the contrast analysis for dependent measures [103]. The following weights were used for each analysis: $\lambda_{\text{Observation 1}} = -1.5$, $\lambda_{\text{Observation 2}} = -0.5$, $\lambda_{\text{Observation 3}} = 0.5$, and $\lambda_{\text{Observation 4}} = 1.5$. Table 7 shows the t-statistics, the corrected p-value, and d. Counter to our hypotheses, SIPA ratings did not increase but decreased and the actual effect of repeated observations was opposite to what we hypothesized. In fact, a follow-up calculation with

Table 7. H4: Contrast Analysis Comparing Repeated SIPA Ratings for Observation Blocks 1–4

Facet	Contrast Analysis for Obs 1–4		
	t	p	d
SIPA transparency	−1.73	.956	0.21
SIPA understandability	−0.95	.827	0.12
SIPA predictability	−0.40	.827	0.05

Note: $^{*}p < .050$. $^{**}p < .010$. $^{***}p < .001$.

Table 8. H5a: Correlations between Trust and SIPA Facets for Each Point of Measurement

	Block	SIPA					
		Transparency		Understandability		Predictability	
		r_S	p	r_S	p	r_S	p
Trust	Observation Block 1	.58	<.001***	.76	<.001***	.64	<.001***
	Observation Block 2	.60	<.001***	.85	<.001***	.80	<.001***
	Observation Block 3	.64	<.001***	.84	<.001***	.82	<.001***
	Observation Block 4	.65	<.001***	.84	<.001***	.79	<.001***
	Performance Block	.72	<.001***	.81	<.001***	.76	<.001***

Note: $^{*}p < .050$. $^{**}p < .010$. $d^{***}p < .001$.

inverted contrasts significantly supported the assumption of decreasing ratings for transparency with $p = .44$, while $p > .050$ for understandability and predictability.

5.5 H5a: SIPA and Trust Correlate Moderately to Strongly

To test H5a, the correlation between the FOST scale scores and each SIPA facet was calculated for each point of measurement. The results are shown in Table 8. The range of effect sizes of the correlation across all facets is between $r_S = .58$ and $r_S = .85$, which indicates a strong relationship. Overall, the hypothesis can therefore be supported by the data.

5.6 H5b: SIPA and Explanation Satisfaction Correlate Moderately to Strongly

To test H5b, the correlation calculated between each SIPA facet for Observation Block 4 with ESS was calculated. All facets of SIPA showed a significant correlation (all $p < .001$), with transparency $r_S = .57$, understandability $r_S = .67$, and predictability $r_S = .65$ indicating a strong correlation, which supports the hypothesis.

5.7 H6: Higher SIPA Ratings before the Performance Block Correlate with Better Performance in the Prediction Task

To test H6, the correlation between each SIPA facet for Observation Block 4 with the overall performance was calculated. No significant correlation was found for transparency ($r_S = -.11$, $p = .850$), understandability ($r_S = -.17$, $p = .355$), or predictability ($r_S = -.08$, $p = .731$). Thus, a correlation between the SIPA ratings before the performance block and the performance cannot be assumed and the hypothesis is not supported.

5.8 H7: Higher Levels of Information Disclosure Lead to Better Performance in the Prediction Task

To test H7, a contrast analysis was performed. The weights correspond to the weights used in H1 with $\lambda_{\text{LowID}} = -2.5$, $\lambda_{\text{MedID}} = -0.5$, and $\lambda_{\text{HighID}} = 3$. A one-tailed significance test with ($t(67) = 1.21$,

Table 9. H8: Contrast Analysis Comparing SIPA Facets before and after Performance Block

Facet	Contrast Analysis for Obs 1–4		
	t	p	d
SIPA transparency	−2.26	.986	−0.28
SIPA understandability	−2.40	.991	−0.29
SIPA predictability	−2.19	.984	−0.27

Note: $^{*}p < .050$. $^{**}p < .010$. $^{***}p < .001$.

Table 10. Results of Explorative Analysis

	Facet	Block	ATI		AIA		Duration of Diabetes	
			r_S	p	r_S	p	r_S	p
SIPA	Transparency	Observation Block 1	.29	.080	.38	.024*	−.29	.098
		Performance Block	.42	.007**	.29	.144	−.10	>.999
	Understandability	Observation Block 1	.24	.150	.36	.115	−.25	.240
		Performance Block	.24	.192	.14	.256	−.01	.961
	Predictability	Observation Block 1	.23	.104	.27	.014*	−.21	.410
		Performance Block	.35	.018*	.18	.099	.03	>.999
Performance			−.04	.731	.05	.666	−.07	>.999

Note: $^{*}p < .050$. $^{**}p < .010$. $^{***}p < .001$.

$p = .116$, $r_{(effect\ size)} = .15$) did not detect a significant difference between the groups, and thus there was no support for the hypothesis.

5.9 H8: SIPA Increases over the Course of the Performance Block

To test H8, multiple contrast analyses were conducted for each SIPA facet following the contrast analysis for dependent measures. The following weights were used for each analysis: $\lambda_{\text{Observation 4}} = -1.5$ and $\lambda_{\text{Performance}} = 1.5$. A one-sample t-test against zero was performed for all contrasts. Table 9 shows the t-statistics, the corrected p-value, and d.

The hypothesis is not supported by the results for any of the SIPA facets. However, all facets show a high negative t-statistic, which suggests that the contrast was chosen in opposition to the real data. This corresponds to the descriptive observation that there was not a successive increase but a decrease in the SIPA ratings for all facets. The calculated effect sizes also indicate a relevant effect at the boundary between small and medium effects. Under the assumption of opposite contrasts, significant effects are shown for transparency ($p = .014$), understandability ($p = .010$), and predictability ($p = .016$).

5.10 EQ: Explorative Analysis of Individual Differences

To examine the relationship between individual differences in human-AI cooperation and user experience, correlations between person characteristics (ATI, AIA, duration of diabetes) and SIPA ratings as well as performance were calculated. The measurements for Observation Block 1 and the performance block were analyzed in order to keep the number of tests (and the resulting loss of power due to correction) low. All values are shown in Table 10. There was no correlation between the duration of the disease and the SIPA ratings or the performance. With regard to the ATI values, no correlation can be found at the beginning of the experiment. At the last time point, there is a small to moderate effect (for SIPA transparency and SIPA predictability). For AIA, no significant effects are found at the end of the study, but at the beginning of the experiment, there are moderate,

significant correlations with SIPA transparency and SIPA understandability. Neither ATI nor AIA shows a significant relationship with performance.

6 DISCUSSION

6.1 Summary of Results

The objective of the present research was to examine the effects of explanations that differ in the amount of disclosed information as well as the effect of repeated interaction on users' subjective perception of trust and traceability in AID systems. Contrast analyses were performed to test directional hypotheses related to the dependent variables SIPA, TOT, and subjective workload.

While results showed a weak tendency for users in the HighID condition to report higher SIPA ratings than users in the LowID condition, the assumed contrast (increasing SIPA ratings with an increasing quantity of disclosed information) was only significant for SIPA transparency after multiple interactions (i.e., after 45 observations) and aligned with hypothesis **H1**. The time users spent on the prediction task was more than twice as high for users in the HighID condition than for users in the LowID condition. Thus, a significant raise of TOT based on higher information disclosure as stated in **(H2)** could be found when participants were asked to predict the insulin needs calculated by the system. In contrast, only non-significant and slight differences were found when people were instructed to observe stimuli displaying the insulin needs calculation. Although the subjective workload did not increase significantly with the level of information disclosure as assumed **(H3)**, an unexpected effect emerged: the perceived workload was higher for the LowID condition than for the HighID, in some cases more than one standard deviation higher. The development of the SIPA rating over time also shows, contrary to our expectation **(H4)**, a decrease. This effect was small for SIPA transparency, while only negligible effects can be observed in the other facets. A strong correlation between all SIPA facets and trust **(H5a)** as well as between all SIPA facets and explanation satisfaction **(H5b)** indicates high convergent validity for the SIPA scale. SIPA ratings prior to the performance block did not correlate with performance itself and also showed very small effects **(H6)**, although SIPA transparency ratings differed significantly before observation for different levels of information disclosure. Although more information was available in the MedId and HighID than in the LowID condition, participants in the MedID or HighID condition did not perform significantly better than participants in the LowID condition **(H7)**. The prediction task in the performance block did not lead to an increase in SIPA but resulted in lower SIPA scores in all facets with a medium to strong effect size **(H8)**. Analysis of intra-individual correlations with SIPA revealed that SIPA was significantly related to attitudes toward AI after the Observation Block1, while ATI showed a significant influence after the Performance Block **(EQ)**.

6.2 Effects of Information Disclosure on User Experience and Cooperation in AID Systems

One focus of the present work was to investigate the effect of different levels of information disclosure on the user experience of AID systems. However, higher information disclosure did not affect SIPA immediately but led to a significant difference in perceived transparency only after 45 observations. The delayed decrease in SIPA transparency ratings suggests that a valid measurement of subjective variables may need an experimental design with sufficient repetitions (cf. for trust [46, 52]). [124] discusses the complexity of trust developed over time and distinguishes between three different phases: learning, adjustment, and fine-tuning. These phases follow each other and could explain the trust development we found after several repetitions as well as explain the effect triggered by the performance block. Our results may indicate that the development of trust at different stages is based on content as well as temporal reasons, i.e., that, for example, new

tasks such as the estimation task trigger a readjustment of trust. While individuals in the LowID condition started with a comparably high level of SIPA transparency, the observed decrease could be related, for example, to the fact that only repeated observations allowed them to recognize that not all necessary information was available. [99] describes that a person's mental model is used to form expectations about the outcome, e.g., of cooperation with automation. As for trust (cf. [50]), individual differences could affect the initial SIPA rating, and only measurement after a system-dependent number of interactions can reveal differences between systems. This is also exemplified by the co-relationship between AIA and SIPA transparency at initial observation and after the performance, which indicates that explanations may be able to offset the effects of initial mistrust of individuals. The relationship between attitudes toward AI systems (such as AIA) and other user diversity factors (such as education level or access to technology) represents another research challenge to explore the effects of explanations more in-depth.

Another reason participants in the HighID or MedID condition did not show better prediction performance could be information overload. Information overload occurs when an increase in the available amount of information leads to negative results, e.g., a decrease in performance or subjective consequences (e.g., as experienced cognitive demand or stress) for the user [71]. Although there were three times as much information available in the HighID condition as in the LowID condition, the TOT for the observation blocks did not differ significantly between the groups. [5] assumes that a high information workload can lead to the use of heuristics (e.g., the representativeness heuristic) or increase the probability of users making biased decisions. This effect is opposed to one goal of XAI design, which is to mitigate errors based on heuristic decision-making [119]. In our AID simulation experiment, the use of heuristics while observing might have been higher for the HighID condition than for the LowID condition. This could explain why TOT did not increase (for the observation blocks) though more attributes were presented. The results of the NASA-TLX on subjective workload allow a parallel conclusion: experienced time demand, cognitive demand, and effort showed no difference between the conditions. It is very unlikely that the participants of the HighID condition did not notice or ignored the additional information, as they partly referred to it in the qualitative comments. While being already discussed [94, 119], the extent to which explanations or the additional information available through explanations creates an information overload and thus influences, for example, the use of heuristics in the evaluation (see also [35]) of an AID system still needs to be investigated more clearly and for users of different levels of expertise. [114] found that, for example, the expertise of users can decrease the probability that they will use heuristics. However, AID systems, in particular, have great potential for individuals with problematic long-term metrics, which in turn may often be due to low engagement with and care for the disease. For an inclusive design of AID systems, the effects of explanations for less experienced users must be understood and avoided, in case they cause, e.g., limited transparency. Representations that lead to a heuristic assessment due to information overload could thus encounter users for whom a heuristic assessment could be particularly problematic. All in all, when designing XAI and in order to act responsibly, developers should consider that more access to information may be harmful to transparency and elaborated context analyses are needed to understand how users will interpret and utilize information or explanations.

Finally, the qualitative results point to another problem, as participants from the LowID conditions explicitly ask for information that was presented to the other groups, e.g., LowID-1: "Please add the probability of hypoglycemia or intensity of correction" or LowID-2: "Please show correction quotas for glucose and carbohydrates." However, the results of the experiment suggest that this does not necessarily allow for higher SIPA or better prediction. In order to achieve higher SIPA, the individual pieces of information presumably need to be put into better proportion, as HighID-1 expresses: "I need refined information, how much insulin is given to correct glucose

levels and how much is given for food." The requirement for a more mathematical description could be due to the fact that users apply their mental models of how they would solve the problem without an AID system to the system's information processing. In the field of AID systems, users potentially perform a complex calculation, through which they have certain expectations, as HighID-2 states: "I would like to see the highlighting of factors that are particularly decisive for the calculation at that moment." In future explanations of AID systems, the representation of the calculation should be as close as possible to the calculation performed by the users (as depicted by [119]) in order to empower users to assess the system's information processing. This would also meet a central criterion for cooperation, where adequate communication of information requires partners to anticipate the relevance of the information for the task of the cooperating partner.

6.3 Fit of Performance Measures and Subjective Measures in XAI

In our experiment, the participants' own assessment of the system's traceability does not correlate with their ability to predict the system's calculation. This is a worrisome correlation since in the best case false expectations arise and people lose confidence in the system. A more serious consequence could be, for example, a misjudgment of the system's performance in extreme situations and the development of overconfidence. Several studies [78, 79] on trust in automation show that a lack of calibration between subjective ratings and objective scores is a well-known phenomenon. This miscalibration can lead to significant problems; e.g., complacency arises and thus the users attribute more competencies to the system than it possesses [89], which is described as an abuse of the system [88]. On the other hand, mistrust can lead to a misuse of the system [88]—in the case of the AID system, suggestions of the system could be corrected frequently and thus lead to an increase of the workload instead of a reduction. Both forms of lack of calibration are significant problems in the AID domain and could help to explain dropout rates [80]. The calibration of SIPA and the correct prediction of an outcome is theoretically more direct than the calibration between prediction and trust (e.g., I can trust the technical competence of a system without understanding how it works; see [76]) and can be used in future studies to show the miscalibration between user experience and the correctness of one's mental model. [99] describes a user's mental model as a "mechanism whereby humans generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future system states." This is also in line with central concepts of SA theory or the idea of so-called situation models [11]: here, mental changes are carried out in order to assess the effects of one's own actions. However, figuring out how changing input variables affects the outcome of an AI's information processing may be complicated in the case of static explanations (c.f. [1]). Also, [27] shows that static explanations have a smaller influence on the ability to understand a system than interactive explanations. The latter allows users to build hypotheses on their own and test them, which is the central approach for knowledge acquisition (c.f. [91]). Interactive explanations should therefore be made possible for AID systems (and other intelligent systems). At the same time, future experiments should focus on observing the formation of hypotheses and their evaluation in the interaction between humans and AI, e.g., to identify when explanations favor confirmation bias or disadvantage individuals with less prior knowledge and how those effects can be mitigated.

This is also supported by the fact that the prediction task had a clear influence on SIPA ratings—all facets of SIPA were reduced, while this was not the case for SIPA understandability and SIPA predictability even after 60 previous repeated (passive) observations. The information provided (i.e., the attributes) was not changed for the performance block. In further studies or development of AID systems, active prediction of AID results should therefore be part of the experimental condition and based thereon considered in training. The role of feedback for SIPA as well as trust should again be considered separately. For example, the diagnosticity [16] or the diagnostic value [122] of

certain attributes (i.e., what informativeness they had in determining insulin needs calculated by the system) might have been misjudged by individuals. This could be corrected by feedback or an interactive simulation. Since participants were not able to provide feedback to the system, this lack of interactivity could also be one factor that led users to rate the systems' trustworthiness as they did (c.f. [55]). The participants' passive role as observers for a large part of the experiment might have influenced trustworthiness ratings. A rather active opportunity to intervene, e.g., a system with adjustable attributes or weights of components such as the current target, could have had an impact on the development of perceived trustworthiness.

Another obstacle, however, is the information overload discussed above, which could also arise in an interactive simulation. While, e.g., explanations on the basis of "counterfactuals" [82] may be well suited for testing hypotheses, more research needs to examine how larger numbers of, e.g., setting possibilities affect the interaction. In the exemplary case of generative visual models, the cognitive load of the user increases with the number of adjustable settings, without a significant effect on performance [31]. Furthermore, it must be considered whether and which additional information is displayed e.g., in a training context or in a daily use context, since these may differ considerably with respect to the available time and cognitive resources. Here, explanations need to be designed for diverse users (i.e., the trainer, which is often a medical professional, as well as the patients). The fact that more attributes lead to a higher time requirement for the derivation of a prediction was also shown in the present experiment (see H4, Performance Block). Overall, context-specific prioritization of information must be made, which could be done based on the following questions: (1) Does the representation of attributes/relationships fit the existing mental model of the users? (2) Does the presentation of attributes/contexts allow for hypothesis generation and testing?

6.4 Research and Design Implications for AID Systems

For the research of experienced traceability of intelligent systems, the SIPA scale with its facets allows for two central observations: (1) a sufficient number of repeated interactions and (2) a differentiation of active interaction from passive observation of explanatory information disclosure are necessary to discuss human-centered AI. The SIPA scale is an appropriate instrument for this context for the following reasons: the SIPA scale shows good scale metrics (i.e., range, standard deviation) on all facets. Additionally, due to the high correlation between all three SIPA facets, a unidimensional application is also possible. Furthermore, the SIPA scale shows a very high convergent validity with measures of perceived trustworthiness and satisfaction with explanations. However, there is a small to medium correlation between ATI and SIPA, and the ATI mean of the present sample is higher than the estimated population mean. Hence, the use of the SIPA scale in groups with lower ATI scores might be different, e.g., shows other correlations with satisfaction. Overall, the SIPA scale with its facets represents a new tool for researching experienced traceability, which can help to underscore and evaluate the effects of explanations on users in detail.

The boundary between Situation Awareness and Performance (i.e., Prediction) has already been raised repeatedly in the discussion of Situation Awareness [90]. While a theoretical discussion of these concepts is beyond the scope of this article (c.f. [77]), a very high crenelation between SIPA understandability and SIPA predictability suggests that the difference between Understanding and Predicting might be too small to provide an impactful analysis. Studies using other explanatory approaches would need to investigate whether this difference can be amplified. In addition, qualitative comments from users suggest that another facet of Traceability may be relevant—the assessment of the relevance of attributes to the information processing, explicated, e.g., from MedID-2: "Display to what extent which information contributed to the result," which possibly refers to the individual attribute's influence or relevance for the prediction (i.e., diagnosticity; c.f. [16]).

The extent to which the presented information has a high, subjective diagnosticity could be distinguished from predictability as a facet. For example, an AID system's user might know that providing information about exercise intensity is more important than providing information about the duration of the physical activity. The user would feel able to instruct the AID system to achieve a more precise prediction, regardless of the user's ability to specify the concrete outcome. Especially for the communicative processes in the field of human-AI cooperation, such an additional facet could enable, e.g., what [28] describes as collegiality.

When designing AID systems, the effects on the experienced traceability as well as on workload and performance must be taken into account. The sole disclosure of additional information cannot be seen as a suitable method to improve the user experience or the basis for human-AI cooperation in AID systems. In the given scenarios, the information used from the AID simulation was relevant for the calculation of the system and mimics information that users themselves need for a calculation. The fact that this approach did not offer a significant advantage for the participants of the HighID condition shows how much human-centered research is still necessary for the XAI area. In XAI research explanatory approaches partly refer to the confidence of the model [10, 17, 86] for a certain result or even to meta-information about the model [75]. Depending on their task, such information might have only low significance for the users. This could lead to erroneous conclusions in the future, especially if the methods to evaluate the performance of human-AI cooperation are based on different processes than the processes supported by the explanation. Regardless of how helpful certain methodologies are to AI method developers, users as well as the constructs and requirements relevant to them may be entirely different and need different explanations. Even among the users of a system (in the broadest sense), there might be differences. That is, the information presented in our experiment might help individuals with medical training who, for example, match the model's approach to guidelines on therapies and for whom a more abstract interaction might provide more information. Individuals, on the other hand, are more likely to want to interact with the system on an individual level, as shown by LowID-3: "I would like to enter an individual target value for physical activity." [73] distinguishes between local and global explanations of an AI system. However, to assume that end-users require only local explanations would be an incorrect simplification: in fact, users express a desire to have more influence at the local level (e.g., adjusting goals for physical activity) as well as match their own calculation with the model at the global level. In any case, explanations need to be aligned and evaluated with the goals of the user.

Furthermore, our experiment shows that subjective effects may only occur after repeated interactions. Both studies and training programs for AID-Systems should take this effect into account. However, our results imply that, e.g., other interaction possibilities could decrease this span if necessary (c.f. [27]). AID systems should therefore ask users for their predictions in the first period of AID therapy so that they can compare their own expectations with the system results with little effort. The testing of hypotheses is also a central task in order to be able to form a correct mental model of information processing. While future studies need to investigate whether interactions with a direct goal of promoting active hypothesis testing can also increase SIPA ratings or experience traceability, it is difficult to integrate this into current AID systems. Actively inducing high or low glucose levels to compare expectations with an AID system's behavior is not recommended for medical reasons. Therefore, for XAI systems to be applicable in medical contexts such as DMT1, simulations of the algorithm need to be developed, for example, that allow this testing of hypotheses before use or as counterfactual during use. Existing approaches for the simulation of glucose level (see [101]) could be supplemented with an interface that offers explanatory variants for situations selected by the users themselves.

6.5 Limitations and Further Research

Several limitations for further research have to be considered. First, the applied method to analyze performance or prediction was not as aligned with potential tasks in real-world applications as possible. That is, in AID systems users do not need to make predictions about the insulin needs calculated by the system. More importantly, they need to be able to estimate the effect of communicated information on, e.g., physical activity to cooperate effectively with the system. A more comprehensive indicator to assess the effect of traceability on the human-machine system performance could be to show a scenario and ask how changes in one or multiple attributes would affect the outcome. This would also open up different possibilities for interpretation (e.g., deviation from the correct value as in this study but also to what extent the direction of the estimate is correct as a non-metric variable). Comparable tasks exist in the area of complex problem solving [108] and could also be used in the area of human-AI interaction.

Second, in an ideal case, it would have been possible to measure the development of user experience on the course over several weeks. The time between observations, interactions, and measurements in our experiment was short compared to everyday applications. In addition, when used in one's own therapy, one's own previous experience can be included to a greater extent. A possibility for further research could be to strive to enable longitudinal designs to allow for results based on longer reflection periods as well as personalization. In addition, participants in this experiment were shown only one condition at a time, whereas patients, for example, may compare different interfaces when deciding on an AID system. As long as the influence of learning experience is taken into account, within-subject analyses of different explanatory and interaction effects could be used in further experimental settings.

Third, the present research only examined one approach to explain to the users the way an AID system calculates insulin needs. To enable users to cope, e.g., with information overload, an interactive simulation may provide counterfactual explanations for scenarios they are interested in or want to understand. Furthermore, depending on the algorithm used to construct the AID system, the concrete depiction of rules applied to calculate insulin needs could lead to important insights into the evolution of mental models in human-AI cooperation. Ideally, further studies provide different explanations to the users in order to render it possible to compare their effectiveness for different goals (i.e., understanding the effects of personalization vs. understanding one own's influence on the system through communicated information).

7 CONCLUSION

Theoretically motivated and impactful research of human-centered AI is still in an early stage of development. Empirical data of potential end-users as a target group in contrast to, e.g., developers or professionals is needed. On top of that, the relationship between subjective experiences and the impact on users' capabilities to cooperate with intelligent systems is crucial for XAI applications in the future: it determines whether explanations truly empower users or, in the worst case, overburden or even deceive them. In this sense, the present work contributes to the development of human-centered XAI on three levels:(1) by refining and applying the SIPA scale, which is derived from theoretical concepts of automation, differentiated statements about the effects of explanations can be made; (2) by developing an experimental environment to examine the interaction of potential end-users with AID XAI, the usefulness of explanations for everyday life can be validly assessed; and (3) by measuring performance at the same time as user experience, the problematic miscalibration between the perceived and actual ability to predict AI behavior can be empirically supported. Based on the empirical study, it is possible to derive design decisions that enable users of medical AI systems to collaborate and understand a system rather than overloading them with information. Future research in AID systems should therefore examine how users actively develop

and test hypotheses on AID information processing to better understand under which conditions reported SIPA ratings may exhibit a better calibration with the actual task performance.

A OVERVIEW OF ATTRIBUTES FOR AID SIMULATION

Table 11. Variables Used within the AID Simulation

Attribute	Description	Relevance
Current Tissue Glucose	The glucose level of interstitial fluid currently measured by the sensor.	It is the proxy for current blood glucose level. Needs to be in a defined range to avoid high and low blood sugar in the short term, as well as long-term problems associated with chronically high blood sugar.
Current Insulin in Body	The amount of active insulin in the body.	Lowers glucose level short term, therefore reduces the amount of insulin needed.
Current Carbohydrates in Body	The amount carbohydrates yet to be used by the body, e.g., carbohydrates in the digestive tract.	Raises glucose level (quickly or slowly depending largely on absorption rate), therefore raises the amount of insulin needed.
Current Activity	The level of physical activity of the user.	A higher activity level raises sensitivity to insulin, leads to carbohydrates being used up more quickly and thus generally lowers blood glucose, meaning it lowers the amount of insulin needed.
Tissue Glucose Target	Target amount of glucose to be measured by the sensor as proxy for blood glucose target.	Trying to reach the blood glucose target is the primary outcome of insulin therapy for T1DM. Target value may depend on current circumstances.
Avoid Hypoglycemia	Lowers risk of low blood sugar (hypoglycemia) when activated.	Automatically reduces aggressiveness and raises glucose target, therefore reduces amount of insulin given.
Duration of Insulin Action	The time in which insulin will still be active in the body.	When insulin stays active longer or has an effect, calculations need to integrate remaining effect or effect of physical activity for remaining insulin levels .
Correction Intensity	How fast the glucose target ought to be reached. Higher aggressiveness means the glucose target ought to be reached fast.	If target glucose is below current glucose reading, high aggressiveness leads to an increased amount of insulin needed. Raises risk of hypoglycemia.
Risk of Hypoglycemia in next hour	Probability of the user experiencing hypoglycemia (low blood sugar, < 3.9 mmol/l) during the next hour.	Hypoglycemia is most likely to interfere with the user's ability to function in everyday life. A high risk of hypoglycemia therefore lets the system reduce the amount of insulin that should be given to mitigate the risk.
Blood Glucose lowering per 1 Unit Insulin	How much 1 insulin unit lowers blood glucose level. High value indicates high insulin sensitivity.	The more 1 insulin unit lowers blood glucose, the less insulin is needed.
Insulin Units per 10 grams Carbohydrates	How many insulin units need to be injected to metabolize 10 grams of carbohydrates. High value indicates low insulin sensitivity.	The more insulin units are needed to metabolize 10 grams of carbohydrates, the more insulin is needed.
Predicted Exercise	System estimate on whether users expected to exercise in the next hours.	Exercise in most cases lowers blood glucose via energy consumption and increasing insulin sensitivity. Raises glucose target automatically and thus reduces the amount of insulin given in preparation for exercise.

B DESCRIPTIVE DATA FOR ALL REPEATED MEASURES VARIABLES

Table 12. Descriptive Data for All Variables Measured Repeatedly at All Points of Measurement

Block	Condition	SIPA Transparency			SIPA Understandability			SIPA Predictability			FOST			NASA-TLX		
		M	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Observation	LowID	5.08	1.31	4.50	4.35	1.46	5.00	3.48	1.21	4.00	4.23	1.31	4.40	4.85	1.11	4.00
Block 1	MedID	4.59	1.34	4.50	3.95	1.40	5.00	3.80	1.31	4.00	4.02	1.01	4.20	3.87	1.25	3.60
	HighID	5.10	0.82	3.00	3.90	1.13	4.00	3.90	0.96	4.00	4.33	0.86	3.40	3.57	1.25	4.60
Observation	LowID	4.40	1.40	4.50	4.04	1.47	5.00	3.35	1.25	4.50	3.90	1.32	4.40	4.82	1.36	5.00
Block 2	MedID	4.36	1.43	4.50	3.59	1.26	4.50	3.25	1.21	4.50	3.72	1.18	3.80	3.72	1.09	4.00
	HighID	5.04	0.79	2.50	3.83	0.97	4.00	3.77	1.07	4.50	4.02	1.04	4.20	3.67	1.42	5.20
Observation	LowID	4.23	1.28	5.00	3.69	1.24	5.00	3.52	1.16	5.00	3.88	1.29	4.40	4.53	1.48	6.00
Block 3	MedID	4.50	1.23	4.50	3.86	1.16	4.50	3.66	1.14	4.00	4.05	1.15	3.80	3.84	1.28	5.00
	HighID	5.02	0.87	3.00	3.94	1.35	4.50	3.67	1.50	5.00	4.13	1.30	5.00	3.84	1.43	5.20
Observation	LowID	4.23	1.36	5.00	3.77	1.32	5.00	3.35	1.13	4.50	3.75	1.41	4.60	4.67	1.43	5.80
Block 4	MedID	4.66	1.24	4.50	3.86	1.34	4.50	3.64	1.38	4.50	4.11	1.39	4.60	4.15	1.39	5.60
	HighID	5.08	0.75	3.00	4.00	1.53	5.00	3.83	1.52	5.00	4.29	1.26	4.60	4.00	1.48	5.20
Performance	LowID	4.08	1.69	5.00	3.42	1.59	5.00	3.06	1.36	4.00	3.86	1.57	4.60	3.97	1.14	4.20
	MedID	4.02	1.59	5.00	3.11	1.30	4.00	2.89	1.13	4.00	3.76	1.12	4.00	2.85	1.26	4.60
	HighID	5.02	0.83	2.50	3.98	1.36	5.00	3.77	1.31	5.00	4.42	0.90	3.60	3.41	1.30	5.40

C ARTIFICIAL INTELLIGENCE ATTITUDE SCALE

Table 13. All Items of the Artificial Intelligence Attitude (AIA) Scale and the Corresponding Instruction

Please indicate the degree to which you agree/disagree with the following statements.	completely disagree	largely disagree	slightly disagree	slightly agree	largely agree	completely agree
01 I feel intimidated by artificial intelligence (AI).						
02 I feel comfortable interacting with an artificial intelligence.						
03 The less contact I have with artificial intelligence, the better.						
04 I would like to work with artificial intelligence as often as possible.						
05 Artificial intelligence makes life more efficient.						
06 Artificial intelligence reduces the relevance of different professions.						

D KNOWLEDGE QUESTIONS ON DIABETES MANAGEMENT

Table 14. Knowledge Questions on Diabetes Management (Translated from German)

	Please indicate whether the following statements are correct or not.	True	False	I don't know
1	Even without eating, type 1 diabetics need insulin.			
2	When treating hypoglycemia, the most important goal is to get back to a level above 70 mg/dl as quickly as possible.			
3	When treating hyperglycemia, the most important goal is to get back to a level below 180 mg/dl as quickly as possible.			
4	If I am unsure of my insulin needs, I should inject too much rather than too little.			
5	Since alcohol consumption causes sugar levels to rise sharply, insulin should be administered particularly generously during a night of partying.			
6	How long insulin has an effect in the body depends, among other things, on the amount administered.			
7	"Rapid" insulin refers to insulin that takes effect immediately after injection without any delay.			
8	I can recognize increased insulin sensitivity by the fact that sugar levels drop more slowly after insulin is administered.			
9	FGM and CGM sensors measure blood glucose.			
10	The Dawn phenomenon describes how some diabetics are at high risk for hypoglycemia early in the morning (around 5 a.m.).			

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–18. DOI: https://doi.org/10.1145/3173574.3174156

[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14. DOI: https://doi.org/10.1145/3313831.3376615

[3] Mary B. Abraham, Martin de Bock, Grant J. Smith, Julie Dart, Janice M. Fairchild, Bruce R. King, Geoffrey R. Ambler, Fergus J. Cameron, Sybil A. McAuley, Anthony C. Keech, Alicia Jenkins, Elizabeth A. Davis, David N. O'Neal, Timothy W. Jones, Australian Juvenile Diabetes Research Fund Closed-Loop Research Group, Ace Choo, Jennifer Nicholas, Leah Laurenson, Alison Roberts, Keely Bebbingtom, Julie Klimek, Kristine Heels, Rebecca Gebert, Shaun Johnson, Stephanie Oats, Jordan Rafferty, Anthony Pease, Sophia Zoungas, Melissa H. Lee, Barbora Paldus, Catriona M. Sims, Richard J. MacIssac, Glenn M. Ward, Peter G. Colman, Neale D. Cohen, Leon Bach, Kavita Kumareswaran, Stephen N. Stranks, Morton G. Burt, Jane D. Holmes-Walker, Roland W. McCallum, Joey Kaye, Jane Speight, Christel Hendreickx, Andrzej Januszewski, Adreinne Kirby, and Sara Vogrin. 2021. Effect of a hybrid closed-loop system on glycemic and psychosocial outcomes in children and adolescents with type 1 diabetes: A randomized clinical trial. *JAMA Pediatrics* 175, 12 (Dec. 2021), 1227. DOI: https://doi.org/10.1001/jamapediatrics.2021.3965

[4] Rebecca N. Adams, Molly L. Tanenbaum, Sarah J. Hanes, Jodie M. Ambrosino, Trang T. Ly, David M. Maahs, Diana Naranjo, Natalie Walders-Abramson, Stuart A. Weinzimer, Bruce A. Buckingham, and Korey K. Hood. 2018. Psychosocial and human factors during a trial of a hybrid closed loop system for type 1 diabetes management. *Diabetes Technology & Therapeutics* 20, 10 (Oct. 2018), 648–653. DOI: https://doi.org/10.1089/dia.2018.0174

[5] Muhammad Aljukhadar, Sylvain Senecal, and Charles-Etienne Daoust. 2010. Information overload and usage of recommendations. In *Proceedings of the ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI'10)*. CEUR Workshop Proceedings, Aachen, 26–33.

[6] Ahlam Alotaibi, Reem Al Khalifah, and Karen McAssey. 2020. The efficacy and safety of insulin pump therapy with predictive low glucose suspend feature in decreasing hypoglycemia in children with type 1 diabetes mellitus: A systematic review and meta-analysis. *Pediatric Diabetes* 21, 7 (Nov. 2020), 1256–1267. DOI: https://doi.org/10.1111/pedi.13088

[7] Yasmeen Alufaisan, Laura Ranee Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. 2020. *Does Explainable Artificial Intelligence Improve Human Decision-making?* Preprint. PsyArXiv. DOI: https://doi.org/10.31234/osf.io/d4r9t

[8] Christiane Attig, Daniel Wessel, and Thomas Franke. 2017. Assessing personality differences in human-technology interaction: An overview of key self-report scales to predict successful interaction. In *HCI International 2017 – Posters' Extended Abstracts*, Constantine Stephanidis (Ed.). Springer International Publishing, 19–29.

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. DOI: https://doi.org/10.1609/hcomp.v7i1.5285

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 1–16. DOI: https://doi.org/10.1145/3411764.3445717

[11] M. Baumann and J. F. Krems. 2007. *Situation Awareness and Driving: A Cognitive Model*. Springer London, London, 253–265. DOI: https://doi.org/10.1007/978-1-84628-618-6_14

[12] Pierre Yves Benhamou, Erik Huneker, Sylvia Franc, Maeva Doron, and Guillaume Charpentier. 2018. Customization of home closed-loop insulin delivery in adult patients with type 1 diabetes, assisted with structured remote monitoring: The pilot WP7 Diabeloop study. *Acta Diabetologica* 55, 6 (June 2018), 549–556. DOI: https://doi.org/10.1007/s00592-018-1123-1

[13] Yoav Benjamini and Henry Braun. 2002. John W. Tukey's contributions to multiple comparisons. *Annals of Statistics* 30, 6 (2002), 1576–1594.

[14] Cari Berget, Halis Kaan Akturk, Laurel H. Messer, Timothy Vigers, Laura Pyle, Janet Snell-Bergeon, Kimberly A. Driscoll, and Gregory P. Forlenza. 2021. Real-world performance of hybrid closed loop in youth, young adults, adults and older adults with type 1 diabetes: Identifying a clinical target for hybrid closed-loop use. *Diabetes, Obesity and Metabolism* 23, 9 (Sept. 2021), 2048–2057. DOI: https://doi.org/10.1111/dom.14441

[15] Cari Berget, Laurel H. Messer, Tim Vigers, Brigitte I. Frohnert, Laura Pyle, R. Paul Wadwa, Kimberly A. Driscoll, and Gregory P. Forlenza. 2020. Six months of hybrid closed loop in the real-world: An evaluation of children and young adults using the 670G system. *Pediatric Diabetes* 21, 2 (March 2020), 310–318. DOI: https://doi.org/10.1111/pedi.12962

[16] Ruth Beyth-Marom and Baruch Fischhoff. 1983. Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology* 45, 6 (1983), 1185–1195. DOI: https://doi.org/10.1037/0022-3514.45.6.1185

[17] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 401–413. DOI: https://doi.org/10.1145/3461702.3462571

[18] Alessandro Bisio, Linda Gonder-Frederick, Ryan McFadden, Daniel Cherñavvsky, Mary Voelmle, Michael Pajewski, Pearl Yu, Heather Bonner, and Sue A. Brown. 2022. The impact of a recently approved automated insulin delivery system on glycemic, sleep, and psychosocial outcomes in older adults with type 1 diabetes: A pilot study. *Journal of Diabetes Science and Technology* 16, 3 (May 2022), 663–669. DOI: https://doi.org/10.1177/1932296820986879

[19] Emanuele Bosi, Pratik Choudhary, Harold W. de Valk, Sandrine Lablanche, Javier Castañeda, Simona de Portu, Julien Da Silva, Roseline Ré, Linda Vorrink-de Groot, John Shin, Francine R. Kaufman, Ohad Cohen, Andrea Laurenzi, Amelia Caretto, David Slatterly, Marcia Henderson-Wilson, S. John Weisnagel, Marie-Christine Dubé, Valérie-Ève Julien, Roberto Trevisan, Giuseppe Lepore, Rosalia Bellante, Irene Hramiak, Tamara Spaic, Marsha Driscoll, Sophie Borot, Annie Clergeot, Lamia Khiat, Peter Hammond, Sutapa Ray, Laura Dinning, Giancarlo Tonolo, Alberto Manconi, Maura Serena Ledda, Wendela de Ranitz, Bianca Silvius, Anne Wojtusciszyn, Anne Farret, Titia Vriesendorp, Folkje Immeker-de Jong, Joke van der Linden, Huguette S. Brink, Marije Alkemade, Pauline Schaepelynck-Belicar, Sébastien Galie, Clémence Tréglia, Pierre-Yves Benhamou, Myriam Haddouche, Roel Hoogma, Lalantha Leelarathna, Angel Shaju, and Linda James. 2019. Efficacy and safety of suspend-before-low insulin pump technology in hypoglycaemia-prone adults with type 1 diabetes (SMILE): An open-label randomised controlled trial. *Lancet Diabetes & Endocrinology* 7, 6 (June 2019), 462–472. DOI: https://doi.org/10.1016/S2213-8587(19)30150-0

[20] Charlotte K. Boughton. 2021. Fully closed-loop insulin delivery—Are we nearly there yet?*Lancet Digital Health* 3, 11 (Nov. 2021), e689–e690. DOI: https://doi.org/10.1016/S2589-7500(21)00218-1

[21] Charlotte K. Boughton, Sara Hartnell, Janet M. Allen, Julia Fuchs, and Roman Hovorka. 2022. Training and support for hybrid closed-loop therapy. *Journal of Diabetes Science and Technology* 16, 1 (Jan. 2022), 218–223. DOI: https://doi.org/10.1177/1932296820955168

[22] Frank A. Buckless and Sue Pickard Ravenscroft. 1990. Contrast coding: A refinement of ANOVA in behavioral analysis. *Accounting Review* 65, 4 (1990), 933–945.

[23] Ruth M. J. Byrne. 2019. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 6276–6282. DOI: https://doi.org/10.24963/ijcai.2019/876

[24] John T. Cacioppo and Richard E. Petty. 1982. The need for cognition.*Journal of Personality and Social Psychology* 42, 1 (Jan. 1982), 116–131. DOI: https://doi.org/10.1037/0022-3514.42.1.116

[25] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. 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*. ACM, 1–14. DOI: https://doi.org/10.1145/3290605.3300234

[26] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. 2021. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The emerging landscape of interpretable agent behavior. *Proceedings of the International Conference on Automated Planning and Scheduling* 29, 1 (May 2021), 86–96. DOI: https://doi.org/10.1609/icaps.v29i1.3463

[27] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. 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*. ACM, 1–12. DOI: https://doi.org/10.1145/3290605.3300789

[28] Erin K. Chiou and John D. Lee. 2023. Trusting automation: Designing for responsivity and resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 65, 1 (Feb. 2023), 137–165. DOI: https://doi.org/10.1177/00187208211009995

[29] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark patterns of explainability, transparency, and user control for intelligent systems. *IUI Workshops*.

[30] Jacob Cohen. 1992. Statistical power analysis. *Current Directions in Psychological Science* 1, 3 (1992), 98–101.

[31] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. GANSlider: How users control generative models for images using multiple sliders with and without feedforward information. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI'22)*. Association for Computing Machinery, New York, NY, Article 569, 15 pages. DOI: https://doi.org/10.1145/3491102.3502141

[32] Mustafa Demir, Nathan J. McNeese, and Nancy J. Cooke. 2019. The evolution of human-autonomy teams in remotely piloted aircraft systems operations. *Frontiers in Communication* 4 (Sept. 2019), 50. DOI: https://doi.org/10.3389/fcomm.2019.00050

[33] L. Dowling, E. G. Wilmot, and P. Choudhary. 2020. Do-it-yourself closed-loop systems for people living with type 1 diabetes. *Diabetic Medicine* 37, 12 (Dec. 2020), 1977–1980. DOI: https://doi.org/10.1111/dme.14321

[34] Jeff Druce, James Niehaus, Vanessa Moody, David Jensen, and Michael L. Littman. 2021. Brittle AI, Causal Confusion, and Bad Mental Models: Challenges and Successes in the XAI Program. DOI: https://doi.org/10.48550/ARXIV.2106.05506

[35] Upon Ehsan, Samir Passi, Qingzi Vera Liao, Larry Chan, I-Hsiang Lee, Michael J. Muller, and Mark O. Riedl. 2021. *The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations*. ArXiv abs/2107.1350.

[36] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*.

[37] M. R. Endsley. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*. IEEE, 789–795. DOI: https://doi.org/10.1109/NAECON.1988.195097

[38] Mica R. Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (March 1995), 32–64. DOI: https://doi.org/10.1518/001872095779049543

[39] Mica R. Endsley and David B. Kaber. 1999. Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics* 42, 3 (March 1999), 462–492. DOI: https://doi.org/10.1080/001401399185595

[40] C. Farrington. 2018. Psychosocial impacts of hybrid closed-loop systems in the management of diabetes: A review. *Diabetic Medicine* 35, 4 (April 2018), 436–449. DOI: https://doi.org/10.1111/dme.13567

[41] Caspar Goeke, Holger Finger, Dorena Diekamp, and Peter König. 2017. Introducing, testing, and evaluating a unified javascript framework for professional online studies. *World Academy of Science, Engineering and Technology, International Journal of Psychological and Behavioral Sciences* 4 (2017).
[42] Erin D. Foster and Ariel Deardorff. 2017. Open Science Framework (OSF). DOI: https://doi.org/10.5195/jmla.2017.88
[43] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human– Computer Interaction* 35, 6 (April 2019), 456–467. DOI: https://doi.org/10.1080/10447318.2018.1456150
[44] David C. Funder and Daniel J. Ozer. 2019. Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science* 2, 2 (June 2019), 156–168. DOI: https://doi.org/10.1177/2515245919847202
[45] Jose Garcia-Tirado, John P. Corbett, Dimitri Boiroux, John Bagterp Jørgensen, and Marc D. Breton. 2019. Closed-loop control with unannounced exercise for adults with type 1 diabetes using the ensemble model predictive control. *Journal of Process Control* 80 (Aug. 2019), 202–210. DOI: https://doi.org/10.1016/j.jprocont.2019.05.017
[46] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (July 2020), 627–660. DOI: https://doi.org/10.5465/annals.2018.0057
[47] Jamie C. Gorman, Nancy J. Cooke, and Jennifer L. Winner. 2006. Measuring team situation awareness in decentralized command and control environments. *Ergonomics* 49, 12–13 (Oct. 2006), 1312–1325. DOI: https://doi.org/10.1080/00140130600612788
[48] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. 2019. Guidelines for reinforcement learning in healthcare. *Nature Medicine* 25, 1 (Jan. 2019), 16–18. DOI: https://doi.org/10.1038/s41591-018-0310-5
[49] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Human Mental Workload*.North-Holland, Oxford, England, 139–183. DOI: https://doi.org/10.1016/S0166-4115(08)62386-9
[50] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (May 2015), 407–434. DOI: https://doi.org/10.1177/0018720814547570
[51] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. arxiv:1812.04608 [cs]
[52] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 164–168. DOI: https://doi.org/10.1145/2856767.2856811
[53] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.
[54] Andreas Holzinger, André Carrington, and Heimo Müller. 2019. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. arxiv:1912.09024 [cs]
[55] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. 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. CEUR-WS, Aachen, 63–72.
[56] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. 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*. ACM, 624–635. DOI: https://doi.org/10.1145/3442188.3445923
[57] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (March 2000), 53–71. DOI: https://doi.org/10.1207/S15327566IJCE0401_04
[58] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna Van Riemsdijk, and Maarten Sierhuis. 2014. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-robot Interaction* 3, 1 (March 2014), 43. DOI: https://doi.org/10.5898/JHRI.3.1.Johnson
[59] P. N. Johnson-Laird. 1986. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press.
[60] Mary Joyce and Jurek Kirakowski. 2013. Development of a general internet attitude scale. In *Design, User Experience, and Usability. Design Philosophy, Methods, and Tools*, Aaron Marcus (Ed.). Springer, Berlin, 303–311.
[61] Barbara Kimbell, David Rankin, Nicole L. Ashcroft, Lidiya Varghese, Janet M. Allen, Charlotte K. Boughton, Fiona Campbell, Atrayee Ghatak, Tabitha Randell, Rachel E. J. Besser, Nicola Trevelyan, Roman Hovorka, Julia Lawton, on Behalf of the CLOuD Consortium. 2020. What training, support, and resourcing do health professionals need to support people using a closed-loop system? A qualitative interview study with health professionals involved in the closed loop from onset in type 1 diabetes (CLOuD) trial. *Diabetes Technology & Therapeutics* 22, 6 (June 2020), 468–475. DOI: https://doi.org/10.1089/dia.2019.0466

[62] G. Klein, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich. 2004. Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems* 19, 6 (Nov. 2004), 91–95. DOI: https://doi.org/10.1109/MIS.2004.74

[63] Rex B. Kline. 2004. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research.*American Psychological Association, Washington, DC. xii, 325 pages. DOI: https://doi.org/10.1037/10693-000

[64] Christine Knoll, Sofia Peacock, Mandy Wäldchen, Drew Cooper, Simran Kaur Aulakh, Klemens Raile, Sufyan Hussain, and Katarina Braune. 2022. Real-world evidence on clinical outcomes of people with type 1 diabetes using open-source and commercial automated insulin dosing systems: A systematic review. *Diabetic Medicine* 39, 5 (2022), e14741. DOI: https://doi.org/10.1111/dme.14741 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/dme.14741

[65] Boris Kovatchev, Peiyao Cheng, Stacey M. Anderson, Jordan E. Pinsker, Federico Boscari, Bruce A. Buckingham, Francis J. Doyle, Korey K. Hood, Sue A. Brown, Marc D. Breton, Daniel Chernavvsky, Wendy C. Bevier, Paige K. Bradley, Daniela Bruttomesso, Simone Del Favero, Roberta Calore, Claudio Cobelli, Angelo Avogaro, Trang T. Ly, Satya Shanmugham, Eyal Dassau, Craig Kollman, John W. Lum, Roy W. Beck, and for the Control to Range Study Group. 2017. Feasibility of long-term closed-loop control: A multicenter 6-month trial of 24/7 automated insulin delivery. *Diabetes Technology & Therapeutics* 19, 1 (Jan. 2017), 18–24. DOI: https://doi.org/10.1089/dia.2016.0333

[66] Joshua A. Kroll. 2021. Outlining traceability: A principle for operationalizing accountability in computing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 758–771. DOI: https://doi.org/10.1145/3442188.3445937

[67] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 3–10. DOI: https://doi.org/10.1109/VLHCC.2013.6645235

[68] B. Kulzer, L. Heinemann, and Timm Roos. 2021. Informationstechnologie in der Diabetesbehandlung – Erleben der Patienten. *Der Diabetologe* 17, 3 (May 2021), 265–274. DOI: https://doi.org/10.1007/s11428-021-00753-9

[69] Richard E. Ladner. 2015. Design for user empowerment. *Interactions* 22, 2 (Feb. 2015), 24–29. DOI: https://doi.org/10.1145/2723869

[70] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. The LRP toolbox for artificial neural networks. *Journal of Machine Learning Research* 17, 114 (2016), 1–5. http://jmlr.org/papers/v17/15-618.html.

[71] David M. Levy. 2008. *Information Overload*. John Wiley & Sons, Ltd., Chapter 20, 497–515. DOI: https://doi.org/10.1002/9780470281819.ch20 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470281819.ch20

[72] Dana Lewis. 2021. How it started, how it is going: The future of artificial pancreas systems (automated insulin delivery systems). *Journal of Diabetes Science and Technology* 15, 6 (Nov. 2021), 1258–1261. DOI: https://doi.org/10.1177/19322968211027558

[73] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–15. DOI: https://doi.org/10.1145/3313831.3376590

[74] L. M. E. Lindner, W. Rathmann, and J. Rosenbauer. 2018. Inequalities in glycaemic control, hypoglycaemia and diabetic ketoacidosis according to socio-economic status and area-level deprivation in type 1 diabetes mellitus: A systematic review. *Diabetic Medicine* 35, 1 (Jan. 2018), 12–32. DOI: https://doi.org/10.1111/dme.13519

[75] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (Jan. 2020), 56–67. DOI: https://doi.org/10.1038/s42256-019-0138-9

[76] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th Australasian Conference on Information Systems*, Vol. 53. ACM, New York, NY, 6–8.

[77] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* 113 (2021), 103655.

[78] John M. McGuirl and Nadine B. Sarter. 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48, 4 (Dec. 2006), 656–665. DOI: https://doi.org/10.1518/001872006779166334

[79] Stephanie M. Merritt, Deborah Lee, Jennifer L. Unnerstall, and Kelli Huber. 2015. Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 1 (Feb. 2015), 34–47. DOI: https://doi.org/10.1177/0018720814561675

[80] Laurel H. Messer, Cari Berget, Tim Vigers, Laura Pyle, Cristy Geno, R. Paul Wadwa, Kimberly A. Driscoll, and Gregory P. Forlenza. 2020. Real world hybrid closed-loop discontinuation: Predictors and perceptions of youth

discontinuing the 670g system in the first 6 months. *Pediatric Diabetes* 21, 2 (March 2020), 319–327. DOI: https://doi.org/10.1111/pedi.12971

[81] Laurel H. Messer, Gregory P. Forlenza, Jennifer L. Sherr, R. Paul Wadwa, Bruce A. Buckingham, Stuart A. Weinzimer, David M. Maahs, and Robert H. Slover. 2018. Optimizing hybrid closed-loop therapy in adolescents and emerging adults using the MiniMed 670G system. *Diabetes Care* 41, 4 (April 2018), 789–796. DOI: https://doi.org/10.2337/dc17-1682

[82] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. DOI: https://doi.org/10.1016/j.artint.2018.07.007

[83] Majid Mobasseri, Masoud Shirmohammadi, Tarlan Amiri, Nafiseh Vahed, and Hossein Hosseini Fard. 2020. Prevalence and incidence of type 1 diabetes in the world: A systematic review and meta-analysis. *Pediatric Diabetes* 10, 2 (2020), 18.

[84] Mobeen Nazar, Muhammad Mansoor Alam, Eiad Yafi, and Mazliham Mohd Su'ud. 2021. A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. *IEEE Access* 9 (2021), 153316–153348. DOI: https://doi.org/10.1109/ACCESS.2021.3127881

[85] Minh Nguyen, Ivana Jankovic, Laurynas Kalesinskas, Michael Baiocchi, and Jonathan H. Chen. 2021. Machine learning for initial insulin estimation in hospitalized patients. *Journal of the American Medical Informatics Association* 28, 10 (Sept. 2021), 2212–2219. DOI: https://doi.org/10.1093/jamia/ocab099

[86] Mahsan Nourani, Chiradeep Roy, Jeremy E. Block, Donald R. Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *26th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, 340–350.

[87] Emil Øversveen. 2020. Stratified users and technologies of empowerment: Theorising social inequalities in the use and perception of diabetes self-management technologies. *Sociology of Health & Illness* 42, 4 (May 2020), 862–876. DOI: https://doi.org/10.1111/1467-9566.13066

[88] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, 2 (June 1997), 230–253. DOI: https://doi.org/10.1518/001872097778543886

[89] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 (May 2000), 286–297. DOI: https://doi.org/10.1109/3468.844354

[90] John Patrick and Philip L. Morgan. 2010. Approaches to understanding, analysing and developing situation awareness. *Theoretical Issues in Ergonomics Science* 11, 1–2 (Jan. 2010), 41–57. DOI: https://doi.org/10.1080/14639220903009946

[91] Margus Pedaste, Mario Mäeots, Leo A. Siiman, Ton De Jong, Siswa A. N. Van Riesen, Ellen T. Kamp, Constantinos C. Manoli, Zacharias C. Zacharia, and Eleftheria Tsourlidaki. 2015. Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review* 14 (2015), 47–61.

[92] Peter Pesl, Pau Herrero, Monika Reddy, Maria Xenou, Nick Oliver, Desmond Johnston, Christofer Toumazou, and Pantelis Georgiou. 2016. An advanced bolus calculator for type 1 diabetes: System architecture and usability results. *IEEE Journal of Biomedical and Health Informatics* 20, 1 (Jan. 2016), 11–17. DOI: https://doi.org/10.1109/JBHI.2015.2464088

[93] Barbara Piccini, Emilio Casalini, Chiara Macucci, and Sonia Toni. 2022. Type 1 diabetes technology management traps in a pediatric patient: Not all that glitters is gold. *Acta Diabetologica* 59, 1 (Jan. 2022), 137–141. DOI: https://doi.org/10.1007/s00592-021-01781-z

[94] Arun Rai. 2020. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science* 48, 1 (Jan. 2020), 137–141. DOI: https://doi.org/10.1007/s11747-019-00710-5

[95] David Rankin, Barbara Kimbell, Janet M. Allen, Rachel E. J. Besser, Charlotte K. Boughton, Fiona Campbell, Daniela Elleri, Julia Fuchs, Atrayee Ghatak, Tabitha Randell, Ajay Thankamony, Nicola Trevelyan, Malgorzata E. Wilinska, Roman Hovorka, and Julia Lawton. 2021. Adolescents' experiences of using a smartphone application hosting a closed-loop algorithm to manage type 1 diabetes in everyday life: Qualitative study. *Journal of Diabetes Science and Technology* 15, 5 (Sept. 2021), 1042–1051. DOI: https://doi.org/10.1177/1932296821994201

[96] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 10 pages. arxiv:1602.04938 [cs, stat]

[97] Lysanne Rivard, Pascale Lehoux, and Hassane Alami. 2021. "It's not just hacking for the sake of it": A qualitative study of health innovators' views on patient-driven open innovations, quality and safety. *BMJ Quality & Safety* 30, 9 (Sept. 2021), 731–738. DOI: https://doi.org/10.1136/bmjqs-2020-011254

[98] Ralph L. Rosnow, Robert Rosenthal, and Donald B. Rubin. 2000. Contrasts and correlations in effect-size estimation. *Psychological Science* 11, 6 (Nov. 2000), 446–453. DOI: https://doi.org/10.1111/1467-9280.00287

[99] William B. Rouse and Nancy M. Morris. 1986. On looking into the black box: Prospects and limits in the search for mental models.*Psychological Bulletin* 100, 3 (1986), 349.

[100] Swati Sachan, Jian-Bo Yang, Dong-Ling Xu, David Eraso Benavides, and Yang Li. 2020. An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications* 144 (April 2020), 113100. DOI: https://doi.org/10.1016/j.eswa.2019.113100

[101] Jana Schmitzer, Carolin Strobel, Ronald Blechschmidt, Adrian Tappe, and Heiko Peuscher. 2022. Efficient closed loop simulation of do-it-yourself artificial pancreas systems. *Journal of Diabetes Science and Technology* 16, 1 (2022), 61–69.

[102] Tim Schrills, Mourad Zoubir, Mona Bickel, Susanne Kargl, and Thomas Franke. 2021. Are users in the loop? Development of the subjective information processing awareness scale to assess XAI. *Proceedings of the ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI.*

[103] Peter Sedlmeier and Frank Renkewitz. 2018. *Forschungsmethoden und Statistik in der Psychologie (nachdr. ed.).* Pearson Studium, München.

[104] Mark Segal. 2004. Machine Learning benchmarks and random forest regression. *UCSF: Center for Bioinformatics and Molecular Biostatistics.* Retrieved from https://escholarship.org/uc/item/35x3v9t4.

[105] Clare Shaban. 2015. Psychological themes that influence self-management of type 1 diabetes. *World Journal of Diabetes* 6, 4 (2015), 621. DOI: https://doi.org/10.4239/wjd.v6.i4.621

[106] Joseph P. Shivers, Linda Mackowiak, Henry Anhalt, and Howard Zisser. 2013. "Turn it off!": Diabetes device alarm fatigue considerations for the present and the future. *Journal of Diabetes Science and Technology* 7, 3 (May 2013), 789–794. DOI: https://doi.org/10.1177/193229681300700324

[107] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human– Computer Interaction* 36, 6 (April 2020), 495–504. DOI: https://doi.org/10.1080/10447318.2020.1741118

[108] Valerie J. Shute, Lubin Wang, Samuel Greiff, Weinan Zhao, and Gregory Moore. 2016. Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior* 63 (Oct. 2016), 106–117. DOI: https://doi.org/10.1016/j.chb.2016.05.047

[109] Madison B. Smith, Anastasia Albanese-O'Neill, Tamara G. R. Macieira, Yingwei Yao, Joseph M. Abbatematteo, Debra Lyon, Diana J. Wilkie, Michael J. Haller, and Gail M. Keenan. 2019. Human factors associated with continuous glucose monitor use in patients with diabetes: A systematic review. *Diabetes Technology & Therapeutics* 21, 10 (Oct. 2019), 589–601. DOI: https://doi.org/10.1089/dia.2019.0136

[110] Aaron Springer and Steve Whittaker. 2018. "I Had a Solid Theory before but It's Falling Apart": Polarizing Effects of Algorithmic Transparency. arxiv:1811.02163 [cs]

[111] Aaron Springer and Steve Whittaker. 2020. Progressive disclosure: When, why, and how do users want algorithmic transparency information?*ACM Transactions on Interactive Intelligent Systems* 10, 4 (Dec. 2020), 1–32. DOI: https://doi.org/10.1145/3374218

[112] Jackie Sturt, Kathryn Dennick, Mette Due-Christensen, and Kate McCarthy. 2015. The detection and management of diabetes distress in people with type 1 diabetes. *Current Diabetes Reports* 15, 11 (Nov. 2015), 101. DOI: https://doi.org/10.1007/s11892-015-0660-z

[113] Sakinah Suttiratana, Jessie J. Wong, Monica S. Lanning, Adrienne Dunlap, Sarah Hanes, Korey K. Hood, Rayhan Lal, and Diana Naranjo. 2021. 518-P: User experiences with loop, an open-source automated insulin delivery (AID) system. *Diabetes* 70, Supplement_1 (June 2021), 518–P. DOI: https://doi.org/10.2337/db21-518-P

[114] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: The effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces.* ACM, 109–119. DOI: https://doi.org/10.1145/3397481.3450662

[115] Richard Taylor. 1989. Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Proceedings of the AGARD AMP Symposium on Situational Awareness in Aerospace Operations, CP478. Seuilly-sur Seine: NATO AGARD.*

[116] Sara Trevitt, Sue Simpson, and Annette Wood. 2016. Artificial pancreas device systems for the closed-loop control of type 1 diabetes: What systems are in development? *Journal of Diabetes Science and Technology* 10, 3 (May 2016), 714–723. DOI: https://doi.org/10.1177/1932296815617968

[117] Daniel Trommler, Christiane Attig, and Thomas Franke. 2018. Trust in activity tracker measurement and its link to user acceptance. In *Mensch und Computer 2018 - Tagungsband. Bonn: Gesellschaft für Informatik e.V.*, R. Dachselt and G. Weber (Hrsg.) (Ed.). DOI: 10.18420/muc2018-mci-0361

[118] Silvia Tulli, Filipa Correia, Samuel Mascarenhas, Samuel Gomes, Francisco S. Melo, and Ana Paiva. 2019. Effects of agents' transparency on teamwork. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Davide Calvaresi, Amro Najjar, Michael Schumacher, and Kary Främling (Eds.). Springer International Publishing, 22–37.

[119] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, 1–15. DOI: https://doi.org/10.1145/3290605.3300831

[120] Kathryn W. Weaver and Irl B. Hirsch. 2018. The hybrid closed-loop system: Evolution and practical applications. *Diabetes Technology & Therapeutics* 20, S2 (June 2018), S2–16–S2–23. DOI: https://doi.org/10.1089/dia.2018.0091

[121] Bert Weijters and Hans Baumgartner. 2012. Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research* 49, 5 (Oct. 2012), 737–747. DOI: https://doi.org/10.1509/jmr.11.0368

[122] Christopher D. Wickens and C. Melody Carswell. 2006. Information processing. In *Handbook of Human Factors and Ergonomics*, Gavriel Salvendy (Ed.). John Wiley & Sons, Inc., Hoboken, NJ, 111–149. DOI: https://doi.org/10.1002/0470048204.ch5

[123] Stefan Wiens and Mats E. Nilsson. 2017. Performing contrast analysis in factorial designs: From NHST to confidence intervals and beyond. *Educational and Psychological Measurement* 77, 4 (Aug. 2017), 690–715. DOI: https://doi.org/10.1177/0013164416668950

[124] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, 307–317.

[125] Tao Zhang, David Kaber, and Maryam Zahabi. 2022. Using situation awareness measures to characterize mental models in an inductive reasoning task. *Theoretical Issues in Ergonomics Science* 23, 1 (Jan. 2022), 80–103. DOI: https://doi.org/10.1080/1463922X.2021.1885083

[126] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 295–305. DOI: https://doi.org/10.1145/3351095.3372852

[127] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (March 2021), 593. DOI: https://doi.org/10.3390/electronics10050593

Received 21 February 2022; revised 21 October 2022; accepted 17 February 2023

6 Study 3: Assessing the influence of Instructions & Trust in AI-supported pattern recognition

6.1 Summary of Study 3

This study examines how measuring trust influences users' confidence in AI systems in pattern recognition tasks. Specifically, it examines the "question-behavior effect," where the act of evaluating a variable-in this case, trust-can change participants' behavior. The research uses Kandinsky Patterns as an experimental paradigm, where participants are asked to identify deviations from a given rule, both with and without AI assistance. The study uses a 2x2 experimental design, varying the timing (early vs. late) and scope (single-item vs. multi-item) of trust assessments. The results show that trust assessments do not significantly affect behavior. Instead, the study highlights the role of communicated AI reliability, finding it to be a better predictor of user trust than self-reported trust measures.

6.2 Relevance within the dissertation

This study is a central contribution of the dissertation, as it deepens the understanding of how user interactions with AI systems can be shaped by factors such as perceived system reliability, i.e., cues for users to develop strategies for using AI systems. By focusing on the question-behavior effect, the presented research

challenges common assumptions in the human-AI interaction literature and adds depth to the dissertation's investigation of automation-related user experiences. It supports the broader discussion of how trust and transparency in AI systems influence user behavior, particularly in diagnostic and decision support systems.

6.3 Contribution to Study 3

I was primarily responsible for the conceptualization of the study, including the experimental design, the development of the Kandinsky Pattern paradigm, and the identification of variables to manipulate trust ratings. I played a key role in translating the idea of using Kandinsky Patterns into the research context of human-AI interaction, effectively bridging AI evaluation with empirical evaluation methods. I also conducted the data collection and analysis to test the hypotheses. In addition, I worked on the entire manuscript, including the introduction, methodology, results, and discussion sections, where I integrated theoretical models of trust and user reliance in AI with practical implications for future research in human-AI interaction.

Questioning Trust in AI Research: Exploring the Influence of Trust Assessment on Dependence in AI-assisted Decision-Making

Tim Schrills[a,**], Thomas Franke[a,*], Steffen Hoesterey[b,*], Eileen Roesler[c,*]

[a] *University of Luebeck, Ratzeburger Allee 160, Luebeck, 23560, Germany*
[b] *Technische Universität Berlin, Straße des 17. Juni 135, Berlin, 10623, Germany*
[c] *George-Mason University, 4400 University Drive, Fairfax, 22030, Virginia, USA*

Abstract

Trust is considered crucial for effective interaction between humans and artificial intelligence (AI), necessitating valid trust assessment methods. The 'question-behavior effect,' however, suggests that by applying a questionnaire subsequent behavior can be influenced, for example participants' dependence on AI. The objective of the present research was to examine the effect of trust assessment on reliance in the context of an AI-supported decision-making task. We designed an AI-supported task, requiring participants to decide on patterns in so-called Kandinsky Figures. In a scripted experiment with a 2x2 between-subjects design, $N = 149$ participants' trust was assessed at different times (before block 1 or block 2) and with different assessment extent (i.e., scale length). Participants' agreement with AI recommendations and task completion time served as behavioral trust indicators. We found no effect of trust assessment on behavior and correlations between trust and dependence were notably low. Participants' dependence matched the instructed reliability level of the AI system and our findings did not suggest the presence of a question-behavior effect of trust assessment. Overall, while the conduction of trust assessment did not influence dependence, our results question the conceptualization of trust as a general predictor for dependence, especially

*These authors contributed equally to this work.
**Corresponding Author.
Email addresses: tim.schrills@uni-luebeck.de (Tim Schrills), thomas.franke@uni-luebeck.de (Thomas Franke), hoesterey@tu-berlin.de (Steffen Hoesterey), eroesle@gmu.edu (Eileen Roesler)

in comparison to instructed reliability.

Keywords: Human-AI Interaction, Explainable Artificial Intelligence, Trust in AI, Trust Assessment

1. Introduction

Systems based on artificial intelligence (AI) affect the world we live in, as they analyze data, suggest decisions, or make predictions [1]. Thereby, AI systems can influence an individual's life crucially, for example, in medical therapy contexts [2], by predicting liability for loans [3] or serving as the basis for governmental decisions [4]. Accordingly, calls for reliable and safe AI have been numerous (e.g., [5]) and the first policies for AI regulation have been positioned (see [6]). However, determining the actual reliability of an AI system can be a challenging task for institutions and individuals, due to the opacity of systems as well as the lack of skills or resources on the part of the user ([7]). In such situations, research has shown that individuals approach the system by assuming certain reliability [8]. If this assumption is incorrect, errors in AI systems could lead to increased compliance by the individual [9]. Hence, the interaction of AI and humans leads to challenges for human-system trust, well known from human-automation research: without knowing how good the AI may perform, users are vulnerable when their expectations are violated. In other words, given the uncertainty and vulnerability in interactions with AI trust might be a central component shaping how individuals behave (see [10]). Research in human-AI interaction (HAII) focuses on understanding how AI system design can enable individuals to easily assess the extent to which they can trust a system. To understand human trust in AI, trust measurement needs to be reliable and validated.

In HAII research, the conceptualization of trust towards AI systems often refers to existing work on trust in automation [10, 11, 12], as AI represents automated information processing with varying levels of task allocation between AI and human users [13]. Trust is commonly defined as an attitude, and a direct assessment thus requires a subjective rating, such as in the form of self-reported trust [14]. Oppositely, research also aims to use indicators such as behavioral [15] or physiological [16] signals, for example, because they can be sampled more frequently and are less intrusive. Moreover, previous studies have repeatedly shown, that the relation between self-reported trust, reliability of systems, or dependence can be low [17, 18]. Hence, further re-

search on the conditions that influence the relationship between self-reported trust and behavior, is needed (c.f. [19, 20]).

In addition to the obvious lack of predictive power, another disadvantage of self-assessments in trust research is that they can be intrusive and alter the participants' response in an experiment. The observation that a questionnaire can change people's behavior is described as the question-behavior effect (QBE). Several studies have shown that measuring an attitude can influence behavior [21, 22, 23]. In particular, the measurement of trust could trigger introspection that affects people's behavior [24]. Accordingly, the exact timing of the trust measurement during the experiment could be a key factor in explaining the different levels of correlation between self-reported trust assessment and trust-related behavior (see [25]). Not only the occurrence of a measurement, but also its scope (i.e., the number of items) could influence subsequent behavior. Repeated reflection on multiple items compared to, for example, a single item could influence a person's attitude, comparable to the effect of mere repeated exposure [26]. As far as we are aware, there are no studies on the measurement effects of self-reported trust assessment on trust-related behavior. However, such results are needed to ensure internal validity and to clarify existing questions in the interpretation of experimental studies (see [27]). Since trust is not subsequently assessed in real-world contexts, it would be problematic if experiments systematically lost their ecological validity as a result.

Accordingly, the objective of the present research was to examine the QBE of self-reported trust assessment on dependence in the context of an AI-supported task. To this end, we designed an abstract pattern recognition task based on current paradigms in AI research [28] and varied at what point in time and to what extent trust was surveyed from the participants. We aimed to investigate the presence of a question behavior effect of trust on dependence, and explore how characteristics such as the timing and length of the measurement may impact this effect.

2. Background

2.1. Trust as an Attitude

Whereas trust is commonly described via different concepts such as beliefs, attitudes intentions, or behavior in the literature [11], there is the overarching consensus in HAII that trust should be conceptualized as an attitude rather than a behavior (especially based on [29]). As an attitude, trust is

an inner tendency of how to evaluate an entity, and therefore trust is not directly observable. This is closely mirrored in the definition from Lee & See (2004, p.51), defining trust as an "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Also, as reflected in multiple models of trust [10, 11, 29], existing HAII research treats trust as a multidimensional concept. However, in various studies in HAII research trust is currently assessed as an unidimensional attitude that can be utilized to predict user behavior such as and especially dependence [30, 31]. The relevance of trust as a construct for HAII [27, 32] and the fact that trust as an attitude can only be assessed through self-reported assessments is currently discussed [33].

Self-report methods of assessing trust are the most common way of assessing trust in research on HAII [14]. Self-report measures of trust mainly consist of self-rated scales, with Jian et al.'s "Trust in Automation" scale being the most cited [34]. However, while there are several models of trust in automation (see e.g., [35, 29]) and scales based on these models (e.g., [36]), most are self-constructed and custom scales that are usually not explicitly validated or grounded on theoretical models [14, 37]. In addition, applied research often aims to achieve economic assessment methods and therefore, relies on single-item scales. However, to improve the methodological reliability of self-reported trust assessment, several studies assess trust with longer scales, for example, [34, 36]. Furthermore, existing self-report measures of trust address heterogeneous constructs: they position trust as an attitude toward a trustee [36, 34], as a perception of a system property such as trustworthiness [38] or reliability [39]. Due to the complexity and lack of standardization in the assessment of trust, as well as the high number of individual assessment methods in the literature, the labeling of trust scales as "measures" of trust could be criticized, as the term measure in its strict meaning is thought to refer to standardized methods that define the extent of a property and are sufficiently validated [40, 37]. That is, while many research papers use the term "trust measure" we refer to the term "trust assessment" in the present paper. Additionally, note that we will use the term "dependence" for behavior related to following system recommendations in line with [41].

Due to the challenges of their psychometric quality, the substitution of self-reported trust assessment with behavioral methods could support the standardization and comparability of trust research. However, while existing models link, for example, dependence to trust [29], dependence can be seen

as a possible behavioral consequence of trust (comparable to, e.g., indirect or implicit methods to assess attitudes, see [42]). Dependence can be analyzed by indicators of consistency, performance, behavior (such as verification), or response bias [43], which compose methods where the actions of a participant are observed.

In addition, while existing research approaches the assessment of trust through physiological methods, it remains unclear what psychological variables these measures truly capture [14]. Other influences such as risk or workload might affect physiological reactions and behavior, thus, overshadowing effects of participants' trust. Moreover, physiological measures can be be expensive (e.g., [16] utilizing eye-tracking) and may not be suitable for field research. Accordingly, self-report measures of trust are likely to remain the primary method for assessing trust. However, research is needed on the relationship between different trust assessment methods and their implications for experimental design.

2.2. The Relationship between Trust & Dependence

The relationship between trust and dependence as a behavior influenced by trust, is discussed among multiple technologies and domains [44, 45, 46]. Exemplary, studies cited by Lee & See (2004) on trust and dependence are based on correlation analysis (see, e.g., [47, 20]). For example, when the self-reported trust assessment takes place after participants could choose to rely on a system or not rely on it, the self-reported trust assessment could only reflect how they perceived their behavior. That is, self-reported trust assessment would be highly correlated with dependence, but could not be used as a predictor (see [17, 19]). While several empirical studies on trust and dependence assume a predictive value of trust assessment for behavior [31, 48, 49], other studies of dependence in automation demonstrate results that indicate a more complex connection, questioning the predictive value of trust assessments [50, 51, 52, 20]. The reason for that may be additional factors that influence behavior but potentially do not affect trust in the same way, for example, workload or the extent of current risk. In addition, the quantity of studies that combine both, self-reported trust assessments as well as behavioral methods to quantify dependence is scarce, especially compared to studies that rely only on self-reported trust assessment [14]. However, studies integrating both, dependence and self-reported assessments, usually try not to understand the predictive value of trust for dependence. In order to do so, those studies are forced to assess self-reported trust before dependence

is measured. The effect of a self-reported trust assessment on dependence, however, has not been studied sufficiently to guarantee the internal validity of the results found in those studies.

In previous studies, changes in dependence as a reaction to AI errors were dependent on the type of error the system made [53, 25]. Accordingly, the predictive value of trust assessment for dependence could be affected by the experimental sequence. [25] modified, whether automated aid in a decision task was shown before or after participants made their own decision. They demonstrated that assessed trust may be more influential on dependence when participants have to make an initial decision, but not when the decision aid is presented first. In any case, the sequence of assessing trust and measuring dependence may affect study results. The experimental conditions under which trust assessment and dependence interact have not been sufficiently studied yet. Hence, sequential assessment effects could be the reason for contradictory or at least inconsistent results.

2.3. Assessing Attitudes may Affect Behavior

Potential effects of assessing trust on trust-related behavior can be explained by different theoretical approaches: On the one hand, previous research, e.g., involving the mere-measurement effect [54], suggests that measuring an attitude can increase behavior consistent with that attitude [55]. It is suggested that this effect is due to the fact that reflecting on the attitude makes it more cognitively accessible, thereby increasing the likelihood of related behavior (e.g., purchase decisions, [56]). Consequently, asking a person about their trust in a system might make them more likely to engage in behavior that demonstrate trust.

While it is possible that the mere-measurement effect also occurs in trust-related behavior, its measurement in the context of HAII is difficult to demonstrate: trust as an attitude is characterized by more dynamics than, for example, the attitude towards blood donation or purchasing decisions, which were associated with the mere-measurement effect in earlier research. In empirical research on trust in AI, the attitude towards certain systems or robots is investigated, which is, however, interaction-dependent and less stable. As a result, a temporally separate measurement is rarely found in research on trust in AI. While in experiments the asessment of trust can also be recorded at the beginning or end, in the mere measurement study the attitude is sometimes measured weeks in advance. In other words, while attitudes towards fundamental decisions such as organ donations are relatively stable, thus,

less dependent on the current situation (e.g., a task or a performance that has just been shown) than the construct of trust.

In the literature on the mere measurement effect (which unfortunately uses other terms such as QBE synonymously in some cases, see [57]), explanatory approaches are given that can also be applied to trust. [58] describe that thinking about behavior may activate cognitive mechanisms such as processing fluency and behavioral simulation. That is, the mental representation of a behavior in which one controls a system or manages the precision of a system may make one more likely to engage in that behavior. For example, behavior associated with trusting or controlling a system may be cognitively activated and thus more likely to be performed, regardless of whether they are consciously intended. Previous research suggests that this cognitive activation increases related behavior in general by making it more accessible. The latter described effects are less specific than the mere measurement effect, as trust-related behavior is not dependent on the individuals' attitude (and its valence) but is generally more accessible. However, the predictions made by the question-behavior effect are less precise than the mere-measurement effect, which focuses specifically on the alignment of attitudes and behaviors [54]. Consequently, the QBE might be more suitable for a more dynamic construct like trust.

Based on previous literature on experimental effects of attitude assessment, we suggest that the introspection, triggered by the evaluation process (i.e., a questionnaire), may slow response times and reduce confidence in decisions, as suggested by [24] and aim to investigate it in an experimental study. Here, it is crucial to investigate whether trust assessments are the primary driver of behavioral change or whether other factors, such as perceived system reliability, play a more important role.

3. Present Research

The objective of the present research was to evaluate the presence of an experimental artifact in human-AI trust research, i.e., effect of trust assessment on dependence in the context of human-AI interaction. Based on previous findings in the literature, we argue that a general question-behavior effect could systemically affect user behavior in experimental studies.

Our first research question focuses on the presence of the question-behavior effect. In particular, we expected higher dependence for the control group (no trust assessment) compared to the experimental groups at both early

and late assessment (H1). Moreover, we aimed to investigate if the extent of assessment influences dependence. We expected lower dependence after a multi-item assessment compared to a single-item assessment (H2). Lastly, we were interested in the influence of time of assessment on dependence. We assumed that participants demonstrated lower dependence overall when trust was assessed at an early point in time of the experiment compared to an assessment at a later time (H3).

In addition, to understand what other variables may influence human dependence, we included a regression analysis of perceived reliability, instructed reliability, and trust assessment. As experience with the system could also affect confidence, we examined the development of confidence over time, independent of the experimental condition. The present research was pre-registered (removed for anonymization).

During the review process, we realized that we could also try to test the mere measurement effect described above, at least to some extent, on the basis of our data. Therefore, we conducted an exploratory investigation to see whether we could better interpret the behavioral effects found if we included the level of the attitude (trust). To do this, we had to assume that trust in the groups without trust assessment is distributed similarly to the group with trust assessment.

Ethics approval for this study was granted by the Ethics Committee of the Local University before the start of the experiment.

4. Method

4.1. Participants

149 participants were recruited via the online platform prolific [59]. Based on our preregistered exclusion criteria (i.e., completion time shorter or longer than median participation time +/- 2 standard deviation, lack of variability in responses, and correct responses below 20% in task block 1 and task block 2) no participant needed to be excluded. Participants were between 19 - 79 years ($Mean$ = 43.9, SD = 14.4). 98 participants identified themselves as female, 59 as male, and one participant preferred not to answer. All participants were residents of the United Kingdom. Additionally, the affinity for technology interaction (ATI) of the participants was assessed (M = 3.24, SD = 1.04). Participants volunteered to participate in the study, and informed consent was obtained. Only native English speakers were recruited for the study. The study was conducted with LimeSurvey [60]. Participants were

instructed to conduct the study only with appropriate screen sizes, i.e., on desktop computers, laptops, or tablets. Participants were compensated £3.50 for their time in the study. In addition, every participant with a performance over 90% was awarded an additional £2. This performance-related reward was offered to incentivize and motivate participants beyond their general compensation.

4.2. Task & Material

In the experiment, we utilized Kandinsky Patterns, which are sets comprised of individual images called Kandinsky Figures (KF). Kandinsky Figures were introduced by [28] and describe images consisting of squares containing 1 to n geometric objects. A set of Kandinsky Figures that follows the same rule (e.g., there are more red than green objects) can be described as a Kandinsky Pattern. Following [28] our implementation of Kandinsky Pattern can be described as follows: each Kandinsky Figure is a square image that contains 1 to n objects defined by their sizes, colors, shapes, and positions in this image. A statement about a KF can be denoted in natural language, linguistically, or mathematically and can either be true or false. A Kandinsky Pattern is defined as a subset of all KF for which a certain set of statements, which we call rules, apply.

We developed a unity-based application that allows for the systematic creation of images of KF (KFgen app). The app can be found and used under https://github.com/JonasJakobi/Kandinsky to create stimuli like the ones used in the present experiment. In the KFgen app, the experimenter can specify which rules should apply to all generated KF. In the present research and version of the KFgen app three possible types of rules can be created: Rules relating to the relative positioning of objects, for example, "Red objects above green objects", rules restricting the maximum amount of a certain type of object, for example, "There can only be two squares", and rules that force a minimum amount of an object, for example, "There have to be more than two circles." KFgen produces a set of KF, randomizing each KF while maintaining the rules. For the rules to stay deducible, each KF will contain at least one of each relevant object (i.e., all objects referred to in the rules). So with a rule like "Squares above triangles", the generated KFs will always contain at least one square and triangle.

In the KFgen app rules can also be inverted. This allows for the generation of a KF that does not follow the same rules as the other KFs and therefore does not belong to the corresponding Kandinsky Pattern. Instead

of just not following the rules, these unfitting KFs will instead explicitly break the rules by also containing each relevant object to break a rule.

We used Kandinsky Patterns as the experimental context for our research. Kandinsky Tasks (KT) are based on the described KFs and were designed as follows: a total of five KFs were displayed. Four of them followed a Kandinsky Pattern, i.e., an underlying rule. A fifth picture explicitly violated this rule. For example, four out of five depicted KFs only contained red objects above green objects. A fifth picture, in contrast, contained red objects under green objects and was therefore violating the rule "All objects are above green objects". These five pictures were displayed horizontally next to each other in a random order. The pictures were labeled a) - e) as shown in Figure 1. The participants' task in the KT was to select the picture that violated the rule. In our experiment, rules were not change for 10 subsequent tasks, i.e., the same rule applied for multiple tasks after each other (similar to the game Mastermind, [61]). Participants did not have the opportunity to revisit previous rules or decisions. Whenever a rule was changed, the previous rule was revealed to participants.

For the trials with AI support, an additional text cue was generated. This read "The AI recommends x", where x was the letter of the picture that violated this rule. In all groups, the AI's reliability was 100% throughout all experimental trials, i.e., the textual cue was always correct.

KT provide a great basis to research HAII, as they canpotentially be AI-assisted, as well as solved by humans. Furthermore, KT are sufficiently abstract that participants cannot benefit from pre-existing knowledge, compared to, for example, studies in the medical domain. To understand the effects of human dependence, the task can be performed correctly, i.e., there is an objectively correct answer, that allows the experimenter to determine performance.

In addition, stimuli are also well suited for investigating the effects of trust assessment on subsequent behavior: although the task is challenging, participants can check for themselves whether they want to comply with the system's suggestions, as they have access to all the information relevant to the task. In order to do this, however, considerably more time must be used in some cases, as all images must be checked in one go. Hence, there is therefore an interest on the part of the participants in complying with the AI and saving resources, but this is not absolutely necessary. In addition, it can always be clearly determined whether the people have followed the system's suggestion or not.

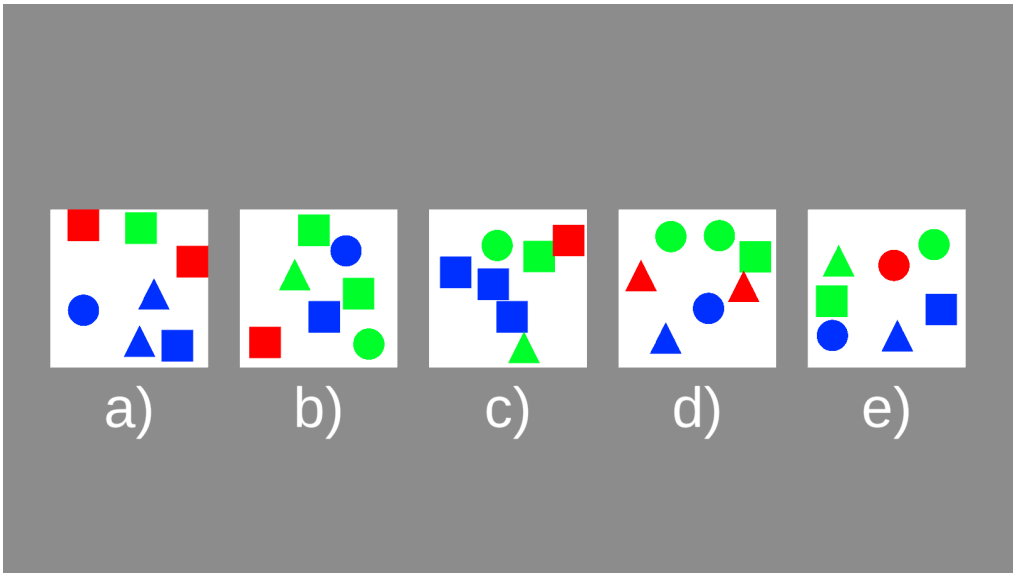


Figure 1: Depiction of a set of five Kandinsky figures. Four of them follow the rule that every red object has to be above blue objects (a Kandinsky Pattern). However, the second Kandinsky Figure, labeled b) does not follow that pattern. Images like these were used to assess participants' baseline performance.

4.3. Pre-Study

We conducted a pre-study with $N = 30$ participants recruited via prolific. Participants were between 22 - 67 years ($M = 37.7$, $SD = 11.3$). 17 identified themselves as female, and 13 as male. All participants were residents of the United Kingdom ($n = 29$) or the United States ($n = 1$). In particular, the task's difficulty was tested. Within the pre-study, the average percentage of correctly answered Kandinsky Tasks of the participants without AI support was 62% and with AI support close to 90%. Within the pre-study, the AI's reliability was stated as 90%, whereby it was 100% correct. As the study in the pre-study lasted over 1.5 hours in some cases, the second block was shortened by 10 stimuli in the main study. The discrepancy between instructed reliability and the system's true reliability was increased, to support users to recognize the difference despite the rather short overall period of experience with the AI system. Accordingly, the system was introduced with only 80% reliability in the main study. In addition, based on an effect size of $\eta^2 = 0.038$ found between task completion time between groups in the first task block and aiming for 80% power, we calculated a required sample size of 150

participants for the main study (based on G*power, [62]).

4.4. Study Design

The study followed a 2x2 design, with the time of trust assessment (early vs. late; i.e., trust assessment prior to block 1 or prior to block 2) and extent of trust assessment (i.e. single-item or multi-item scale) as manipulated variables, and a control group as depicted in Figure 2. Accordingly, each participant was assigned to one of five conditions: control condition (Con), early trust assessment, single item (E-S), early trust assessment, multi-item (E-M), late trust assessment, single item (L-S), or late trust assessment, multi-item (L-M).

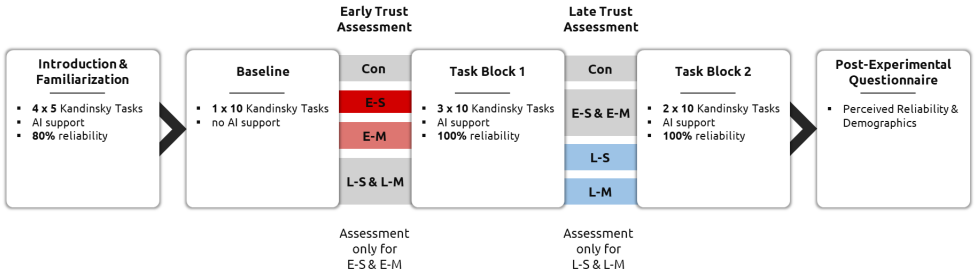


Figure 2: Depiction of study design, highlighting the different points in time for trust assessment.

4.5. Dependent Variables

Trust. In the multi-item conditions, trust was assessed using the Trust in Automation (TiA) scale from Jian et al. (2000) as it is one of the most frequently used scales for self-reported trust assessment [14]. Accordingly, for the single item condition, we used the following item: "How much do you trust the system?" rated on a six-point Likert scale from 1 (not at all) to 6 (completely), to compare it to the TiA scale as done in previous research (e.g., [7]). We decided to use a six-point Likert scale in both cases instead of a scale ranging from 0 to 100 (single item as used in [7]) or 1 to 7 (multi item as originally introduced by [34]) compared to previous research, to achieve higher consistency between the response scales used in our study.

Dependence. We measured dependence based on two variables 1) agreement between users and AI and 2) task completion time. First, the extent to which people followed the AI recommendation for their answers was calculated. One commonly used measure to quantify the agreement of two raters is Cohen's Kappa coefficient [63]. Cohen's Kappa compares the observed agreement between the AI recommendation and user responses with the expected agreement that would occur by chance alone. It takes into account both the proportion of agreement and the possibility of agreement by random chance. The possible values of Cohen's Kappa range from -1 to 1. A positive value indicates agreement beyond chance, where 1 indicates perfect agreement, and 0 indicates agreement equal to chance. For consistency, we report Cohen's Kappa in percent. Second, task completion time was computed by capturing how long each participant spent in a block divided by the number of presented stimuli in each block (i.e., the average time for one Kandinsky task, i.e. five pictures to compare, in each block).

4.6. Procedure

After obtaining informed consent, KF and tasks were explained. Participants were then asked to familiarize themselves with the task. They were presented with 20 Kandinsky tasks, with the Kandinsky pattern changing after every fifth task. The rule was shown to the users and they only had to decide whether the AI recommendation shown was correct or not. The AI system's reliability was introduced to the participants at 80%. In order to have a perfect fit between description and experience the AI's reliability in this introductory section was also scripted to be 80% [64]. Afterwards, during the Baseline block the participant's ability to correctly identify the unfitting image without AI support was assessed. They were presented with ten Kandinsky tasks following the same rule (red objects must be below blue objects). The rule was revealed to participants after they had completed all ten Kandinsky tasks. After the baseline phase before the actual experiment started, a trust assessment was conducted for both groups in the early condition.

In the following first experimental block, participants were presented with 30 Kandinsky tasks. The rules on which the KFs were based changed after ten tasks, i.e., three different rules were presented in this block. The order of the rules presented was randomized within the block. Participants received feedback on each rule after completing ten tasks, thus, they were able to understand whether the AI made correct recommendations or not. Next, a

trust assessment was administered to all groups assigned to the late trust assessment conditions. The same procedure as in block 1 was repeated, but only with 20 Kandinsky tasks and two different rules. Again, the order of the rules presented was randomized and the rules were revealed after the tasks had been completed. After completing block 1 and block 2, participants were asked to estimate the AI's reliability in percent as an indicator of perceived reliability. Variables to assess interpersonal differences were also administered at this point. Finally, participants were given a code to receive compensation.

5. Results

5.1. Descriptive & baseline values

We calculated the baseline values of participants' ability in the Kandinsky task that were not supported by AI for dependence and task completion time. We then calculated the Cohen's Kappa coefficients ($M = 62.43$, $SD = 39.45$). Afterwards, we calculated the task completion time ($M = 244.94$, $SD = 139.74$) using the time the participants spent on the task for the baseline measure. The descriptive data for all groups divided into the two blocks is depicted in Table 1.

Group		Recommendation Agreement				Task Completion Time				Trust			
		Block 1		Block 2		Block 1		Block 2		Block 1		Block 2	
	n	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Con.	23	75.54	24.23	79.53	19.57	205.26	148.90	167.57	111.26				
L-S	32	76.79	16.38	84.05	17.09	251.01	127.60	168.68	88.61			4.19	0.79
L-M	35	76.80	19.93	79.85	15.85	191.94	92.84	181.73	230.66			3.62	0.75
E-S	24	74.43	21.60	78.43	23.20	203.08	127.21	120.95	60.46	3.67	1.15		
E-M	35	74.82	19.55	83.09	20.01	192.97	117.83	128.64	107.65	4.01	0.68		

Note. $N = 149$.

Table 1: Descriptive statistics of dependence assessed through Cohenś kappa coefficient, time on task measured in seconds, and trust for each condition and measurement block. Con = Control, LS = Late Assessment, Single Item, LM = Late Assessment, Multi Item, ES = Early Assessment, Single Item, EM = Early Assessment, Multi Item

5.2. The effect of trust assessment on dependence (H1)

To analyze the effect of trust assessment on dependence, we compared dependence between all groups in a one-factor ANOVA for each task block, separately.

Recommendation Agreement. For recommendation agreement in task block 1, we found no difference between control (Con; M = 75.54, SD = 24.23), single item (L-S & E-S; M = 75.78, SD = 18.65) and multi-item (L-M & E-M; M = 75.81, SD = 19.63) groups ($F(2, 146) = 0.002$, $p = .998$). For the model the calculated η^2 value of 0.00002 indicates no effect according to Cohen.

In parallel, for task block 2, we did not find significant difference in dependence measurement between control (Con; M = 79.53, SD = 19.57), single item (L-S & E-S; M = 81.64, SD = 19.94) and multi-item (L-M & E-M; M = 81.47, SD = 17.99) groups ($F(2, 146) = 0.11$, $p = .895$). For the model the calculated η^2 value of 0.002 indicates a very small effect.

Task completion time. For the time on task comparison between the groups in task block 1, we found no difference between the groups (Con; M = 205.26, SD = 148.90, L-S & E-S; M = 230.47, SD = 128.52, L-M & E-M; M = 192.46, SD = 105.30) as well ($F(2, 146) = 1.53$, $p = .220$). Here, the generalized η^2 value of 0.02 would show a small effect.

For the task completion time between groups (Con; M = 167.57, SD = 111.26, L-S & E-S; M = 148.22, SD = 80.76, L-M & E-M; M = 155.19, SD = 180.67), we did not find a significant difference between the groups regarding the time needed to complete the tasks in block 2 ($F(2, 146) = 0.16$, p = .856). Here, the generalized η^2 value of 0.002 showed a very small effect.

Based on the results of the ANOVAs for the dependence measure in both blocks, there is no sufficient evidence to support H1, stating that there is a difference in dependence after trust assessment has been carried out. The results are depicted in Figure 3.

5.3. The effect of trust assessment extent on dependence (H2)

Recommendation Agreement. We compared the dependence of all groups that were assessed with single-item (L-S & E-S; M = 78.37, SD = 17.40) and all groups that were assessed with a multi-item (L-M & E-M; M = 80.03, SD = 16.18) questionnaire. The recommendation agreement was calculated based on the responses of both task blocks. Due to the larger sample size we conducted a t-test even though the normality assumption was violated [65]. Our results indicate that the observed difference ($t(105)$ = 0.52, $p = .699$) is not statistically significant. The effect size (Cohen's d = -0.1) indicates a small effect according to ([66]). Accordingly, our results did not support the assumption of any difference in recommendation agreement

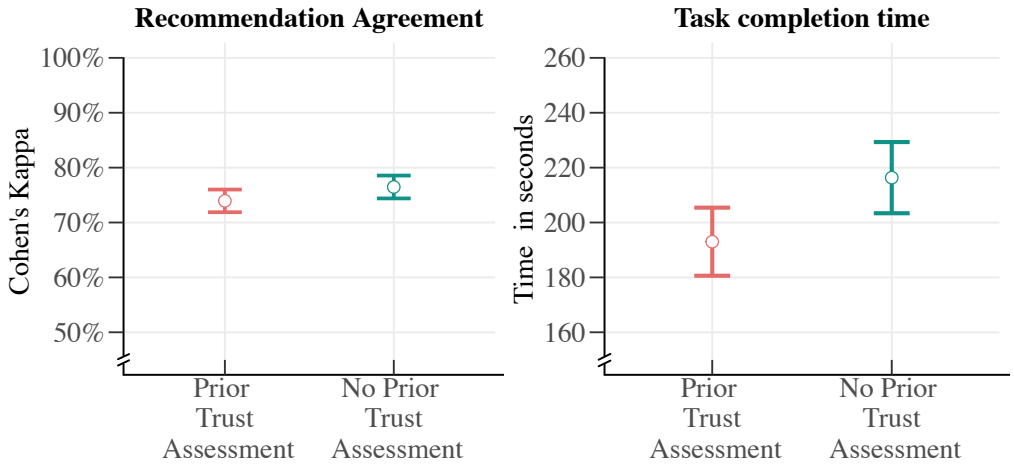


Figure 3: Comparison of user behaviour means and standard errors for the early and late trust measurement groups of the first measurement block. Task dependence is shown through the average Cohen's kappa coefficients. The time needed to complete the tasks is measured in seconds.

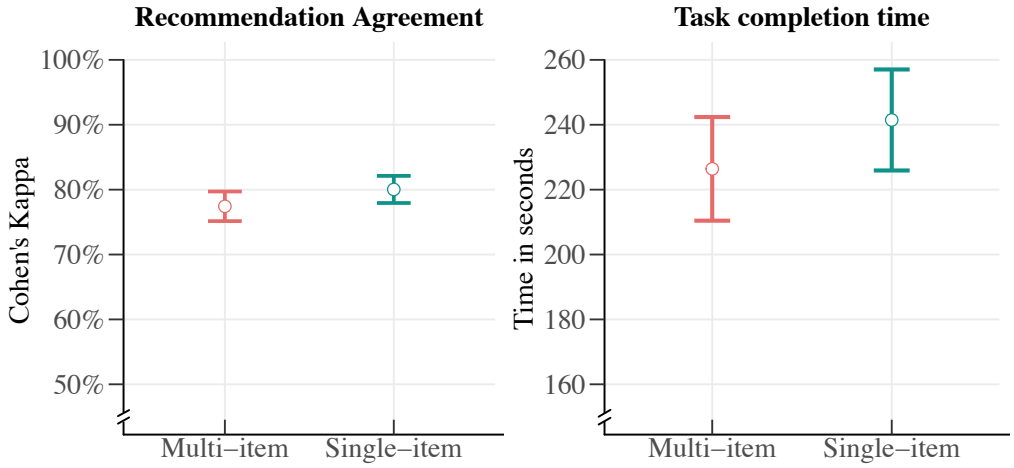


Figure 4: Comparison of user dependence measures with corresponding standard errors for multi- and single-item trust assessment groups for blocks 1 and 2 through the average Cohen's kappa coefficients for task dependence as well as the time needed to complete the tasks in seconds.

between multi-item and single-item assessments. The results are depicted in Figure 4.

Task completion time. We did not find a significant difference in task completion time between single- (L-S & E-S; M = 246.92, SD = 122.32) and multi-item (L-M & E-M; M = 226.41, SD = 123.74) groups ($t(108)$ = -0.88, p = .810). Here, Cohen's d = 0.17 also indicates a small effect. Accordingly, based on our results H2 that the "multi-item measure" has significantly higher task completion time than the "single-item measure" cannot be supported.

5.4. The effect of timing of trust assessment on dependence (H3)

To assess whether the time of trust assessment affects user behavior we differentiated between all groups that were assessed in Block 1 and all groups that were assessed in Block 2. The dependence for the blocks following immediately after the trust assessment were compared.

Recommendation Agreement. To analyze the effect of time on recommendation agreement, we compared the Cohenś kappa values of single- and multi-item measures of the first block (E-S & E-M; M = 74.66; SD = 20.23) with the Cohenś kappa values of single- and multi-item measures of the second block (L-S & L-M; M = 81.85; SD = 20.23). The results indicate significantly lower dependence in the early assessment block than in the late assessment block ($t(112)$ = -2.17, p = .032, Cohens' d = -0.39) and do therefore support H3.

Task completion time. We then compared the time needed by the participants of single- and multi-item measure groups of the first block (E-S & E-M; M = 197.08; SD = 120.75) with the participants time of the single- and multi-item groups of the second block (L-S & L-M; M = 175.50; SD = 176.47). For this comparison, the results show, that there is no significant difference between the task completion time of the blocks ($t(117)$ = 0.81, p = .420, Cohens' d = 0.14), and do not support H3.

5.5. Exploratory analysis

5.5.1. Dependence development over time

To understand, whether effects in dependence were based on experience over time, we additionally compared the dependence measures between the first (M = 75.76, SD = 19.90) and second (M = 81.23, SD = 18.87) block. Hence, Cohen's Kappa coefficient for both blocks was calculated and used for analysis. The development over all conditions is depicted in Figure 5 (and

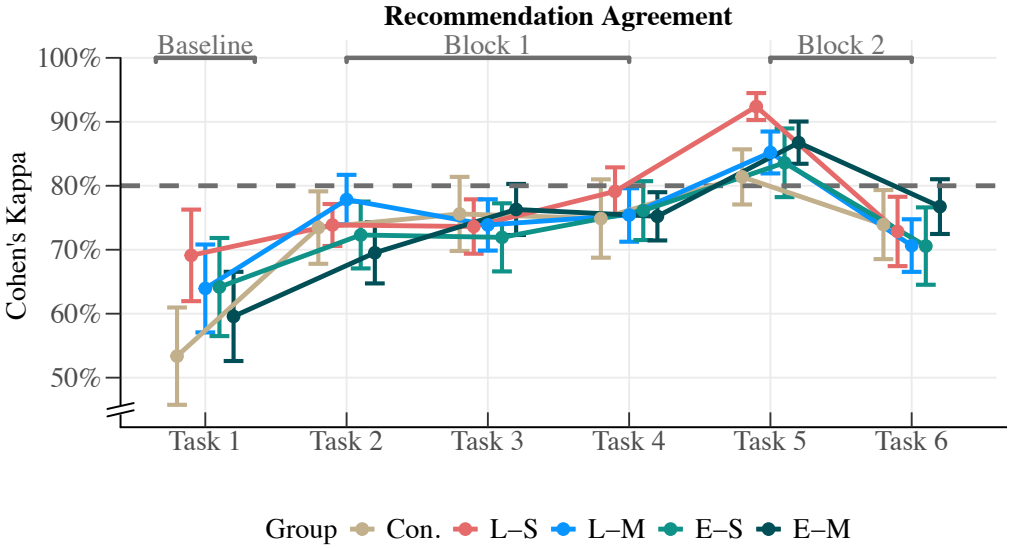


Figure 5: Comparison of recommendation agreement represented by the average Cohen's kappa coefficients and the standard error of the mean for each measurement block and task presented in the chronological sequence in which the tasks were given. Additionally, the instructed reliability of the AI system of 80% is marked.

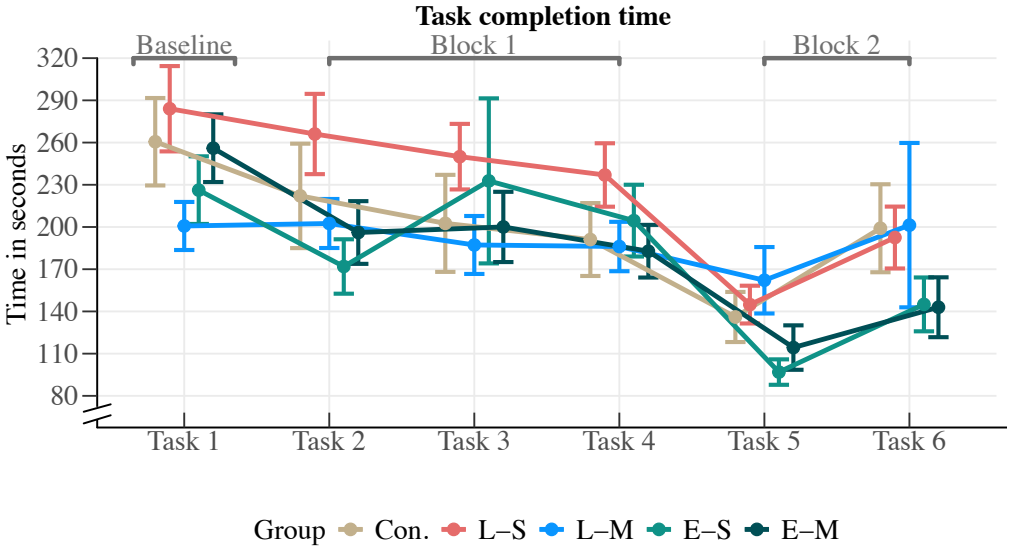


Figure 6: Comparison of task completion time for each measurement block and task presented in the chronological sequence in which the tasks were given.

Model	R^2	AIC	BIC
Instructed Reliability	-	-64.44	-61.61
Perceived reliability	0.00002	-62.45	-56.77
Trust	0.00122	-67.29	-61.85
Propensity to trust	0.00190	-63.75	-58.09
ATI	0.00080	-63.61	-57.95

Table 2: Akaike information criterion and Bayesian information criterion comparison of regression models for the prediction of recommendation agreement.

for Time on Task in Figure 6, respectively). We found a significant difference in dependence between the first and the second block ($t(148)$ = -4.77, $p <$ .001). Here, the t-test shows a small effect size of Cohen's $d = 0.28$ regarding the comparison of dependence between the first and second block.

5.5.2. Relationship between dependence and perceived reliability

In order to understand possible other factors for dependence demonstrated by participants, we also estimated, to what extent the perceived reliability (M = 73.09, SD = 20.01) of the AI system was connected to recommendation agreement (M = 78.16, SD = 18.14) as an indicator of dependence. The Pearson correlation coefficient for the estimated AI performance and dependence of the participants shows a moderate positive correlation (r = 0.37, $p <$ .001).

5.5.3. Prediction of dependence based on perceived reliability and trust assessment

Multiple linear regression was used to test whether perceived reliability assessed through the estimated AI performance (E-S, E-M, L-S, L-M; M = 73.63, SD = 19.46) and both single- and multi-item trust measurements (E-S, E-M, L-S, L-M; M = 3.89, SD = 0.86) with a fixed intercept given by the instructed reliability of 0.8 significantly predict recommendation agreement determined through the Cohen's kappa coefficient (E-S, E-M, L-S, L-M; M = 78.49, SD = 18.60). The results show that the overall regression was not significant ($R^2 = 0.00808$, $F(2, 110) = 0.45$, $p = .640$). We can therefore conclude that predicting the recommendation agreement cannot be significantly improved by integrating perceived reliability and trust assessment.

5.5.4. Prediction of dependence based on individual differences

Another multiple linear regression model was calculated to test whether the propensity to trust (E-S, E-M, L-S, L-M; $M = 3.68$, $SD = 0.69$) and affinity for technology interaction (E-S, E-M, L-S, L-M; $M = 3.32$, $SD = 1.04$) predict recommendation agreement assessed through Cohen's kappa coefficients (E-S, E-M, L-S, L-M; $M = 78.49$, $SD = 18.60$). The intercept was again given by the instructed reliability and set to the constant value of 0.8. The results of the model indicate again that the overall regression was not significant ($R^2 = 0.00405$, $F(2, 123) = 0.25$, $p = .779$).

5.5.5. Direct comparison of trust assessment

Measurement extent. To explore potential effects of assessment extent on trust values, the trust assessment of the single- (L-S & E-S; $M = 3.98$, $SD = 0.98$) and multi-item (L-M & E-M; $M = 3.82$, $SD = 0.73$) groups of both blocks were compared using the Welch two sample t-Test. The results showed, that there is no significant difference between the trust values of single- and multi-item assessment groups ($t(94) = 0.99$, $p = .324$, Cohens' $d = 0.19$).

Measurement timing. Furthermore, we compared the trust measures of the early (E-S & E-M; $M = 3.87$, $SD = 0.91$) and late (L-S & L-M; $M = 3.91$, $SD = 0.82$) measurement groups of the first block, to analyze whether the measurement time point has an impact on trust values. Again the results indicate, that the observed difference between the trust values of early and late assessment is not significant. ($t(101) = -0.26$, $p = .793$, Cohens' $d = -0.05$).

5.5.6. Correlation coefficients between dependence and trust assessments

Furthermore, the correlation coefficients between dependence and trust values were calculated for the general comparison of single- and multi-item measures. With a Pearson correlation coefficient of $r = 0.45$ and $p < .001$ for the single-item measure and a $r = 0.42$ and $p < .001$ for the multi-item measure, the analysis shows a moderate positive correlation between dependence and trust values. Additionally, the Pearson correlation coefficient between the trust assessments of the multi- and single-item scales was calculated.

5.5.7. Multi-Level-Model based analysis of the mere-measurement effect

We conducted an exploratory analysis to test the level of trust combined with the impact of trust assessment. Specifically, we examined whether in-

cluding the assessment of trust as another predictor to the level of trust could better explain participants recommendation agreement and task completion time. We assumed that trust values of participants for whom trust was not assessed are distributed similarly to the values of participants with trust assessment. Thus, we first simulated trust data for the control condition based on the distribution of trust scores within the experimental condition. The average trust score (M = 3.89) and its standard deviation (SD = 0.86) were then used to generate normally distributed trust data for measurements without trust assessment.

For both recommendation agreement and task completion time, a null model including only the fixed intercept and two baseline models including variables for the grouping context were created to explore the random effects (i.e. the variability due to grouping contexts) structure of the data and to assess the appropriateness of a multi-level approach. The first baseline model considered only participants as the grouping context, which already accounts for a considerable amount of variance based on the intraclass correlation coefficient (ICC) as a measure of variance between groups (Recommendation Agreement: ICC = .71, Task Completion Time: ICC = .58) while the second model additionally included the measurement block as a second grouping factor (Recommendation Agreement: ICC = .74, Task Completion Time: ICC = .74). To compare the goodness of fit and determine whether adding complexity to the model improves its accuracy, we used likelihood ratio tests. The results indicate that the baseline model considering participants as a grouping context is significantly better compared to a model without any random effects for both the recommendation agreement and task completion time (Recommendation Agreement: $\chi^2(3) = 102.45, p < .001$, Task Completion Time: $\chi^2(3) = 60.94, p < .001$. However, in both cases, considering participants and measurement blocks shows a further improvement in explanatory power (Recommendation Agreement: $\chi^2(4) = 15.98, p < .001$, Task Completion Time: $\chi^2(4) = 63.14, p < .001$). Thus, we can assume that the multi-level modeling approach with the more complex model can be considered appropriate in both cases.

To analyze recommendation agreement and task completion time Linear Mixed-effect Models were used to account for the nested structure of repeated measures and measurement blocks. Analyses were conducted in R using the packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) to fit the multi-level regression models. Satterthwaite's approximation method (Luke, 2017) for adjusted estimates of the degrees of freedom was used for

the fixed effects. To allow for model comparisons through LRT, the models were fitted using maximum likelihood.

For a simple model only considering trust, we found a significant effect of trust level on recommendation agreement ($b = 0.11, t(148.52) = 3.89, p < .001$). This suggests that when the level of trust is positive, recommendation agreement increases by approximately $b = 0.11$ units compared to a negative trust level. However, adding whether the trust was assessed or not does not significantly improve the model's explanatory power any further ($\chi^2(6) = 1.29, p = .256$).

6. Discussion

6.1. Summary of Results

The objective of the present research was to evaluate and examine the magnitude of the question-behavior effect of trust assessment on dependence in the context of an AI-supported task. Our results showed no support for the presence of the question-behavior effect in trust assessment and dependence. That is, assessing trust in participants did not affect their subsequent dependence behavior (H1 was not supported). Further, we did not find support for an effect of either trust assessment with single or multiple items in comparison to a control condition (H2 was not supported). While we found higher dependence after a late trust assessment (H3 was supported partially for agreement, not for task completion time), that was potentially due to an overall increase of dependence over time. Interestingly, the predictive value of self-reported trust assessment for dependence was surprisingly low, i.e., the prediction of agreement with AI recommendations did not benefit from analyzing trust. Further exploratory analyses revealed that participants followed AI recommendations in around 80% of all decisions during the course of the study. That mirrors the instructed reliability of the AI system, as participants were told in the introduction that the AI was only correct 80% of the time and were interacting with such a system during familiarization, but were interacting with a system with 100% reliability in the experimental task blocks. That is, participants' behavior seems to reflect this instructed reliability. In addition, we found higher dependence in the second experimental block, compared to the first experimental block. Interestingly, single- and multi-item assessments did not differ significantly nor were they significantly correlated.

6.2. Implications

Contrary to our hypotheses, trust assessment showed no significant influence on dependence in our study. This finding was independent of scale extent, i.e., we did not find any effect of scale length on dependence, and suggests that it is possible to assess trust without affecting the users' system-related behavior. Accordingly, we did not find support for any effects connected to the question-behavior effect [57, 54]. However, an alternative explanation for this could be based on the comparable low correlation between trust assessment and dependence. If behavior does not correlate as strongly with assessed attitude as initially assumed, it stands to reason that the question-behavior effect would not be observable in this context. A low relevance of the question-behavior effect for behavior in contrast to higher importance for intention [67] has been discussed before and could also explain the results of the present study.

Due to the descriptive match between instructed reliability (80%) and dependence (mean over both task blocks and groups is 78 %), instructed reliability appears to be more influential in determining behavior compared to assessed trust. By assuming the instructed reliability as a fixed intercept in a regression analysis, adding other variables (such as trust, perceived reliability, ATI, or propensity to trust) did not yield any substantial improvement in predicting recommendation agreement. While this insight aligns in general with the broader research that emphasizes the importance of perceived reliability as a key factor influencing user behavior (e.g., [68]), it highlights potential limitations of self-reported assessments. Consistent with the concept of "probability matching" as described in [69], our study found that participants complied with system recommendations depending on the instructed reliability of the system across various conditions. This suggests a general tendency of users to match their dependence to the anticipated system reliability [70]. This strategy can account for results in previous studies, however, there is no research on probability matching as described in the present study. While probability matching is an ineffective strategy for the given experiment (as the AI was 100% correct in the task blocks), participants potentially assumed that the instruction of 80% reliability definitely applied to the tasks presented to them. Accordingly, they may not question the instructed reliability but incorrectly assume it to be true. That is, after a series of correct answers from the AI system, they may have believed that due to the instructed reliability, an incorrect recommendation was guaranteed and looked for more closely. That behavior, which is similar

to the gambler's fallacy [71] could explain why participants reject correct AI recommendations. Our results would support a comparable strategy in participants, which would in turn be aligned with research on probability matching [69]. Further research needs to examine to what extent and under which conditions users reflect on the instructed reliability, as instructing reliability of systems might be central in shaping dependence.

Yet, in the context of our research question, the occurrence of probability matching is particularly interesting when considering the heterogeneity of trust assessment methods. These include the extent to which system performance and reliability are addressed within different assessments (e.g., in [34, 38, 36]). Therefore, the extent to which the participants' perception of reliability contributes to trust assessment could also be a factor influencing the (diverging) predictive value of the trust assessments. In our explanatory analysis, we did not find a significant correlation between dependence and the single item *"How much do you trust the system?"* of the multi-item trust scale. However, the relationship between perceived reliability reported by users in percent at the end of the study and dependence within the study did not exceed the relationship between trust assessment and dependence. This leaves the question, of which self-reported assessment - of trust, reliability, or even another related construct - may demonstrate higher predictive value for dependence. Interestingly, while the AI had 100% reliability, the majority of participants did not report perceived reliability over 80% (only 48 of 149) but mostly a perceived reliability under 80%. The massive gap between perceived and actual reliability has been already reported in earlier research ([72]. Although we expected to observe a description-experience gap (see [73]), participants reported perceiving 80% reliability of the system, despite this perception being incorrect. While we found a higher recommendation agreement over time, there is no clear trend or indication that participants approximated higher levels of recommendation agreement based on their individual experience. That strengthens the need for further research to understand the role of instructed (or outside of research settings, reported) reliability in comparison to participant experience and perception. It needs to be noted, however, that the perceived reliability in percent was only surveyed at the end of the study and not between experimental blocks and the study was potentially too short to clearly identify effects over time [74]. In addition, participants recruited via Profilic potentially perceived this question as the attention check and reported the instructed reliability here to keep eligible for compensation.

Our results imply that the instructed reliability of a system might be a better predictor of human dependence than self-reported trust assessment at least in tasks where AI systems give a clear recommendation for a specific action and AI recommendation is presented before participants make their own decision. Therefore, the predictive value of instructed reliability is important to consider when designing AI-based support systems: further research needs to examine how humans estimate the correctness of information provided by systems on their reliability and factors influencing reliability when assessing dependence (e.g., by using explanations, see [75]. In our experiment, no detailed information on factors for the system's reliability was given to the participants, for example, low confidence in the system when color-based rules are involved. Participants' estimation of AI reliability was lower than the actual AI reliability, which has been illustrated in earlier research [72, 76]. Further research on AI-supported pattern detection should explore methods to improve users' perception of system reliability.

As demonstrated by existing studies, the lack of predictive value of trust assessment for behavior can be contingent on the specific task design. In the present study, the AI system provided clear recommendations without considering the individual's state or hypotheses, setting it apart from evaluative AI as defined in [77]. As discussed before [25], trust might be more influential on dependence when users need to act before the systems or need to utilize system information to make a decision instead of receiving a recommendation. This suggests that the role of trust and its influence on the question-behavior effect might vary with different HAII modalities and sequences within a task. The differences in the predictive value of trust in the early group vs. the late group indicate that the role of trust could also change with growing familiarity with the system (see also [36]). It is possible that a different interaction dynamic between humans and AI could result in an altered role of trust and consequently, also, the question-behavior effect. This possibility underlines the need for further research into varying types of HAII and their implications for trust dynamics as well as for the question-behavior effect.

The results of our study show that trust research in HAII needs to examine the predictive value of established constructs when trying to reliably predict behavior. The use of trust assessment in an interaction where the AI provides a clear recommendation, however, does not influence the subsequent dependence demonstrated and does not result in a question-behavior effect. However, the predictive value of trust assessment remains a topic for

further study [27], and the findings need to be applied to different types of interactions (i.e., different types of collaboration/ combination of information processing and action regulation by the AI versus the user) to verify their generalizability. In particular, it remains an open question, under which conditions the instructed reliability of a system is more important than users' experience, for example, dependent on the frequency of feedback, the sequence of AI and human actions in a shared task, or the length of an experiment. As more regulation demands AI development to provide information on an AI system's reliability and limitations (e.g., the EU AI Act, [78]) the impact on dependence of such information needs to be studied in detail.

6.3. Limitations and Further Research

In interpreting the results of the present research, several aspects have to be kept in mind, especially regarding the design of the presented task. The potential complexity of Kandinsky Tasks may have been challenging for participants, limiting the interpretability of recommendation agreement and affecting the comparability of results with other studies, that utilize easier tasks. As participants' baseline (i.e., the number of tasks solved without the AI system) was at 60% correctly solved Kandinsky Tasks on average, the goal of 90% may have been an unrealistic high demand to function as an additional incentive. If participants did not expect to be able to meet the criterion for additional reward even with AI help, the vulnerability necessary to experience trust [29] was potentially not present. As previous research demonstrated, trust might decline when automated systems are not reliable enough [79]. For the attitude of the participants to be sufficiently strong to influence behavior, a better performing system (i.e., with reliability levels of 90% and more) could be needed. In addition, previous research demonstrated that only medium or low levels of an attitude may have drastically fewer effects than strong attitudes (see [80]). Differentiating to what extent participants were "giving up" and just following the AI recommendation is difficult to examine. However, as the difficulty was the same in all conditions, we do not believe that this impacted our experimental manipulation to a great extent. However, our exploratory results remain limited to tasks with higher complexity and strict AI recommendations (i.e., the AI system does not deliver information to determine the correct answer, but recommends specific answers directly).

It is important to note the specificity of the task as well as its abstract nature. While it was essential for our research to have a task where users could

not differ in pre-existing knowledge, collaborative human-AI tasks involve experienced humans, and expertise can change the result of the interaction [81]. Previous studies already demonstrated how assigning similar tasks to human and AI systems may lead to decreased performance of human users, comparable to social loafing [82, 83]. Accordingly, future research should aim to present tasks that require, for example, a combination of skills or exchange of information (see [84]).

As another important point, further research is needed to investigate varied levels of automation and interaction types to understand how these factors impact trust assessment and dependence. For example, users always saw the AI recommendation at the same time they saw the Kandinsky Figures - changing the sequence of decision-making could change the role of trust or even improve decision-making in general (see [85] compared to [76]).

7. Conclusion

In conclusion, the present study contributes to the understanding of trust dynamics in human-AI interaction and raises further questions for trust assessment and communication of information on the reliability of AI systems. Contrary to our initial hypotheses, we found that self-reported trust assessment has no discernible impact on reliance behavior in AI-supported tasks. This observation is underpinned by the absence of substantial predictive power of trust assessment in forecasting reliance. Notably, the strongest predictor of behavior in our experimental setup was the instructed reliability of the AI system, which is more relevant than self-reported trust assessment or perceived reliability. These findings suggest a need for a reevaluation of the role of trust assessment in AI interactions and point toward the greater influence of perceived system reliability on user behavior.

References

[1] Y. Duan, J. S. Edwards, Y. K. Dwivedi, Artificial intelligence for decision making in the era of big data–evolution, challenges and research agenda, International journal of information management 48 (2019) 63–71.

[2] J. Waring, C. Lindvall, R. Umeton, Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, Artificial intelligence in medicine 104 (2020) 101822.

[3] K. Arun, G. Ishan, K. Sanmeet, Loan approval prediction based on machine learning approach, IOSR J. Comput. Eng 18 (3) (2016) 18–21.

[4] D. Hadwick, S. Lan, Lessons to be learned from the dutch childcare allowance scandal: a comparative review of algorithmic governance by tax administrations in the netherlands, france and germany, World tax journal.-Amsterdam 13 (4) (2021) 609–645.

[5] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, International Journal of Human–Computer Interaction 36 (6) (2020) 495–504.

[6] M. Veale, F. Zuiderveen Borgesius, Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach, Computer Law Review International 22 (4) (2021) 97–112.

[7] T. Rieger, E. Roesler, D. Manzey, Challenging presumed technological superiority when working with (artificial) colleagues, Sci. Rep. 12 (1) (2022) 3768.

[8] J. L. Wright, J. Y. Chen, S. G. Lakhmani, Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability, IEEE Transactions on Human-Machine Systems 50 (3) (2019) 254–263.

[9] P. Madhavan, D. A. Wiegmann, Effects of information source, pedigree, and reliability on operator interaction with decision support systems, Human factors 49 (5) (2007) 773–785.

[10] K. A. Hoff, M. Bashir, Trust in automation: Integrating empirical evidence on factors that influence trust, Human factors 57 (3) (2015) 407–434.

[11] E. K. Chiou, J. D. Lee, Trusting automation: Designing for responsivity and resilience, Human factors 65 (1) (2023) 137–165.

[12] B. M. Muir, Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems, Ergonomics 37 (11) (1994) 1905–1922.

[13] S. Paul, L. Yuan, H. K. Jain, L. P. Robert Jr, J. Spohrer, H. Lifshitz-Assaf, Intelligence augmentation: Human factors in ai and future of work, AIS Transactions on Human-Computer Interaction 14 (3) (2022) 426–445.

[14] S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, T. H. Shaw, Measurement of trust in automation: A narrative review and reference guide, Frontiers in psychology 12 (2021) 604977.

[15] D. Miller, M. Johns, B. Mok, N. Gowda, D. Sirkin, K. Lee, W. Ju, Behavioral measurement of trust in automation: the trust fall, in: Proceedings of the human factors and ergonomics society annual meeting, Vol. 60, SAGE Publications Sage CA: Los Angeles, CA, 2016, pp. 1849–1853.

[16] S. Hergeth, L. Lorenz, R. Vilimek, J. F. Krems, Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving, Human factors 58 (3) (2016) 509–519.

[17] R. Wiczorek, D. Manzey, Is operators' compliance with alarm systems a product of rational consideration?, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 54, SAGE Publications Sage CA: Los Angeles, CA, 2010, pp. 1722–1726.

[18] H. Elder, C. Canfield, D. B. Shank, T. Rieger, C. Hines, Knowing when to pass: The effect of ai reliability in risky decision contexts, Human Factors: The Journal of the Human Factors and Ergonomics Society (2022) 001872082211006doi:10.1177/00187208221100691.
URL http://dx.doi.org/10.1177/00187208221100691

[19] E. A. Bustamante, A reexamination of the mediating effect of trust among alarm systems' characteristics and human compliance and reliance, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 53, SAGE Publications Sage CA: Los Angeles, CA, 2009, pp. 249–253.

[20] C. E. Patton, C. D. Wickens, The relationship of trust and dependence, Ergonomics 67 (11) (2024) 1535–1552.

[21] S. A. Bone, K. N. Lemon, C. M. Voorhees, K. A. Liljenquist, P. W. Fombelle, K. B. Detienne, R. B. Money, "mere measurement plus":

how solicitation of open-ended positive feedback influences customer purchase behavior, Journal of Marketing Research 54 (1) (2017) 156–170.

[22] P. Chandon, V. G. Morwitz, W. J. Reinartz, The short-and long-term effects of measuring intent to repurchase, Journal of Consumer Research 31 (3) (2004) 566–572.

[23] A. P. Zwane, J. Zinman, E. Van Dusen, W. Pariente, C. Null, E. Miguel, M. Kremer, D. S. Karlan, R. Hornbeck, X. Giné, et al., Being surveyed can change later behavior and related parameter estimates, Proceedings of the National Academy of Sciences 108 (5) (2011) 1821–1826.

[24] T. D. Wilson, J. W. Schooler, Thinking too much: introspection can reduce the quality of preferences and decisions., Journal of personality and social psychology 60 (2) (1991) 181.

[25] K. van Dongen, P.-P. van Maanen, A framework for explaining reliance on decision aids, International Journal of Human-Computer Studies 71 (4) (2013) 410–424. doi:10.1016/j.ijhcs.2012.10.018.
URL http://dx.doi.org/10.1016/j.ijhcs.2012.10.018

[26] R. B. Zajonc, Attitudinal effects of mere exposure., Journal of personality and social psychology 9 (2p2) (1968) 1.

[27] M. L. Bolton, Trust is not a virtue: Why we should not trust trust, Ergonomics in Design (2022) 10648046221130171.

[28] H. Müller, A. Holzinger, Kandinsky patterns, Artificial intelligence 300 (2021) 103546.

[29] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human factors 46 (1) (2004) 50–80.

[30] L. Bergkvist, J. R. Rossiter, The predictive validity of multiple-item versus single-item measures of the same constructs, Journal of marketing research 44 (2) (2007) 175–184.

[31] P. De Vries, C. Midden, D. Bouwhuis, The effects of errors on system trust, self-confidence, and the allocation of control in route planning, International Journal of Human-Computer Studies 58 (6) (2003) 719–735.

[32] H. Choung, P. David, A. Ross, Trust in ai and its role in the acceptance of ai technologies, International Journal of Human–Computer Interaction 39 (9) (2023) 1727–1739.

[33] M. Brzowski, D. Nathan-Roberts, Trust measurement in human–automation interaction: A systematic review, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 63, SAGE Publications Sage CA: Los Angeles, CA, 2019, pp. 1595–1599.

[34] J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an empirically determined scale of trust in automated systems, International journal of cognitive ergonomics 4 (1) (2000) 53–71.

[35] R. C. Mayer, J. H. Davis, The effect of the performance appraisal system on trust for management: A field quasi-experiment., Journal of Applied Psychology 84 (1) (1999) 123–136. doi:10.1037/0021-9010.84.1.123.
URL http://dx.doi.org/10.1037/0021-9010.84.1.123

[36] M. Körber, E. Baseler, K. Bengler, Introduction matters: Manipulating trust in automation and reliance in automated driving, Applied ergonomics 66 (2018) 18–31.

[37] Y. S. Razin, K. M. Feigh, Converging measures and an emergent model: A meta-analysis of human-machine trust questionnaires, ACM Transactions on Human-Robot Interaction 13 (4) (2024) 1–41. doi:10.1145/3677614.
URL http://dx.doi.org/10.1145/3677614

[38] D. Trommler, C. Attig, T. Franke, Trust in activity tracker measurement and its link to user acceptance (2018).

[39] S. M. Merritt, D. Lee, J. L. Unnerstall, K. Huber, Are well-calibrated users effective users? associations between calibration of trust and performance on an automation-aided task, Human Factors 57 (1) (2015) 34–47.

[40] B. D. Wright, Fundamental measurement for psychology, The new rules of measurement: What every psychologist and educator should know (1999) 65–104.

[41] J. Meyer, Conceptual issues in the study of dynamic hazard warnings, Human Factors: The Journal of the Human Factors and Ergonomics Society 46 (2) (2004) 196–204. doi:10.1518/hfes.46.2.196.37335.
URL http://dx.doi.org/10.1518/hfes.46.2.196.37335

[42] W. A. Cunningham, K. J. Preacher, M. R. Banaji, Implicit attitude measures: Consistency, stability, and convergent validity, Psychological science 12 (2) (2001) 163–170.

[43] L. Wang, G. A. Jamieson, J. G. Hollands, Selecting methods for the analysis of reliance on automation, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 52, SAGE Publications Sage CA: Los Angeles, CA, 2008, pp. 287–291.

[44] C. X. Kerasidou, A. Kerasidou, M. Buscher, S. Wilkinson, Before and beyond trust: reliance in medical ai, Journal of medical ethics 48 (11) (2022) 852–856.

[45] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, H. P. Beck, The role of trust in automation reliance, International journal of human-computer studies 58 (6) (2003) 697–718.

[46] E. Solberg, M. Kaarstad, M. H. R. Eitrheim, R. Bisio, K. Reegård, M. Bloch, A conceptual model of trust, perceived risk, and reliance on ai decision aids, Group & Organization Management 47 (2) (2022) 187–222.

[47] B. M. Muir, N. Moray, Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation, Ergonomics 39 (3) (1996) 429–460.

[48] S. M. Merritt, K. Huber, J. LaChapell-Unnerstall, D. Lee, Continuous calibration of trust in automated systems, Air Force Research Laboratory Technical Report AFRL-RH-WP-TR-2014-0026 (2014).

[49] S. M. Merritt, H. Heimbaugh, J. LaChapell, D. Lee, I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system, Human factors 55 (3) (2013) 520–534.

[50] A. Papenmeier, D. Kern, G. Englebienne, C. Seifert, It's complicated: The relationship between user trust, model accuracy and explanations in ai, ACM Trans. Comput.-Hum. Interact. 29 (4) (mar 2022). doi:10.1145/3495013.
URL https://doi.org/10.1145/3495013

[51] T. Ueno, Y. Kim, H. Oura, K. Seaborn, Trust and reliance in consensus-based explanations from an anti-misinformation agent, in: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA '23, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3544549.3585713.
URL https://doi.org/10.1145/3544549.3585713

[52] S. Hoesterey, L. Onnasch, The effect of risk on trust attitude and trust behavior in interaction with information and decision automation, Cognition, Technology & Work 25 (1) (2023) 15–29.

[53] J. Sanchez, W. A. Rogers, A. D. Fisk, E. Rovira, Understanding reliance on automation: effects of error type, error distribution, age and experience, Theoretical Issues in Ergonomics Science 15 (2) (2011) 134–160. doi:10.1080/1463922x.2011.611269.
URL http://dx.doi.org/10.1080/1463922X.2011.611269

[54] V. G. Morwitz, G. J. Fitzsimons, The mere-measurement effect: Why does measuring intentions change actual behavior?, Journal of consumer psychology 14 (1-2) (2004) 64–74.

[55] G. Godin, P. Sheeran, M. Conner, M. Germain, Asking questions changes behavior: Mere measurement effects on frequency of blood donation., Health Psychology 27 (2) (2008) 179–184. doi:10.1037/0278-6133.27.2.179.
URL http://dx.doi.org/10.1037/0278-6133.27.2.179

[56] G. J. Fitzsimons, P. Williams, Asking questions can change choice behavior: Does it do so automatically or effortfully?, Journal of Experimental Psychology: Applied 6 (3) (2000) 195–206. doi:10.1037/1076-898x.6.3.195.
URL http://dx.doi.org/10.1037/1076-898X.6.3.195

[57] T. S. A. P. R. L. C. W. E. M. G. G. Sarah Wilding, Mark Conner, P. Sheeran, The question-behaviour effect: A theoretical and methodological review and meta-analysis, European Review of Social Psychology 27 (1) (2016) 196–230. arXiv:https://doi.org/10.1080/10463283.2016.1245940, doi:10.1080/10463283.2016.1245940.
URL https://doi.org/10.1080/10463283.2016.1245940

[58] H. Song, N. Schwarz, If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation, Psychological Science 19 (10) (2008) 986–988. doi:10.1111/j.1467-9280.2008.02189.x.
URL http://dx.doi.org/10.1111/j.1467-9280.2008.02189.x

[59] Prolific Team, Prolific, Prolific Academic Ltd, London, United Kingdom (2014).
URL https://www.prolific.com

[60] LimeSurvey Project Team / Carsten Schmitz, LimeSurvey: An Open Source survey tool, LimeSurvey Project, Hamburg, Germany (2012).
URL http://www.limesurvey.org

[61] J. Stuckman, G.-Q. Zhang, Mastermind is np-complete (2005). arXiv:cs/0512049.

[62] F. Faul, E. Erdfelder, A. Buchner, A.-G. Lang, Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses, Behavior research methods 41 (4) (2009) 1149–1160.

[63] N. J.-M. Blackman, J. J. Koval, Interval estimation for cohen's kappa as a measure of agreement, Statistics in Medicine 19 (5) (2000) 723–741. doi:10.1002/(sici)1097-0258(20000315)19:5¡723::aid-sim379¿3.0.co;2-a.
URL http://dx.doi.org/10.1002/(SICI)1097-0258(20000315)19:5<723::AID-SIM379>3.0.CO;2-A

[64] N. Haines, P. D. Kvam, B. M. Turner, Explaining the description-experience gap in risky decision-making: learning and memory retention during experience as causal mechanisms, Cogn. Affect. Behav. Neurosci. 23 (3) (2023) 557–577.

[65] L. M. Sullivan, R. B. D'Agostino, Robustness of the t test applied to data distorted from normality by floor effects, Journal of dental research 71 (12) (1992) 1938–1943.

[66] J. Cohen, Statistical power analysis, Current directions in psychological science 1 (3) (1992) 98–101.

[67] K. J. Chapman, Measuring intent: There's nothing ?mere? about mere measurement effects, Psychol. Mark. 18 (8) (2001) 811–841.

[68] J. Hutchinson, L. Strickland, S. Farrell, S. Loft, Human behavioral response to fluctuating automation reliability, Appl. Ergon. 105 (103835) (2022) 103835.

[69] M. L. Bartlett, J. S. McCarley, Benchmarking aided decision making in a signal detection task, Hum. Factors 59 (6) (2017) 881–900.

[70] L. Tikhomirov, M. L. Bartlett, J. Duncan-Reid, J. S. McCarley, Identifying inefficient strategies in automation-aided signal detection, J. Exp. Psychol. Appl. (Jul. 2023).

[71] M. Lawrence, S. Makridakis, Factors affecting judgmental forecasts and confidence intervals, Organizational Behavior and Human Decision Processes 43 (2) (1989) 172–187.
URL https://EconPapers.repec.org/RePEc:eee:jobhdp:v:43:y:1989:i:2:p:172-187

[72] T. Rieger, E. Roesler, D. Manzey, Challenging presumed technological superiority when working with (artificial) colleagues, Scientific Reports 12 (1) (Mar. 2022). doi:10.1038/s41598-022-07808-x.
URL http://dx.doi.org/10.1038/s41598-022-07808-x

[73] J. Chen, S. Mishler, B. Hu, N. Li, R. W. Proctor, The description-experience gap in the effect of warning reliability on user trust and performance in a phishing-detection context, International Journal of Human-Computer Studies 119 (2018) 35–47. doi:10.1016/j.ijhcs.2018.05.010.
URL http://dx.doi.org/10.1016/j.ijhcs.2018.05.010

[74] D. Holliday, S. Wilson, S. Stumpf, User trust in intelligent systems: A journey over time, in: Proceedings of the 21st international conference on intelligent user interfaces, 2016, pp. 164–168.

[75] M. Nourani, S. Kabir, S. Mohseni, E. D. Ragan, The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7 (2019) 97–105.

[76] T. Rieger, L. Kugler, D. Manzey, E. Roesler, The (im)perfect automation schema: Who is trusted more, automated or human decision support?, Human Factors: The Journal of the Human Factors and Ergonomics Society (Aug. 2023). doi:10.1177/00187208231197347.
URL http://dx.doi.org/10.1177/00187208231197347

[77] T. Miller, Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 333–342. doi:10.1145/3593013.3594001.
URL https://doi.org/10.1145/3593013.3594001

[78] European Commission, Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, COM(2021) 206 final (2021).

[79] N. Moray, T. Inagaki, M. Itoh, Adaptive automation, trust, and self-confidence in fault management of time-critical tasks, J. Exp. Psychol. Appl. 6 (1) (2000) 44–58.

[80] Y. J. Kim, J. U. Chun, J. Song, Investigating the role of attitude in technology acceptance from an attitude strength perspective, International Journal of Information Management 29 (1) (2009) 67–77. doi:10.1016/j.ijinfomgt.2008.01.011.
URL http://dx.doi.org/10.1016/j.ijinfomgt.2008.01.011

[81] K. Inkpen, S. Chappidi, K. Mallari, B. Nushi, D. Ramesh, P. Michelucci, V. Mandava, L. H. Vepřek, G. Quinn, Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making, ACM Trans. Comput.-Hum. Interact. 30 (5) (sep 2023).

doi:10.1145/3534561.
URL https://doi.org/10.1145/3534561

[82] I. Inuwa-Dutse, A. Toniolo, A. Weller, U. Bhatt, Algorithmic loafing and mitigation strategies in Human-AI teams, Computers in Human Behavior: Artificial Humans (100024) (2023) 100024.

[83] D. H. Cymek, A. Truckenbrodt, L. Onnasch, Lean back or lean in? exploring social loafing in human-robot teams, Front. Robot. AI 10 (2023) 1249252.

[84] N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, I. Dunning, S. Mourad, H. Larochelle, M. G. Bellemare, M. Bowling, The hanabi challenge: A new frontier for AI research, Artif. Intell. 280 (103216) (2020) 103216.

[85] M. T. Dzindolet, L. G. Pierce, H. P. Beck, L. A. Dawe, The perceived utility of human and automated aids in a visual detection task, Human Factors: The Journal of the Human Factors and Ergonomics Society 44 (1) (2002) 79–94. doi:10.1518/0018720024494856.
URL http://dx.doi.org/10.1518/0018720024494856

7 Study 4: Examining how to improve integrated & interdependent information processing

7.1 Summary of Study 4

The fourth study included in the present dissertation focused on interdependence in information processing in diagnostic AI-supported tasks. When examining blood vessel to diagnose deep vein thrombosis, pressure is a central component of the examination process. However, since AI currently available AI systems may struggle to detect whether users apply the correct level of pressure, instructions were developed to highlight informational interdependence. The study utilized a between-subjects design assessing both, performance in terms of diagnostic quality as well as automation-related user experience. The results demonstrate that instructing users to experience how pressure, i.e., information not available to the system but solely to the human, can positively influence diagnostic quality.

7.2 Relevance within the dissertation

While other studies presented within this dissertation provides empirical results on how XAI can affect automation-related UX and performance, they mainly focus on

explanations as an artifact produced by a technical system. Studying the effect of instructions widens that narrow perspective on XAI while simultaneously offering insights into novices' usage of medical AI. Also, since this dissertation also presents a conceptual model of integrated information processing, the results of the fourth study demonstrate the importance of integrating information processing of both, AI and human.

7.3 Contribution to Study 4

I developed the idea to modify instructions and integrated both applied theories, interdependence in Human-AI interaction and seamful design, to provide a basis for the experiment. I developed the instructions and experimental design and developed all experimental material. While the study was conducted with support of study assistants, I conducted the analysis and preparation of all data for a manuscript.

Information Interdependence in Human-AI Collaboration: Learners' Perception and Performance in Cooperative Ultrasound Diagnosis

TIM SCHRILLS, University of Lübeck, Germany
NIELS VAN BERKEL, Aalborg University, Denmark
THOMAS FRANKE, University of Lübeck, Germany

Novices may particularly benefit from artificial intelligence (AI) in diagnostic tasks, e.g., in ultrasound-based diagnosis where images are highly volatile. However, due to challenges in capturing image-relevant pressure information, AI systems results may be affected by novices' ability to physically conduct ultrasound-based examinations. In a between-subject experiment, we instructed medical novices to manipulate pressure levels during ultrasound-based diagnosis and observed how they may cause failures and compared them to a control group without pressure-related instructions. We assessed perceived system traceability and the relationship between confidence and diagnostic quality. Our results (N = 42) indicated that instructing novices did not affect perceived traceability and confidence. However, having users test the system on themselves and manipulate pressure led to higher-quality ultrasound images. Overall, our study demonstrates the importance of communicating information interdependencies between human and AI partners in diagnostic tasks.

CCS Concepts: • **Human-centered computing → Empirical studies in interaction design**; *Empirical studies in HCI*; *Interactive systems and tools.*

Additional Key Words and Phrases: Human-AI Cooperation, Decision Support Systems, Explainable AI, Medical AI

ACM Reference Format:
Tim Schrills, Niels van Berkel, and Thomas Franke. 2025. Information Interdependence in Human-AI Collaboration: Learners' Perception and Performance in Cooperative Ultrasound Diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 06, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/xx.xxxx/xxxxxxx.xxxxxxx

1 INTRODUCTION

Automated information processing is progressively supporting medical decisions, affecting professionals, as well as patients [1]. The influence of artificial intelligence (AI) can be found in both, diagnostic [24] as well as therapeutic contexts [65]. Multiple studies suggest that the support of AI can increase e.g., diagnostic accuracy (e.g. [9]). However, as human users stay responsible for decisions that may have severe consequences, there is a growing demand for transparency [6] in automated information processing, even on regulatory levels ([54]). Especially in systems not operating based on symbolic rules, but, e.g., neural networks, medical users might perceive low levels of understandability [15] due to a lack of information representation that users can integrate into their decisions. As a result, research on explainable AI (XAI) has been carried out to develop ways of making AI systems understandable to their users [56] and allowing users to detect biases and errors in automated information processing (see [13]). As such, efforts undertaken in XAI

Manuscript submitted to ACM

research resemble earlier work in human-automation interaction, that focused on methods to prevent miscalibration of trust [38] or users experiencing out-of-the-loop unfamilarity (see [61]).

However, existing approaches in XAI do not lead to coherent results. On the one hand, multiple studies demonstrate positive effects of explanations, e.g., on experienced trustworthiness, accuracy or bias detection [29, 55, 58]. For example, Weitz et al. [59] demonstrated that a speech recognition system with explanations given by an agent leads to higher levels of trust than without. For the medical domain, Naiseh et al. [35] demonstrated that explanations influence user engagement with decision support systems (DSS), leading to higher frequency of interaction and better calibration. However, multiple studies also demonstrated that explanations can lead to undesired effects, especially in cases where automated information processing might lead to wrong results. Bansal et al. [4] demonstrated the likelihood of accepting an AI's recommendation is higher after users received an explanation, making it more likely that users followed the wrong suggestions. Similarly, Chromik et al. [10] introduced the term 'placebic' explanations, which did not offer helpful information to the user but was still rated as such (see also [11]). Potential reasons could be the expertise of the users [52] or effects of information overload associated with additional information given by explanations [42].

In contrast the present work expands explanations that visually demonstrate to users how their behavior affects an intelligent system to allow them to directly experience this effect to instructions about how they initially interact with system. Specifically, we designed instructions that highlight a system's diagnostic information processing and allow users to experience the significance of their own actions and the information provided by them. To this end, we introduced participants to an automated diagnostic system for detecting deep vein thrombosis and instructed them to intentionally try to let the system fail as well as use it on themselves. Our goal was to explore how this approach affects users' understanding and mental model of the shared information processing [3] and their confidence in its diagnostic abilities. We were particularly interested in observing changes in their subjective experience and the quality of the examination. By adding instruction as a method of explanation, our research diverges from traditional XAI strategies (see [5]), focusing on experiential and skill-based [51] learning rather than explanatory frameworks, and aims to contribute to the broader discourse on the development of human-centered AI.

Hence, the objective of our research was to examine how instruction focusing on informational interdependence and users' influence on information processing can affect user experience and diagnostic quality. To this end, we introduced a functioning AI system for detection of deep vein thrombosis to novices and varied how it was introduced, i.e., whether system capabilities and limits were explicitly experienced or not. We aimed to investigate whether explicitly experiencing one's effect on the systems information processing can support the development of confidence, raise examination quality and improve user experience. All in all, the present paper explore an innovative approach to develop instructions for novice users of AI-based DSS, which will support design of AI systems that sufficiently comply with ethical and regulatory requirements. It demonstrates that designing instructions to appropriately demonstrate interdependence between human users and AI systems can positively affect the utilization of AI-based DSS.

2 RELATED WORK

2.1 AI Support for Ultrasound Analysis in Deep Vein Thrombosis Detection

While concrete fatality rates of deep vein thrombosis (DVT) are generally elusive [2], DVT is assumed to be one of the most frequent reasons for death in hospitals worldwide [21]. DVT is a condition characterized by the formation of a blood clot in a deep vein, typically in the legs [21]. The presence of this clot can result in a range of symptoms, including leg pain and swelling. In some cases, the condition may manifest without any discernible signs. The condition

is serious in that the clot has the potential to dislodge and travel through the bloodstream, resulting in a potentially life-threatening pulmonary embolism if it reaches the lungs (see [22]). DVT affects approximately 8 in 100 individuals over the course of their lifetime [25], making it one of the most prevalent conditions with severe medical consequences if left untreated [16]. In the absence of treatment, DVT can result in significant complications. The formation of a blood clot can obstruct blood flow, leading to tissue damage and an increased risk of further clot formation. The standard treatment for DVT involves the administration of anticoagulant medications, also known as blood thinners, which serve to prevent the growth of existing clots and to reduce the risk of new clots forming [60]. Given the potential adverse effects of treatment and the associated costs, precise diagnostics are of paramount importance.

The primary diagnostic approach for DVT is a duplex ultrasound, a non-invasive test that utilizes sound waves to visualize blood flow and detect clots in the veins [31]. The ultrasound probe is positioned at a relatively elevated point on the leg in proximity to the vein. It is crucial to ensure that an adequate amount of ultrasound gel and pressure is applied to effectively visualize the vessels, as illustrated in Figure 1. Identifying the optimal angle, pressure, and position can often be challenging for users, i.e., medical handlers (see [48]). Once the ultrasound device has been correctly positioned, the vessel is compressed by applying pressure. The requisite degree of pressure is contingent upon factors such as body mass index (BMI). A completely closed vein indicates the absence of a clot obstructing the vessel. Therefore, it is essential to apply sufficient pressure but to avoid compressing the vein prior to the examination too much (see [48]). Subsequently, the procedure is repeated on other parts of the leg, in accordance with the standard operating procedure. In sum, this process is one that requires both experience and skill, given the necessity of precise positioning and the application of optimal pressure.

Unfortunately, the necessity for highly qualified personnel to perform examinations for DVT and the high prevalence of DVT are typically not met by the availability of sufficiently trained medical experts. Therefore, the provision of technological assistance represents a promising avenue for facilitating the widespread use of skill-dependent diagnostic processes, such as those employed in the diagnosis of DVT. The integration of AI-based decision support systems is particularly advantageous for those new to the medical field. In the case of DVT, a number of AI systems have already been developed which can provide assistance in the diagnostic process. Such systems comprise predictive models that estimate the probability of DVT occurrence based on identified risk factors at various time points, including the elevated risk following surgical procedures. For instance, [57] illustrated that DVT can be accurately predicted by integrating physiological data, thereby demonstrating the potential for DSS to support the monitoring of high-risk patients. The present paper, however, is concerned with systems designed for the examination of DVT by ultrasound devices. To the best of our knowledge, apart from existing guidelines [37], there have been no other guiding systems, especially no DSS, in ultrasound diagnosis of DVT so far.

All in all, the visual examination of veins represents the best-known diagnostic approach when there is an acute suspicion of deep vein thrombosis, with ultrasound being the least invasive technique available. The performance of a venous ultrasound entails the navigation of numerous intricate tasks that demand a high degree of expertise. This examination process is characterized by a continuous cycle of information processing and action regulation [7], both of which are aimed at establishing an accurate diagnosis.

The system utilized in the present study (see Section 3.1 for a detailed description) provides assistance to the user throughout the process. The examination of the sensory system, specifically the control of the ultrasound device, is conducted by the examiner. Even advanced systems today are unable to perform the compression itself, and there is no pressure sensor that provides information about the pressure exerted. This is because the optimal pressure required depends heavily on the BMI and the anatomy of the person being examined, and therefore such information would only

be useful to a limited extent. However, individuals lacking the requisite experience are particularly reliant on guidance in correctly positioning the ultrasound probe and applying compression. The ultrasound-based DVT examination thus represents a particularly promising area of human-AI cooperation, given the interdependencies inherent to the information processing involved.

2.2 Accounting for Interdependence in AI-supported DVT Diagnosis for Diagnostic Information Processing

Despite the expansion of sensor technology, which has led to greater automation in information processing, AI systems often remain reliant on human input [20]. Concurrently, AI systems are employed to automate the information processing procedure, whereby they supplant human responsibilities and, in turn, provide data to human users. This implies that AI systems are contingent upon the data they receive, and humans are similarly reliant on the information processed by AI systems. This is due to the complementary capabilities that both partners contribute to the collaborative information processing (see [17]): in the case of DVT, humans are able to position the ultrasound probe and perform compression. Conversely, the system is capable of analyzing the generated image data based on the training data and determining whether the probe is correctly positioned.

If both parties in a collaborative endeavor are reliant on one another, we may characterize this as interdependence (see [30]). In the case of DVT, this interdependence is evident in the processing of information, particularly in the acquisition and analysis of data. One consequence of this interdependence is that, despite the automation, human users must fulfill a monitoring and control function, especially over the information they provide. One manifestation of complacency (see [62]) or false security in automation could be a lack of awareness of informational interdependence. In that case, a user might not realize that an AI system cannot monitor the information provided by the user.

A lack of awareness of interdependence can result in disadvantages in cooperation. For instance, if the user of a DSS in the field of DVT is not aware that insufficient pressure in the system is not recognized and can lead to a misdiagnosis, the performance of the human-AI team can deteriorate. Consequently, this indicates that the user has an inadequate mental model, which gives rise to erroneous expectations regarding the system's performance. The system's trust and subjective information processing awareness may be negatively affected by an incomprehensible expectation violation. Existing literature (e.g., [49]) suggests that the safety of using an AI system is contingent upon its comprehensibility. Consequently, an incomplete understanding of interdependencies may potentially lead to a reduction in safety. Furthermore, mutual predictability (as discussed by [23]) is also affected if an accurate assessment of the system's dependence on the correct execution of the examination by the user cannot be made, which in turn can limit the effectiveness of cooperative information processing and thus the diagnostic quality.

While Johnson et al. [18] discuss methods for overcoming interdependence in the context of co-active design, these are also designed with particular consideration for physical capabilities and sequential processes. The case of DVT illustrates the necessity of considering informational interdependence in diagnostic processes. In accordance with the concepts discussed by Parasuraman et al. [39], distinct processes are posited to occur during the course of information processing: (1) information acquisition, (2) information analysis, (3) decision selection, and (4) action implementation. In the context of diagnosis, the first three of these stages of information processing are particulary relevant. In the case of DVT, these processes are also highly interdependent and interconnected. For example, compression data can only be recorded when the ultrasound probe is in the correct position. In turn, prior analysis of the current image information is necessary to determine the positioning. However, the design of an AI system in the context of DVT diagnosis must convey this highly interwoven and not purely sequential informational interdependence.

2.3 Seamful Design as Approach in Information Interdependence

To address informational interdependence in human-AI cooperation, design approaches must address where a partner's exclusive information has an effect (see [23]). This means that a user must be able to see where the system has redundant information for checking and where this is not possible. The aim of the design is thus to correct the mental model of information processing and to draw the user's attention to interdependencies [18]. Hence, the aim of seamful design [8] is to optimize mental models. Mismatches between the conceptualization of the system (often as a system that works as automatically as possible) and its use in reality (with interdependency between humans and the system) should be actively presented. Chalmers [8] also speaks of the fact that those mismatches should be strategically revealed, i.e., designs should actively bring them to users attention.

Specifically for the comprehensibility of AI systems, Ehsan et al. [14] describe Seamful Explainable AI as an approach for developing human-centered and comprehensible XAI systems. The aim of the present work was to integrate the idea of seams, i.e. design decisions that in this context reveal the interdependencies in information processing, into the diagnosis of DVT and to examine whether this has a positive effect on the experience and performance of human-AI cooperation. When implementing Seamful Explainable AI [14] for novices, explicit attention was paid to focus on training and instruction interventions for novices.

2.4 Research Contribution

To explore the application and implications of seamful design in AI-supported diagnostic tasks with informational interdependence, we address several key research objectives (RO). Firstly, we investigate how seamful design can be effectively applied in diagnostic tasks where human and AI systems rely on shared and distinct information sources (**RO1**). Our initial investigation includes examining the role of seamful design in enhancing user understanding of the collaborative information processing required for accurate diagnostics. Secondly, we assess whether Information Relevance Training, a specific intervention designed to improve the comprehensibility of AI systems, influences perceived trustworthiness, perceived usefulness, and traceability, operationalized through Subjective Information Processing Awareness (SIPA) in users (**RO2**). Further, we examine if such training can lead to improvements in the quality of diagnostic assessments (**RO3**), enhance user confidence in their diagnostic decisions (**RO4**), and impact the relationship between traceability and user confidence ratings (**RO5**). These research objectives aim to provide insights into optimizing human-AI cooperation in the diagnostic process by enhancing user awareness of informational interdependencies through targeted design and training interventions.

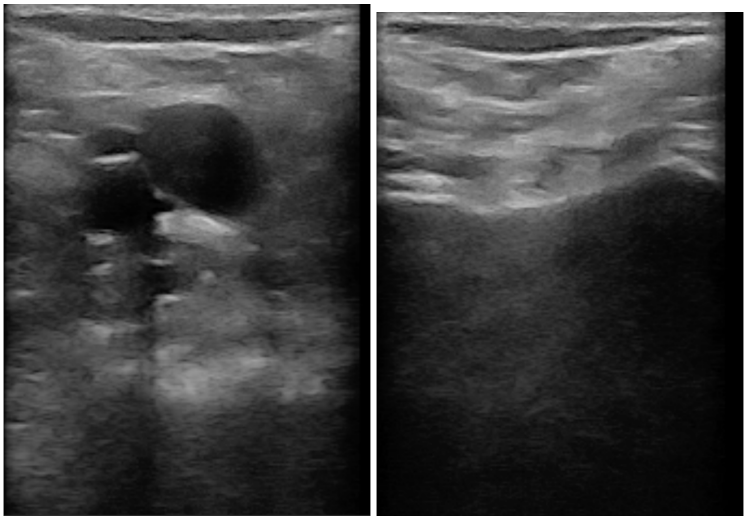
2.5 Designing Seamful Instructions for Deep Vein Thrombosis

The first of our research objectives was to identify informational interdependencies in the AI-supported diagnosis of DVT and to design approaches to represent these interdependencies in the mental models of the users. To do this, we first identified all the information that is available exclusively to humans or systems. Figure 2 summarizes the elements that are crucial for the design.

One piece of information available exclusively to the system is the direction in which the ultrasound device must be moved to reach the correct position for the examination. We assume that the users are novices who have little experience with DVT examinations. However, the system can provide information via the display – for example, a description to move towards the direction of the patient's foot. The system-exclusive information can thus enter the shared information space.

Furthermore, there is a range of information that both partners possess and, therefore, does not trigger any interdependency. For example, the current position of the probe (i.e., the ultrasound handheld device) can be visually perceived by the human and reliably calculated by the system based on the current image. This also applies to the angle at which the probe sits on the body to be examined. At the time of the examination, further information has already been shared (e.g., BMI, age, and which leg is involved), which is done, for example, via an entry mask. This data is, therefore, also in the shared information space.

Therefore, our design concentrated on the relevant information on the printout. On the one hand, this cannot be given to the system by sensors. On the other hand, in contrast to data such as BMI or age, it changes during the examination and, therefore, cannot be continuously transmitted from the person to the system. Since pressure remains exclusively in the human information space, but is crucial for a successful examination and for the analysis of the system, we developed instructions that emphasize the special role of the person in terms of the information provided by pressure. Two approaches were chosen for this: on the one hand, novices were asked to try out the ultrasound probe on themselves. The additional perceptive information helps them to understand how much pressure is needed to achieve changes in the ultrasound image and to reduce uncertainty about the pain threshold of the applied pressure. The exact instruction was (translated from German): "Please carry out the examination on yourself. Please note exactly how much pressure you need to exert in order to completely close the vein found." On the other hand, they should consciously apply too little pressure as part of a training examination and thus create errors in the system. This should help novices to improve their mental model and understand how pressure changes can affect system performance. The exact instruction was (translated from German): "Please deliberately apply too little pressure during the next examination and try to perform an incomplete compression. Observe how the system reacts to this. " The instructions, which we refer to as information relevance training, were examined in a laboratory experiment.



(a) Ultrasound image of uncompressed vein without overlay **(b)** Vein compressed no overlay

Fig. 1. Ultrasound image of compressed vein without overlay

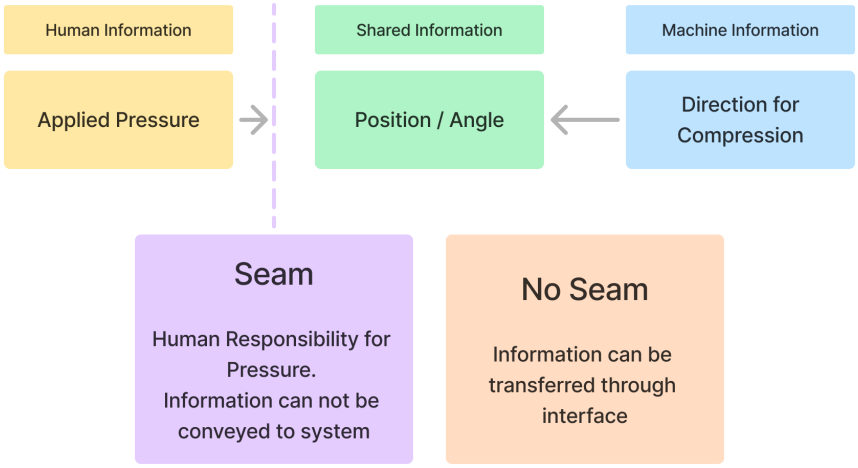


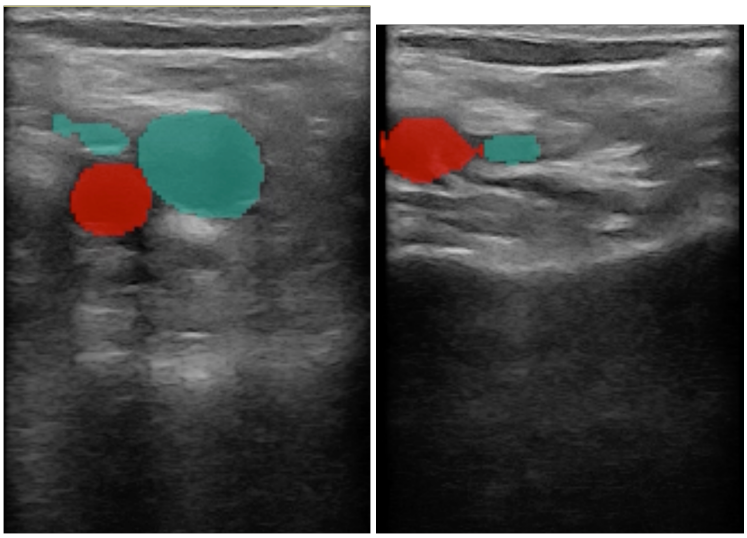
Fig. 2. Depiction of Information available only to Human and Machine and Identification of Seam to address informational interdependence

3 METHODS

For the investigation of informational interdependence, a system was used that is already in clinical studies [19]. Our focus was particularly on a modification of the instructions, which is why the specific system is only briefly explained below.

3.1 Hardware and Software

In our study, we utilized the ThinkSono system [53], which is a medical software that combines a mobile application and a cloud-based dashboard to assist healthcare personnel who are not trained in ultrasound technology in acquiring and storing compression ultrasound data for the detection of DVT. The system employs compatible handheld ultrasound devices, such as the Clarius models (e.g., [12] as used in our study), to capture high-resolution images, which are then analyzed using a machine learning model that segments veins and arteries in real time. The mobile application provides users with on-screen prompts to guide them through the examination process as depicted in Figure 3. The system is trained on annotated ultrasound images to identify anatomical landmarks (such as the branching of veins) as a basis for this.



(a) Ultrasound image of uncompressed vein with AI-based overlay **(b)** Vein compressed no overlay

Fig. 3. Ultrasound image of compressed vein with AI-based overlay

3.2 Recruitment and Procedure

Participants (N = 42) were recruited at the local university via online forums and consisted solely of medical students in their third year of study or above. All participants had prior experience with ultrasound diagnosis through their education, but n = 11 stated that they had no general experience with ultrasound examinations. We decided to include them in the analysis anyway, as their level of experience with ultrasound is representative of novices in the ultrasound-based diagnosis of DVT. Prior to participation, an instruction was handed out and informed consent was given by all participants. This study was conducted in accordance with the Declaration of Helsinki and positively reviewed by the [anonymous for submission] ethics committee.

We applied a randomized within-subject design with a control group, as visualized in Figure 4. In the experimental groups, the order of tasks was varied to control for sequence effects, effectively leading to three different conditions (control, fail-first, and self-first condition). All participants first watched a standardized instructional video, followed by a questionnaire that constituted the baseline and captured their initial expectations.

Following that, participants in the control group took two more exams. All exams were carried out on a simulating patient, who was the same for all participants and had a (protoypical) BMI of 22. The first of these exams was assessed as a baseline (Exam 1) in terms of diagnostic quality. After the second exam in the control group, the questionnaire was presented again, which counts as 'T1' for the variables trust, SIPA, usefulness, and confidence. The same procedure was then repeated with the third exam (Exam 3) for diagnostic quality and the questionnaire after the fourth exam as 'T2' for automation-related user experience. After that, the eamined leg was switched to induce some level of transfer requirements from participants and exams 5 and 6 were both evaluated in terms of diagnostic quality. The sixth exam was followed by the final questionnaire, which represents 'T3' for automation-related UX.

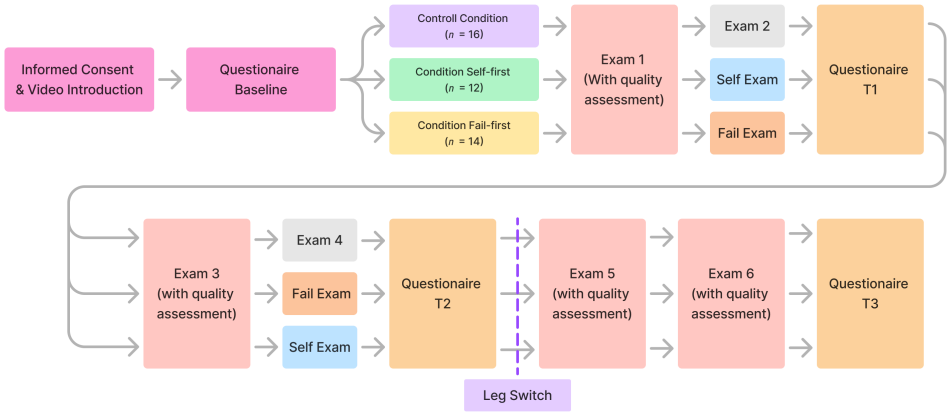


Fig. 4. Participant flow through the study.

The first experimental group followed a 'self-first' condition. This group differed from the control group at the time of exam 2 and exam 4. Instead of exam 2, this group was instructed to test the system on themselves. Instead of exam 4, this group was instructed to voluntarily apply incorrect levels of pressure to fail the exam. For the 'fail-first' group, the interventions were reversed, i.e. instead of exam 2, they were instructed to voluntarily apply incorrect levels of pressure to fail the exam and instead of exam 4, they were instructed to test the system on themselves. The measurement of the questionnaires, the change of leg, and, in particular, exams 5 and 6 were the same for these groups as for the control group. Despite random assignment, not all groups were exactly equally distributed (n = 16 for the control condition, n = 12 for Self-first, and n = 14 for Fail-first).

In order to examine our hypotheses, perceived trust, perceived usefulness, SIPA, as well as perceived confidence were assessed via questionnaires. All scales used 6-point Likert response items ranging from completely disagree = 1, largely disagree = 2, slightly disagree = 3, slightly agree = 4, largely agree = 5, to completely agree = 6. Trust was measured using the same questions as [46], as well as perceived usefulness. SIPA assessment utilized the question published in [47]. Finally, for perceived confidence, five items were developed by two of the authors, that also used the same 6-point Likert response items. The items were: (1) "I am not sure to what extent my behavior during the investigation was correct", (2) I am sure that I can detect errors in the investigation, (3) "I have the feeling that I can produce useful images during the examination", (4) "I know how to carry out a correct examination with the system", and (5) "I feel confident when carrying out the examination with the system".

Quality of medical diagnostic assessment was graded following standards of CIT by two independent reviewers who were, at the point of the study, medical students with experience in ultrasound diagnosis. Quality was rated from 0 to 5, with 5 implying that all conditions of a high qualitative ultrasound recording were fulfilled. Standards include the clear visibility of anatomic structures as well as a focus on the compression (i.e., the point of interest does not leave the visible area when a compression is performed). In general, their rating achieved a Cohen's Kappa ranging between .21 and .39,

Metric	Condition	*Mdn*	*M*	*SD*	*min*	*max*
Trust	Control	4.42	4.49	0.63	3.75	5.6
	Self-first	4.79	4.58	0.57	3.65	5.29
	Fail-first	4.38	4.46	0.53	3.81	5.31
SIPA	Control	4.27	4.26	0.57	3.17	5.13
	Self-first	4.40	4.52	0.83	3.12	5.75
	Fail-first	3.79	4.01	0.77	2.88	5.08
Usefulness	Control	2.22	2.21	0.54	1.50	3.06
	Self-first	2.28	2.38	0.51	1.81	3.38
	Fail-first	2.41	2.42	0.43	1.31	3.00

Table 1. Descriptive statistics of perceived trust, SIPA and usefulness for each condition across all measurement points.

demonstrating a rather medium inter-rater agreement, which represents a satisfying agreement for the given context [28]. For analysis, R version 4.3.2 [40],including the packages psych [64], targets [27], and tidyverse [63] were used.

4 RESULTS

In this section, we present the results from our experimental study in the order of the research objectives formulated before. We first analysed whether constructs of automation-related user experience differed between conditions and over time, using a repeated-measures ANOVA for each variable. For effect sizes, we report η_g^2, which can be interpreted as small with $\eta_g^2 = 0.01$, as medium with $\eta_g^2 = 0.06$, and as large with $\eta_g^2 = 0.14$, [26].

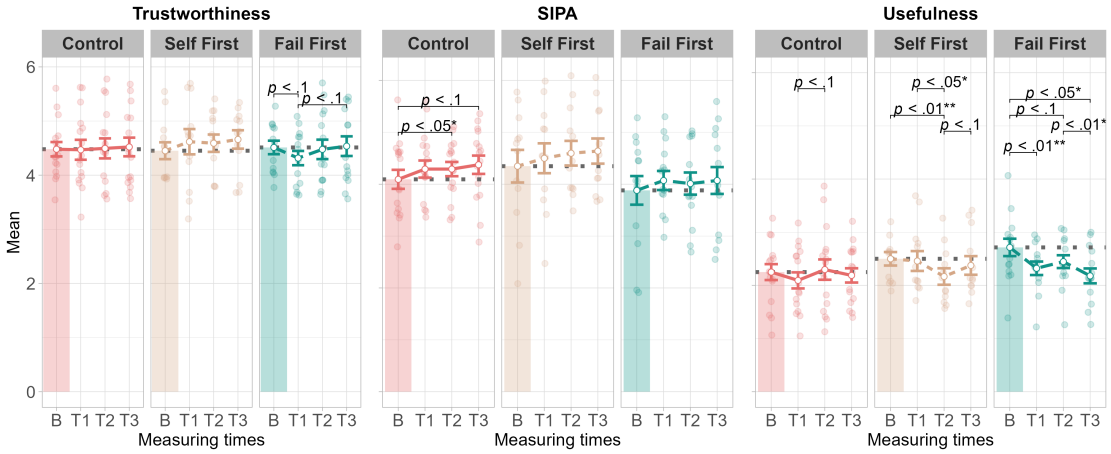
Regarding RO2, we did not find any significant differences in automation-related user experience. There were no effects on trust with $F(2, 39) = 0.15$, $p = .864$, $\eta_g^2 = .006$ for a main effect of condition, $F(3, 117) = 1.03$, $p = .383$, $\eta_g^2 = .004$ for an effect over time, and $F(6, 117) = 0.88$, $p = .515$, $\eta_g^2 = .007$ for a possible interaction of time and condition.

Our data showed no significant difference for for SIPA with $F(2, 39) = 1.65$, $p = .206$, $\eta_g^2 = .060$ for a main effect of condition, $F(3, 117) = 2.25$, $p = .086$, $\eta_g^2 = .014$ for an effect over time and $F(6, 117) = 0.09$, $p = .997$, $\eta_g^2 = .001$ for a possible interaction of time and condition.

While we did not find a main effect of condition in usefulness with $F(2, 39) = 0.74$, $p = .482$, $\eta_g^2 = .027$, we found significant differences in point of time with $F(3, 117) = 4.10$, $p = .008$, $\eta_g^2 = .026$ and interaction with $F(6, 117) = 2.82$, $p = .013$, $\eta_g^2 = .036$. As can be seen in Figure 5c, this result was mainly caused by the Fail-first group. Post-hoc tests are depicted in Fig. 5c.

Next, we analysed the quality of diagnostic assessment, referring to RO3. Again, a repeated measures ANOVA was conducted. With $F(2, 39) = 0.18$, $p = .838$, $\eta_g^2 = .004$ we did not find a main effect of condition, with $F(3, 117) = 0.64$, $p = .638$, $\eta_g^2 = .009$ also no effect of time, and also with $F(6, 117) = 0.31$, $p = .307$, $\eta_g^2 = .037$ no interaction of time and condition. While the ANOVA did not reveal a significant improvement over time, visual analysis of Figure 7 indicated a rise in the Fail-first condition. A post-hoc test revealed a nearly significant difference between the baseline and the final assessment with $t(11) = -2.06$, $p = .064$.

For user confidence as variable of RO4, we found no difference between conditions $F(2, 39) = 0.33$, $p = .718$, $\eta_g^2 = .013$, but did find a significant difference in time with $F(2, 39) = 14.83$, $p < .001$, $\eta_g^2 = .083$. Our results do not support an interaction between time and condition with $F(2, 39) = 0.70$, $p = .648$, $\eta_g^2 = .009$. Significance levels of Post-Hoc tests can be found in Fig. 6.



(a) Development of percevied trust ratings **(b)** Development of SIPA ratings **(c)** Development of perceived usefulness rating

Fig. 5. Development of perceived trustworthiness, SIPA and perceived usefulness over time for each condition. Significance levels of post-hoc tests within conditions (i.e., over time) are shown when $p < .010$.

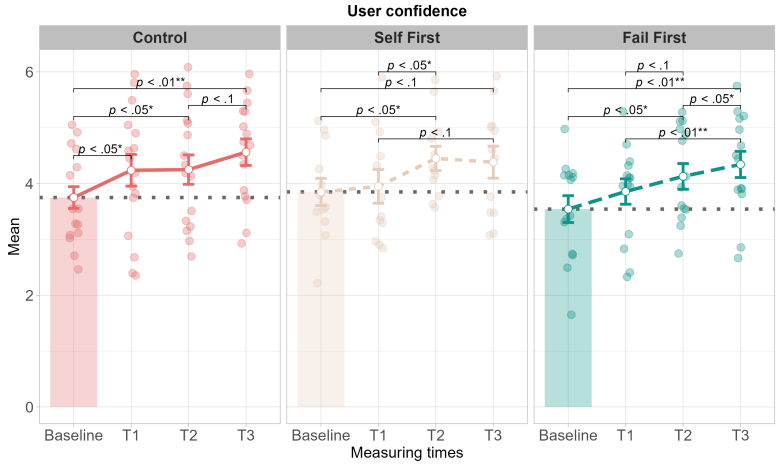


Fig. 6. Development of confidence over time for each condition. Significance levels of post-hoc tests within conditions (i.e., over time) are shown when $p < .010$.

Metric	Condition	*Mdn*	*M*	*SD*	*min*	*max*
Quality of diagnostic assessment	Control	3.00	3.00	0.58	2.00	4.00
	Self-first	3.00	2.98	0.56	2.00	4.00
	Fail-first	3.00	2.86	0.67	1.50	4.00

Table 2. Descriptive statistics of the quality of diagnostic assessment for each condition across all measurement points.

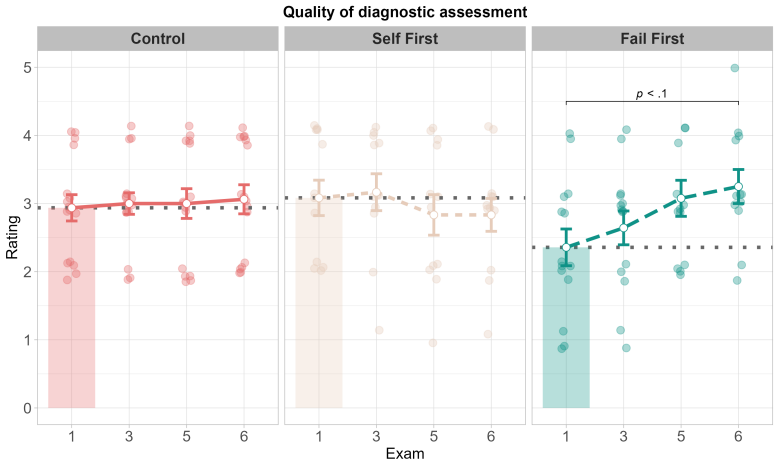


Fig. 7. Development of medical diagnostic quality over time for each condition. Significance levels of post-hoc tests within conditions (i.e., over time) are shown when $p < .010$.

Metric	Condition	*Mdn*	*M*	*SD*	*min*	*max*
Confidence	Control	4.30	4.20	0.85	3.05	5.70
	Self-first	4.05	4.16	0.79	3.00	5.75
	Fail-first	4.00	3.97	0.78	2.35	5.25

Table 3. Descriptive statistics of user confidence for each condition across all measurement points.

Concluding, we analysed how different instructions affect the relationship between usefulness and SIPA (RO5). That is, a high correlation indicates that high levels of perceived traceability, operationalized by SIPA, occur with high levels of perceived usefulness. A graphical representation as well as all statistics can be found in Fig. 8. For the baseline questionnaire, we found significant correlations for the control condition, the self-first condition and the fail-first condition. On the contrary, we only found significant correlations for control and self-first condition at T3.

5 DISCUSSION

In this work, we examined how instructions focusing on informational interdependence affect diagnostic tasks in a human-AI cooperation concerning compression-based DVT diagnosis. We first identified information that was exclusive to human users but nevertheless affected the system's information processing and developed instructions that focused on that information (RO1). Instructions aiming to improve users' understanding of how pressure affects information processing did not affect automation-related user experience, i.e., we did not find any differences in terms of perceived trustworthiness, subjective information processing awareness, or perceived usefulness between conditions (RO2). However, our analysis revealed that a correction of perceived usefulness was strongest when users were first instructed to apply incorrect levels of pressure. While we did not find higher levels of diagnostic accuracy based on conditions, our results indicate that letting users experience failure early on showed the highest improvement in diagnostic quality (RO3). Reported confidence did not differ between groups (RO4). Interestingly, the relationship between confidence and

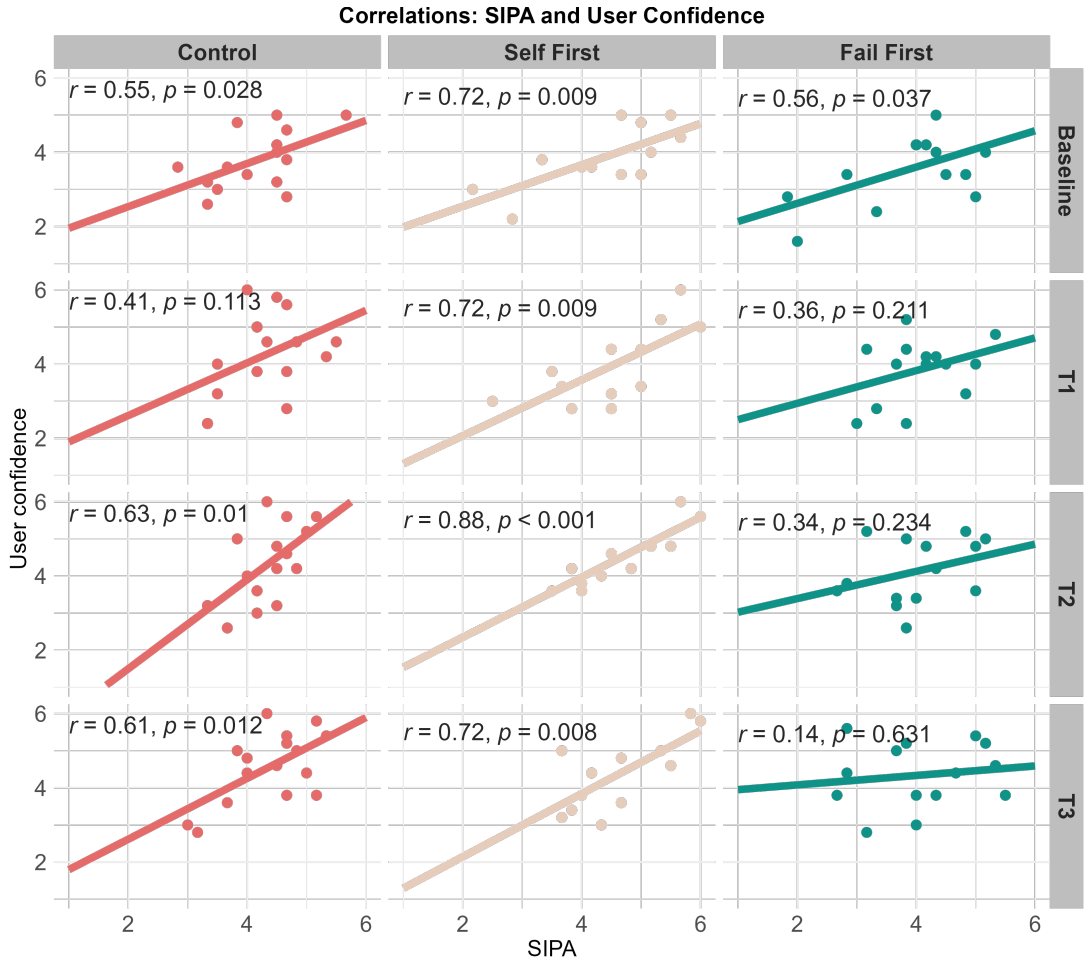


Fig. 8. Calculation and plots of the correlations of SIPA and user confidence for all conditions at all measurement times.

SIPA seems to be dependent on the condition. Although our sample was too small to reveal statistically significant effects when comparing correlation coefficients, our results give some indication that experiencing failure early weakens the relationship between SIPA and perceived confidence (RO5).

5.1 Information Relevance Communication for Novices in Human-AI Cooperation

Our results can be condensed into two key points: first, actively introducing an error early on has a negative effect on the correlation between one's own confidence and the perceived traceability of the system. It is possible that this early error perception strengthens the sense of agency or responsibility. With a better mental model of one's own influence on the result, confidence is less dependent on the understanding of the system. Previous studies have already shown that better mental models in team situations can improve performance [33] and lead to a better understanding of roles [44]. That is, users' increase in confidence in fail-first condition may be related to the fact that they increasingly

adapted the task to provide high quality diagnostic images and optimally perform the compression, resulting in an improved human-AI collaboration.

Second, our study only constitutes a first exploration to address the communication of informational interdependence in diagnostic tasks with AI to improve mental models. Based on our results, we propose to continue using this approach in the future: in particular, when users' mental models of information processing are incorrect, approaches like the one we present can be followed. We argue that experiencing failure early on may weaken the relationship between subjective information processing awareness and confidence. However, participants in this condition had the strongest improvement in diagnostic quality. Hence, an early experience of system failure may lead to the impulse to better understand one's own role and constant performance improvement. Since novices are often even less able to monitor and control the system themselves, this applies in particular to them as a target group.

Furthermore, we have systematically tried to identify gaps in information processing by defining which information is available exclusively to a partner, human, or AI. This allowed us to select information for which the system cannot monitor the quality provided by the human. This can lead to a first-failure effect [32] in the real-world use of systems that still work well in training. To counter this effect, users are actively instructed to provoke a failure. This can dampen potential first-failure effects. In general, we suggest identifying informational interdependencies and considering them as seams when designing the system and training.

5.2 Contribution to XAI in Medical Systems

Legal frameworks, such as the EU AI Act [54] or ethical requirements for the transparency of AI systems [36], could lead to increased use of XAI methods in medical diagnostic systems in the future. In the sense of human-centered development of AI systems [50], however, it should always be defined what effect explanations should have and how the higher controllability or transparency of the systems should be ensured. In studies on human-centered XAI systems, variables of automation-related user experience are regularly used (e.g., [34, 41, 43]). We also integrated these into the present study with perceived trustworthiness, SIPA, and usefulness, but could not determine any differences.

The lack of correlation between automation-related user experience and performance has already been shown by existing studies [10, 45] and can be described as the illusion of explanatory depth [11]. In this case, although the automation-related user experience increases, the performance, e.g. in the prediction of the AI system, does not. In our case, we see the opposite example: in the 'fail first' condition, performance increases, but perceived trustworthiness and SIPA do not. Only perceived usefulness decreases in the fail-first condition. For the use of UX measurement instruments, it should be clear that an increase in performance is not necessarily reflected in the UX experience.

At the same time, the present study sets a task for the development of XAI systems, since revealing informational interdependence can be seen as a subgoal of human-centered AI systems. While only training measures were examined as an intervention in the present example, technical solutions could also be considered. For example, a simulation could be used to show how information exclusively available to the user affects system performance. Overall, explanations based on interdependency should be pursued as a goal of XAI approaches and should be increasingly designed for integrated information processing.

5.3 Limitations & Future Work

When considering our results, it is important to account for the characteristics of the sample, particularly that the participants were novices. For more experienced individuals, the impact of informational interdependence may be reduced due to their advanced skills and familiarity with the specific diagnostic process [52]. This suggests that the

findings might not generalize well to expert populations. To address this, future research should include participants with varying levels of expertise to explore whether experience moderates the effects of informational interdependence on diagnostic performance and user experience.

Additionally, our study's task involved a high degree of interconnected and sequential information processing between humans and AI, which is not typically representative of all diagnostic processes. In fields such as radiology or MRI, the acquisition and analysis of information are often more distinct and separated, making the emphasis on interdependencies potentially less applicable. Future studies should examine the role of informational interdependence in a wider array of diagnostic contexts, including those with more discrete stages of human and AI involvement, to assess the broader relevance and applicability of our approach.

Another limitation is the variability in the baseline knowledge and skills of participants, which could have influenced the results. This variability suggests that a replication of the study with more standardized baseline conditions would be valuable to confirm the robustness of the findings. Moreover, exploring interventions that can account for or adapt to individual differences in initial skill levels may help tailor instructional strategies to better support users across a range of competencies.

Future research should also explore alternative methods for enhancing user understanding of informational interdependence beyond instructional interventions. For example, implementing interactive tools or simulations that visually represent the flow and impact of human-specific information on AI decision-making could provide deeper insights and foster more effective human-AI cooperation. Additionally, studies could investigate the potential of real-time feedback mechanisms that adapt to user performance and highlight interdependencies dynamically, thereby enhancing learning and system transparency.

Finally, it is crucial to consider how these findings translate into practical applications within the design of human-AI systems, particularly in medical diagnostics. While our study focused on novice users and specific task designs, future work should examine how these concepts can be integrated into the development of AI systems that support both novice and expert users. This could involve designing AI systems that not only explain their reasoning but also actively involve users in the decision-making process by highlighting their unique contributions, thereby reinforcing the collaborative nature of human-AI partnerships.

6 CONCLUSION

In this study, we set out to examine how novice users of AI-based DSS in diagnosis can be supported by instruction reflecting interdependencies in Human-AI collaboration. By designing an experiment that emphasized the informational interdependence between human actions and AI outputs, we demonstrated the pivotal role that early encounters with system limitations and failures can play in fostering a deeper understanding of system dynamics. Our approach involved to understand which gaps between users self-concept of information processing and their role in a shared information processing between human and AI needs to be highlighted. Here, we utilized seamful design to develop instructions revealing how user action may lead to system failure. In an experiment, we demonstrated how instruction can affect diagnostic quality in AI-based DSS systems, even though automation-related UX was largely unaffected. All in all, our research emphasizes designers to design for informational interdependence between human actions and AI outputs and offers a promising alternative to traditional explanatory frameworks used in XAI.

7 ACKNOWLEDGEMENTS

We used LLM-based support (ChatGPT 4.0 from OpenAI) to provide description of all figures and tables to improve accessibility.

REFERENCES

[1] Ahmed Al Kuwaiti, Khalid Nazer, Abdullah Al-Reedy, Shaher Al-Shehri, Afnan Al-Muhanna, Arun Vijay Subbarayalu, Dhoha Al Muhanna, and Fahad A Al-Muhanna. 2023. A review of the role of artificial intelligence in healthcare. *Journal of personalized medicine* 13, 6 (2023), 951.

[2] Frederick A Anderson, H Brownell Wheeler, Robert J Goldberg, David W Hosmer, Nilima A Patwardhan, Borko Jovanovic, Ann Forcier, and James E Dalen. 1991. A population-based perspective of the hospital incidence and case-fatality rates of deep vein thrombosis and pulmonary embolism: the Worcester DVT Study. *Archives of internal medicine* 151, 5 (1991), 933–938.

[3] Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science* 24, 2 (2023), 129–175.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[5] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2023. On Selective, Mutable and Dialogic XAI: a Review of What Users Say about Different Types of Interactive Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 411, 21 pages. https://doi.org/10.1145/3544548.3581314

[6] Sebastian Bruckert, Bettina Finzel, and Ute Schmid. 2020. The next generation of medical decision support: A roadmap toward transparent expert companions. *Frontiers in artificial intelligence* 3 (2020), 507973.

[7] Charles S Carver and Michael F Scheier. 1985. A control-systems approach to the self-regulation of action. In *Action control: From cognition to behavior*. Springer, 237–265.

[8] Matthew Chalmers. 2003. Seamful design and ubicomp infrastructure. In *Proceedings of Ubicomp 2003 workshop at the crossroads: The interaction of HCI and systems issues in Ubicomp*. 577–584.

[9] Po-Hsuan Cameron Chen, Krishna Gadepalli, Robert MacDonald, Yun Liu, Shiro Kadowaki, Kunal Nagpal, Timo Kohlberger, Jeffrey Dean, Greg S. Corrado, Jason D. Hipp, Craig H. Mermel, and Martin C. Stumpe. 2019. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature Medicine* 25, 9 (Aug. 2019), 1453–1457. https://doi.org/10.1038/s41591-019-0539-7

[10] Michael Chromik and Andreas Butz. 2021. Human-XAI interaction: a review and design principles for explanation user interfaces. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18*. Springer, 619–640.

[11] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 307–317.

[12] Clarius Mobile Health. 2024. Clarius C7 Scanner. https://clarius.com/scanners/c7/ Accessed: 2024-09-13.

[13] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–19.

[14] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daumé III. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–29.

[15] Feng-Lei Fan, Jinjun Xiong, Mengzhou Li, and Ge Wang. 2021. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences* 5, 6 (2021), 741–760.

[16] Samuel Z Goldhaber and Henri Bounameaux. 2012. Pulmonary embolism and deep vein thrombosis. *The Lancet* 379, 9828 (2012), 1835–1846.

[17] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS* (2021), 78.

[18] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. 2014. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction* 3, 1 (2014), 43–69.

[19] Bernhard Kainz, Mattias P Heinrich, Antonios Makropoulos, Jonas Oppenheimer, Ramin Mandegaran, Shrinivasan Sankar, Christopher Deane, Sven Mischkewitz, Fouad Al-Noor, Andrew C Rawdin, et al. 2021. Non-invasive diagnosis of deep vein thrombosis from ultrasound imaging with machine learning. *NPJ Digital Medicine* 4, 1 (2021), 137.

[20] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.. In *IJCAI*. 4070–4073.

[21] Emeka Kesieme, Chinenye Kesieme, Nze Jebbin, Eshiobo Irekpita, and Andrew Dongo. 2011. Deep vein thrombosis: a clinical review. *Journal of blood medicine* (2011), 59–69.

[22] Levi Kitchen, Matthew Lawrence, Matthew Speicher, and Kenneth Frumkin. 2016. Emergency department management of suspected calf-vein deep venous thrombosis: a diagnostic algorithm. *Western Journal of Emergency Medicine* 17, 4 (2016), 384.

[23] Gary Klein, Paul J Feltovich, Jeffrey M Bradshaw, and David D Woods. 2005. Common ground and coordination in joint activity. *Organizational simulation* 53 (2005), 139–184.

[24] Yogesh Kumar, Apeksha Koul, Ruchi Singla, and Muhammad Fazal Ijaz. 2023. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of ambient intelligence and humanized computing* 14, 7 (2023), 8459–8486.

[25] Paul A Kyrle and Sabine Eichinger. 2005. Deep vein thrombosis. *The Lancet* 365, 9465 (2005), 1163–1174.

[26] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology* 4 (2013). https://doi.org/10.3389/fpsyg.2013.00863

[27] William Michael Landau. 2021. The targets R package: a dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software* 6, 57 (2021), 2959. https://doi.org/10.21105/joss.02959

[28] J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* (1977), 363–374.

[29] Retno Larasati, Anna De Liddo, and Enrico Motta. 2023. Meaningful Explanation Effect on User's Trust in an AI Medical System: Designing Explanations for Non-Expert Users. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 30 (dec 2023), 39 pages. https://doi.org/10.1145/3631614

[30] William F Lawless, Ranjeev Mittu, Don Sofge, and Laura Hiatt. 2019. Artificial intelligence, autonomy, and human-machine teams—interdependence, context, and explainable AI. *Ai Magazine* 40, 3 (2019), 5–13.

[31] R Mani, F Regan, J Sheridan, and V Batty. 1995. Duplex ultrasound scanning for diagnosing lower limb deep vein thrombosis. *Dermatologic surgery* 21, 4 (1995), 324–326.

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

[33] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. 2000. The influence of shared mental models on team process and performance. *Journal of applied psychology* 85, 2 (2000), 273.

[34] Nathan J McNeese, Mustafa Demir, Nancy J Cooke, and Manrong She. 2021. Team situation awareness and conflict: A study of human–machine teaming. *Journal of Cognitive Engineering and Decision Making* 15, 2-3 (2021), 83–96.

[35] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2023. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies* 169 (2023), 102941.

[36] Sidra Nasir, Rizwan Ahmed Khan, and Samita Bai. 2024. Ethical framework for harnessing the power of ai in healthcare and beyond. *IEEE Access* 12 (2024), 31014–31035.

[37] Laurence Needleman, John J Cronan, Michael P Lilly, Geno J Merli, Srikar Adhikari, Barbara S Hertzberg, M Robert DeJong, Michael B Streiff, and Mark H Meissner. 2018. Ultrasound for lower extremity deep venous thrombosis: multidisciplinary recommendations from the Society of Radiologists in Ultrasound Consensus Conference. *Circulation* 137, 14 (2018), 1505–1515.

[38] Raja Parasuraman and Christopher A Miller. 2004. Trust and etiquette in high-criticality automated systems. *Commun. ACM* 47, 4 (2004), 51–55.

[39] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30, 3 (2000), 286–297.

[40] R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[41] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence* (2023).

[42] Lindsay Sanneman and Julie A Shah. 2022. The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. *International Journal of Human–Computer Interaction* 38, 18-20 (2022), 1772–1788.

[43] Lindsay Sanneman, Mycal Tucker, and Julie A Shah. 2024. An Information Bottleneck Characterization of the Understanding-Workload Tradeoff in Human-Centered Explainable AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2175–2198.

[44] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's think together! Assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–29.

[45] Tim Schrills and Thomas Franke. 2023. How do users experience traceability of AI systems? Examining subjective information processing awareness in automated insulin delivery (AID) systems. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 1–34.

[46] Tim Schrills, Lilian Kojan, Marthe Gruner, André Calero Valdez, Thomas Franke, et al. 2024. Effects of User Experience in Automated Information Processing on Perceived Usefulness of Digital Contact-Tracing Apps: Cross-Sectional Survey Study. *JMIR Human Factors* 11, 1 (2024), e53940.

[47] Tim Schrills, Marvin Sieger, Marthe Gruner, and Thomas Franke. 2024. Evaluation of a Scale to Assess Subjective Information Processing Awareness of Humans in Interaction with Automation amp; Artificial Intelligence. In *Artificial Intelligence and Social Computing (AHFE)*. AHFE International. https://doi.org/10.54941/ahfe1004640

[48] Ariel L Shiloh, Christa McPhee, Lewis Eisen, Seth Koenig, and Scott J Millington. 2020. Better with ultrasound: detection of DVT. *Chest* 158, 3 (2020), 1122–1127.

[49] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[50] Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

[51] Timo Speith, Barnaby Crook, Sara Mann, Astrid Schomäcker, and Markus Langer. 2024. Conceptualizing understanding in explainable artificial intelligence (XAI): an abilities-based approach. *Ethics and Information Technology* 26, 2 (2024), 40.

[52] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 109–119.

[53] ThinkSono Ltd. 2024. *ThinkSono*. http://thinksono.com Accessed: 2024-09-13.

[54] André Calero Valdez, Moreen Heine, Thomas Franke, Nicole Jochems, Hans-Christian Jetter, and Tim Schrills. 2024. The European Commitment to Human-Centered Technology: The Integral Role of HCI in the EU AI Act's Success. *arXiv preprint arXiv:2402.14728* (2024).

[55] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial intelligence* 291 (2021), 103404.

[56] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[57] Qi Wang, Lili Yuan, Xianhui Ding, and Zhiming Zhou. 2021. Prediction and diagnosis of venous thromboembolism using artificial intelligence approaches: A systematic review and meta-analysis. *Clinical and Applied Thrombosis/Hemostasis* 27 (2021), 10760296211021162.

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

[59] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do You Trust Me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 7–9. https://doi.org/10.1145/3308532.3329441

[60] Philip S Wells, Melissa A Forgie, and Marc A Rodger. 2014. Treatment of venous thromboembolism. *Jama* 311, 7 (2014), 717–728.

[61] Christopher D Wickens. 1999. Automation in air traffic control: The human performance issues. *Automation technology and human performance: current research and trends* (1999), 2–10.

[62] Christopher D Wickens, Benjamin A Clegg, Alex Z Vieane, and Angelia L Sebok. 2015. Complacency and automation bias in the use of imperfect automation. *Human factors* 57, 5 (2015), 728–739.

[63] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4, 43 (2019), 1686. https://doi.org/10.21105/joss.01686

[64] William Revelle. 2024. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. https://CRAN.R-project.org/package=psych R package version 2.4.3.

[65] Meina Zhang, Linzee Zhu, Shih-Yin Lin, Keela Herr, Chih-Lin Chi, Ibrahim Demir, Karen Dunn Lopez, and Nai-Ching Chi. 2023. Using artificial intelligence to improve pain assessment and pain management: a scoping review. *Journal of the American Medical Informatics Association* 30, 3 (2023), 570–587.

8 General Discussion

8.1 Summary of Results

The objective of the present dissertation was to examine how the information displayed by an AI system in diagnostic tasks that integrate human and machine information processing affects automation-related user experience and behavior. The dissertation is structured around three empirical studies, each addressing different aspects of interaction between human and AI and applying different methods of empirical research (i.e., experimental study with users, structural modeling and design of a paradigm to study HAI).

The first study examined the importance of automation-related user experience for the intention to use automated digital contact tracing. The results highlight the critical role of result diagnosticity — the ability of a system to help users distinguish between different outcomes and make a decision about their next action. In addition, the findings put forward a conceptualization of human-automation interaction grounded in cybernetic control theory, where two loops of information processing - human and machine - are integrated. The main conclusion of this study is that the quality of integration of human and machine information processing is an important driver for the intention to use automated systems.

In the second study, the relationship between information disclosure, intended to provide transparency and users' subjective information processing awareness as an integral part of automation-related user experience was examined. The study found that while SIPA validly captured higher experiences of transparency, it did not enhance users' ability to predict system behavior. Moreover, increased information disclosure did not lead to higher engagement (i.e., longer time spent interacting

with the system). This suggests that simply providing more information may be insufficient to improve users' predictive capabilities or engagement levels which contrasts their increased experience of transparency.

The third study introduced a new research paradigm for integrated information processing, drawing on existing research frameworks in the field of explainable AI, called Kandinsky patterns (Müller & Holzinger, 2021). The results did not confirm the presence of a question behavior effect, where the act of assessment itself influences behavior. However, the study revealed that instructed reliability — defined as a measure of user confidence in the reliability of a system — significantly predicted user behavior in diagnostic tasks involving shared information processing. This finding emphasizes the importance of communicated system reliability in shaping user reliance and interaction patterns.

The fourth study was able to demonstrate that instructions adressing the interdependency between humans and AI can have a positive effect on the use of AI systems. In this study with novices in medicine, an improvement in performance was shown when using an ultrasound-based AI diagnostic system with instruction reflecting the human role in information processing. At the same time, it was also shown that the relationship between SIPA and individuals' action confidence could be altered by information-related instructions. Overall, the findings of study 4 underline the need to systematically examine human and machine roles in information processing and to disclose them, e.g., in the context of instructions.

Overall, the empirical results of this research underscore the necessity of focusing on human information processing tasks and theoretical work on how explanations influence human information processing. The findings indicate that XAI research must be grounded in a model of integrated information processing that can directly inform design, linking specific design features to their effects on information processing but also allowing to examine users' action regulation when interacting with intelligent systems.

By exploring these dynamics, the empirical research presented as part of this dissertation contributes to the development of human-centered AI systems that are not only technically proficient and have the potential to explain their decisions but are

also aligned with the psychological needs of their users, thereby enhancing usability and automation-related user experience.

Based on the results of the studies and the state of the relevant literature at the time this dissertation was written, the following points will be addressed in more detail in the discussion: (1) the importance of the diagnosticity of information (e.g., an explanation) and the perception of diagnosticity should be central to the design of explanations and require an understanding of human information processing in automation contexts. (2) The integration of human and machine information processing can be supported, e.g., by explanations or interactions. A model of integrated information processing can support interdisciplinary research from HCI, engineering psychology, and AI. (3) The discrepancy between human experience, the associated self-rated user experience, and the observation of (performance-related) behavior poses a risk for the use of XAI systems, as users might be led to act under wrong impressions (i.e., about their ability to effectively control and interact with an XAI system). Since explanations can potentially enhance the risk of information overload, it is necessary to investigate the effect of explanations under different levels of available cognitive resources.

8.2 The Underestimated Role of Diagnosticity

The aim of a diagnosis - and therefore the aim of diagnostic tasks - is to select a correct hypothesis (e.g., identify a disease, select a strategy, or localize a fault). This means that information is used to assess the probability of a hypothesis being correct (see Wickens and Scott, 1983). Diagnosticity describes the extent to which evidence changes the relative probabilities of different hypotheses. It can be expressed as a likelihood ratio (Nelson, 2005). At the beginning of this dissertation, the term diagnosticity (and in particular perceived diagnosticity) was still little associated with XAI, as the negative effects of XAI only became better known through corresponding publications. The focus on diagnosticity as a benchmarking variable in XAI systems thus represents a central theoretical contribution of this work. Since diagnosticity is discussed in all studies that are part of this dissertation, this concept is discussed more intensively in this section.

In medical diagnosis, diagnosticity plays a crucial role. For instance, when diagnosing whether a patient has the flu or a common cold, a doctor might ask, 'Do you have a fever?' Suppose that among 100 patients with the flu, 80 have a fever, while only 10 out of 100 patients with a common cold have a fever. This high diagnosticity means that the presence of a fever significantly increases the likelihood of the flu compared to the common cold. The likelihood ratio in this scenario would demonstrate how much this evidence shifts the relative probabilities of the two hypotheses. In the first study presented in this dissertation, users expressed the desire for more detailed information about potential contacts during a pandemic, i.e., whether they wore a mask or not. That is, they needed the contact tracing to present information with higher diagnosticity.

Diagnosticity is fundamentally about how evidence affects the probabilities of different hypotheses. In the previous example, the presence of a fever shifts the probability towards the flu hypothesis and away from the common cold hypothesis. This shift helps in making a more informed decision and potentially affects confidence in a positive way. From a psychological perspective, diagnosticity is important because it helps in selecting an action, that is, it is integral for a successful output function. If a symptom significantly favors one hypothesis over another, it guides, e.g., a healthcare professional towards a specific diagnosis and subsequent treatment, rather than leaving the probabilities of both hypotheses close to each other, which would result in uncertainty and indecision.

As diagnosticity as a term is not used often in the discussion of XAI so far (e.g., key works like T. Miller, 2019 or Shneiderman, 2020a do not discuss it), it is important to distinguish it from close concepts discussed in the literature, e.g., information value, as they may seem to be almost the same but affect HAI differently. Information value is connected to the general concept of entropy. Entropy, in the context of diagnostic decisions and information, represents uncertainty. For instance, asking about fever is in general a valuable question because it is related to many diseases and can significantly reduce uncertainty. In contrast, asking about a symptom like 'the presence of a specific skin rash' might help distinguish between two specific diseases, such as measles and rubella, but is otherwise less informative and doesn't generally reduce overall entropy in a diagnostic context.

Thus, whether to focus on information gain or entropy in comparison to diagnosticity might be context-dependent. In scenarios with a wide range of possible diagnoses, reducing overall entropy (information gain) might be more effective for a user's task. However, in more targeted diagnostic tasks where the goal is to distinguish between a few specific hypotheses, diagnosticity might be favored. The integration of diagnostic tasks into standard operating procedures or diagnostic processes allows for the prioritization of information gain or diagnosticity based on the number of available hypotheses. For instance, in a setting where quick and accurate differentiation between a few common hypotheses is needed, diagnosticity can streamline the decision-making process and is therefore what AI systems should support when presenting information.

As already discussed with SIPA in the present dissertation (and discussed in the concept of the illusion of explanatory depth, Chromik et al., 2021), there can be a difference between perceived diagnosticity and actual diagnosticity. This means that the degree to which information changes the relative probability of two hypotheses is underestimated or overestimated (cf. definition from Nelson, 2005). This can be the case, e.g., if information with a high information gain, regardless of the specific hypotheses, is assessed. Let's assume that a distinction needs to be made between a COVID-19 infection and the flu: in this case, information about fever does not represent information with a high diagnostic value, although fever is general information that can result in a high information gain when diagnosing illnesses. E.g a so-called pseudo-diagnostic decision can arise in which people select diagnostically worthless data in an opinion revision task and then to make a judgment based on those data (Kern & Doherty, 1982). Similar to the illusion of explanatory depth, this can lead to (critical) misjudgments of one's own performance and in the evaluation of the system. In particular, results from the second study of this dissertation demonstrate that users' perception of a system can significantly differ, but their ability to predict system information processing does not increase. According to Speith et al. (2024) who argue that understanding is expressed in a user's abilities, it is important to distinguish between users' perception and understanding of diagnosticity, and their ability to utilize diagnostic information.

Although diagnosticity is a central characteristic of information in a diagnostic task,

the diagnosticity of information provided by an AI system cannot be equated with the usefulness of the system. As already described in Section 2.3.2, the usefulness of a person also includes other goals that depend on the specific task. For instance, if one assumes that a certain diagnosis made by a physician can be treated by the person in their own practice, but another diagnosis means a (potentially tedious) journey to another practice for the patient. A system can support the weighing-up process by displaying the potential consequences of decisions and thus being useful in the task - at the same time, the diagnostic task is not influenced by this. Thus, diagnosticity contributes to a system's usefulness.

In the third study of this dissertation, the diagnosticity of the system was not analyzed as a variable. However, the exploratory analysis indicated that the instructed reliability may be relevant information in determining the choice of action of the individuals. Due to the lack of integration of machine and human information processing in this study, participants used task-independent information (e.g., on the reliability of the system) to regulate their own behavior. This raises the question of what information persons use to select actions in a human-AI interaction when no diagnostic information is provided by the system. In their work, Bartlett and McCarley (2017) describe various strategies that people can select to solve a task (in the discussed study, deciding which color is prevalent in a visual stimulus). One problematic strategy is probability matching, in which the reliability of the AI system is used to adjust one's own frequency of agreement. Other strategies, e.g., optimal weighing strategy, require better diagnostics of the systems in my view. Overall, this means that the sub-optimal action regulation in the result of study 3 could also be attributed to the lack of diagnosticity of the implemented system.

The fourth study of this dissertation turns the idea of diagnosticity around to a certain extent: the use of seamful design should reveal where a system (and its performance) is particularly dependent on information from users. Within the study, pressure during compression was identified as a variable, i.e. human information that modifies the diagnosticity of the image evaluated by the machine. Diagnosticity was not manipulated here, but rather the assumptions under which diagnostic information can be incorrectly assessed were shown. For the AI system, the remaining volume of the vessel on the ultrasound image has high diagnostic accuracy - assuming that

enough pressure has been applied. Maintaining the diagnosticity of information provided for in the process is therefore a central component in the investigation of informational interdependence.

In summary, when discussing diagnostic tasks, assessing the diagnosticity of information provided by machines for humans is essential when humans are responsible for a diagnostic decision. Accordingly, diagnosticity is a highly relevant construct referred to in all studies presented within this dissertation and discussed across all research objectives.

8.3 Integrated Human-AI Information Processing for Action Regulation

In chapter 2 multiple evolving perspectives on the interaction between humans and AI were presented and the topic of explainable AI as well as metrics of XAI were introduced. However, as discussed in the studies included in this dissertation, a theoretical framework to study the effects of XAI and guide the design of XAI is necessary. The first approaches in this direction have already been discussed: e.g., Yang et al. (2022) discuss theoretical foundation of explanation based on similarity with human explanation. The authors conclude that human users process and utilize a machine-based explanation in reference to a potential explanation they would expect from a person. While this approach can support understanding users' reactions to explanations, it does not specifically help to guide the design and evaluation of HAI. That is, the discussed research is focused on understanding humans' reception of explanations but does not aim to provide a theoretical contribution that addresses both, machine and human processes. In general, previous research demonstrated well how humans generate explanations (e.g., Baehrens et al., 2010; Chin-Parker and Bradner, 2010; Malle, 2022) and how cognitive processes of humans can be emulated to produce effective explanations (see Wilkenfeld and Lombrozo, 2015). However, to the best of my knowledge, no models exist to describe how and when explanations can affect human information processing while using an AI system (in contrast to studies examining the effect of XAI on technical information processing, e.g., Bell

et al., 2022). It is not necessary - and also not the aim of this dissertation - that all approaches labeled as XAI can be covered by such a model. Rather, a model of integrated information processing should support the assessment of XAI's impact on integrated information processing.

All three of the studies conducted as part of this dissertation had one disadvantage for the participants - the results or explanations displayed were not interactive. In Study 1 and 2 in particular, participants criticized the fact that no interaction allowed them to evaluate their own assumptions or hypotheses based on the system outputs. For example, while the information processing model depicted in Study 1 already connects human and machine information processing, it does not conceptualize possible interactions within the information processing. However, in their review, Bertrand et al. (2023) found that interactive explanations are perceived as more useful and may lead to better performance compared to non-interactive explanations but also possibly increase the demand for users' resources (e.g., time spent to understand the interactive explanation). Our results in study 4 also demonstrated that interactive approaches to improve users' understanding of a system may be beneficial in terms of performance. That is, a model of integrated information processing must describe the interaction between human and AI and its influence on shared information processing and informational interdependence.

An integrated approach to human and machine information processing offers significant advantages for experimental research on AI systems. Artificial cognition (see Ritter et al., 2017), a concept derived from applying experimental psychology paradigms to AI, emphasizes the necessity of understanding AI behavior through the same rigorous methods used in human cognitive research. By doing so, AI research can bridge the gap between XAI and empirical research, ensuring that the results meet the stringent requirements of causal argumentation (see T. Miller, 2019).

Artificial cognition (J. E. T. Taylor & Taylor, 2021) involves using psychological theories and methodologies to interpret and predict AI behavior, making AI systems more transparent and trustworthy. This approach helps in validating AI decisions through empirical studies that mimic human cognitive processes. For instance, when AI systems are exposed to similar experimental conditions as human participants, it becomes easier to draw parallels and understand the underlying mechanisms driving

AI decisions (see Ritter et al., 2017).

In the following - in view of research objective 5 - a conceptual model of the integrated information processing (the *Integrated Information Processing Model*, IIP) of humans and automated systems is presented. This model combines both the existing model of information processing by Parasuraman et al. (2000) and models of human action regulation by, e.g., Carver and Scheier (2000). In addition, it takes into account results on human-AI interaction from the field of explainable AI. It is shown in Fig. 8.1.

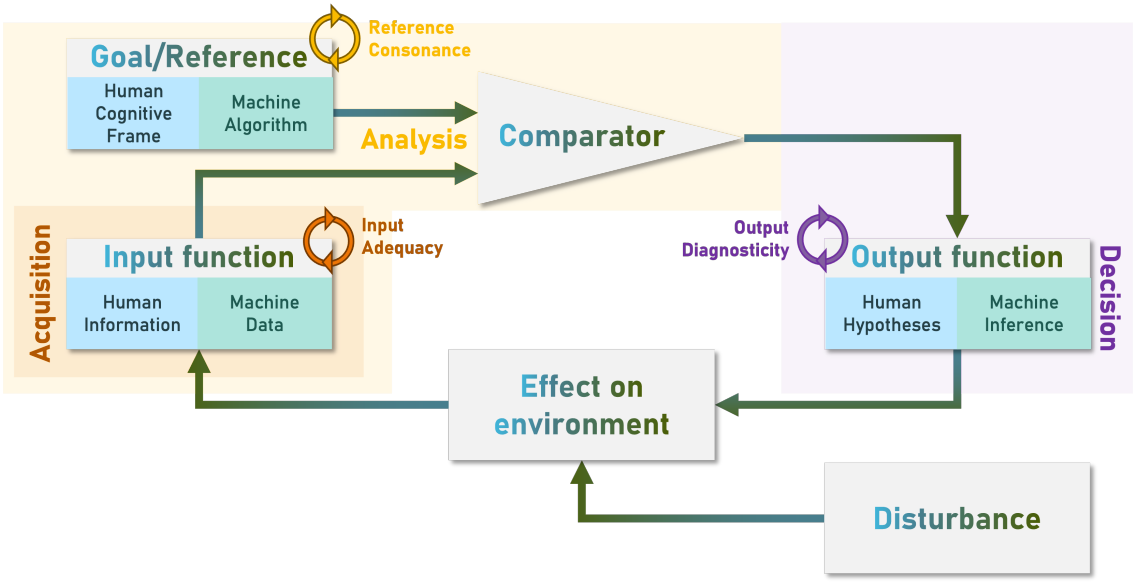


Figure 8.1: The conceptual model of integrated information processing

8.3.1 Integration of Automated Information Processing in Diagnosis

The basic structure of the model is based on the cybernetic control of human and machine actions in the context of information processing. Cybernetic control loops, as articulated by Carver and Scheier (2000), are fundamental mechanisms in behavioral self-regulation. These loops consist of four core components: the input function (perception), the reference value (goal), the comparator, and the output function (behavior). The comparator continuously assesses the input against the reference

value, triggering adjustments in the output function when discrepancies are detected. Two primary types of feedback loops are identified: negative feedback loops, which aim to reduce discrepancies between input and reference, and positive feedback loops, which work to increase discrepancies from a negative reference point.

Based on Carver and Scheier (2000), the concept of behavior is not merely a sequence of actions; rather, it is a dynamic process of maintaining alignment with goals through constant adjustments based on feedback and updated information from the environment. This feedback mechanism is of critical importance for both achieving desired outcomes and avoiding undesired states. The nested structure of cybernetic control loops allows for complex, multi-layered regulation, whereby higher-level goals set the standards for lower-level behaviors. Thereby nested control loops create a hierarchical system of control (Carver & Scheier, 2000). From my point of view, an understanding of the structure of cybernetic control is fundamental to grasping the manner in which human and AI systems interact and regulate actions within a shared information processing task. Consequently, cybernetic control is addressed in studies 1 and 2 of the present dissertation. The following section will elaborate on how this framework supports integrated information processing in dynamic environments.

The selection of this basic structure of cybernetic control takes into account the fact that integrated information processing is not a single, sequential process, but a circular regulatory process of actions towards a goal. For instance, a diagnosis found can become obsolete due to changes in the environment (e.g., new, available data) requiring the process to be repeated. It is also crucial to monitor processes, as conditions can change due to external factors that are unrelated to the initial information processing. For instance, in the context of diabetes management as described in study 2, the constant monitoring of glucose levels in diabetes can be affected by external variables such as illness and fever. This illustrates the necessity of continuously integrate external factors (i.e., that are introduced into the input function only after a first diagnosis was made and reactions were observed) into the information processing cycle.

Furthermore, Carver and Scheier (2000) posits that cybernetic control loops are nested within each other. This implies, for instance, that there is a higher-level loop that is concerned with maintaining overall health, with nested loops that are

dedicated to obtaining specific information about one's health status. This nesting also permits us to examine the interactive components of integrated information processing. For instance, the exchange of information between humans and machines can be described in a separate loop that is on another level of action regulation compared to the overall information processing. The nested loops for input, reference and output functions are an indispensable component of this model.

Lastly, the cybernetic action regulation model may incorporate the information processing model proposed by Parasuraman and colleagues in the context of diagnostic procedures. The initial three phases of the information processing cycle can be situated within the context of action regulation. The initial stage is the acquisition of information (stage 1), which is addressed by the machine or human input function. Subsequently, the information is subjected to analysis, whereby the reference function is compared with the input function. The selection of a decision (stage 3) can be described in the context of the output function, whereby a decision is selected. In the event that a direct action is also linked to the diagnosis, the action implementation (stage 4) is also addressed as part of the output function.

The model also diverges from existing action regulation models at the highest level, as it incorporates both machine and human entities into the input reference and output functions. Previous models were focused on one actor - the human - while the IIP model is specifically designed to allow the integration of both entities. While there can be more than two entities in general, the model assumes the presence of at least one human and one machine entity. The relationship between these two entities gives rise to three distinct qualities of integrated information processing: *(1) input adequacy, (2) reference consonance, and (3) output diagnosticity.* The relative priority of these three information processing qualities may vary depending on the task and individual. To illustrate, a higher level of input adequacy may be necessary for high-risk tasks in the medical field than compared to recommender systems (see also the description of 'individual standards' in the TRaM model by Schlicker et al., 2022.

The three qualities of integrated information processing are described below, and information processing is illustrated using two examples. A general definition of each quality will be provided, along with an overview of how it relates to the machine and

human states within the information processing. In the context of machine-based processing, the term 'estimated' is used prior to the description of the machine perspective. In contrast, the human perspective is indicated by the term 'perceived'. Subsequently, an examination is conducted of how explanations are (or can be) related to information processing qualities.

8.3.2 Input Adequacy

The first information processing quality concerns the input function or an integrated information acquisition process. 'Input', on the one hand, includes information that a human actor can possess: e.g., how much pain a person is currently experiencing or how much food they have consumed. All information a human is aware of and that is relevant for the given information processing task, e.g., making a diagnosis, is considered part of human information. On the other hand, there is machine data that is available to the automated system, e.g., X-ray images or a current glucose value. For better differentiation, a distinction is made in the model between human information and machine data. The available machine data and human information represent the overall input for information processing, which means that anything that is not part of the human information or machine data is not processed. It is therefore possible, for example, that the input is not sufficient to make a diagnosis, e.g., if there are no values for fever in a medical diagnosis. Furthermore, data could be outdated (e.g., no current glucose value) or not precise enough (e.g., only the indication of whether food was eaten instead of an exact amount of nutrients). All these factors could lead to the input being not adequate for the information processing task. In the fourth study of the present dissertation, informational interdependence could lead to inadequate input, as the system was not able to register how much pressure a user applied in a diagnostic examination. Additionally, the user could not communicate the amount of pressure to the system, which might led to insufficient input adequacy.

In mathematics, for example, adequacy describes the state of a model to be suitable for a given objective: Adequacy can be used to describe the state of a set of information that is sufficient to allow a conclusion (Tedeschi, 2006). **In light of mathematical**

description of adequacy, input adequacy as information processing quality is defined: 'Input adequacy in integrated information processing exists when both entities have in total a sufficient amount of sufficiently correct information to process the task'. Accordingly, perceived input adequacy describes the human 'perception of the extent of the sufficiency of information in terms of completeness and correctness for all entities within the information processing'. In turn, machines may calculate an Estimated Input Adequacy, which is the 'calculation of the extent of the sufficiency of information in terms of completeness and correctness for all entities within the information processing'.

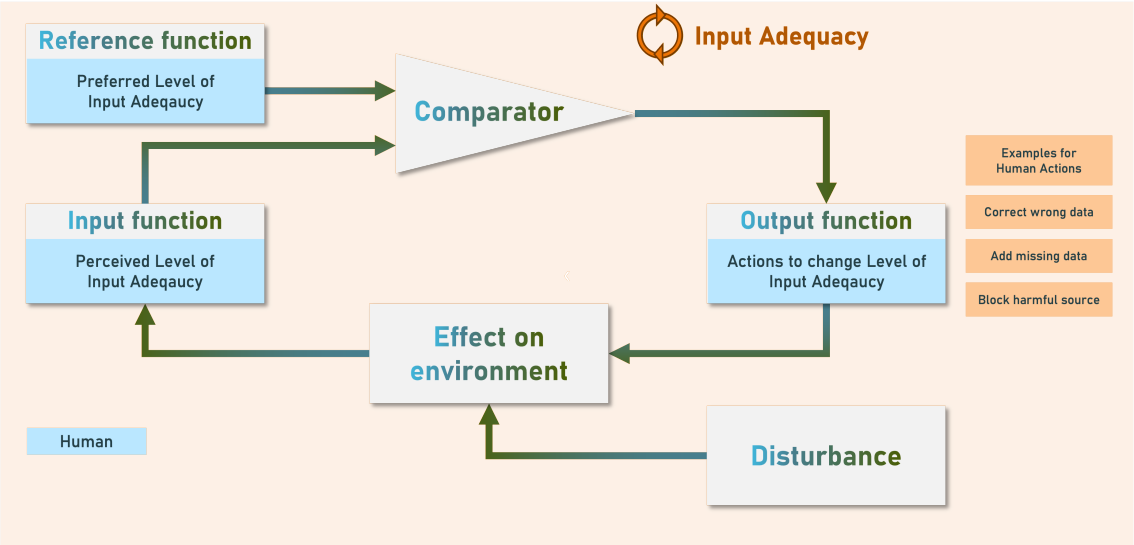


Figure 8.2: Nested loop of human action regulation to reach preferred level of input adequacy

This definition implies that both human and machine may conclude that input adequacy is insufficient, making further action necessary (reflected also in the idea of information asymmetry, see Hemmer et al., 2022). Fig. 8.2 shows the nested loop of input adequacy, in which a comparison is made with the preferred level of input adequacy on the basis of the available information. In the event of deviations, an output can be generated to increase the input adequacy. For example, the input adequacy could be below the human's preferred level because the system does not have the necessary information, e.g., the current glucose value. To correct this, the human could add this information, if possible. Table 8.1 shows examples of different situations with a non-preferred level of information adequacy.

Table 8.1: Examples of low levels of input adequacy with examples from the context of AID systems (referring to study 2)

Perceiving Entity	Target to Improve Input Adequacy	Control Action	Example
Human	Human Information	Request data from machine, acquire information	Human does not know the current level of insulin to calculate insulin demand and requests current insulin level from insulin pump
	Machine Data	Correct, delete or extend machine data, request update from system	AID system has no information about meals, human communicates amount of carbohydrates
Machine	Human Information	Communicate data to human, provide human with list of required information	AID systems informs human user about long period of high glucose levels, which were previously undetected
	Machine Data	Update data / validate sensory input, request input from human	AID system requests information about physical activity before correcting high glucose levels

Once the preferred level of input adequacy has been reached, the person or system does not take any further action to increase the input adequacy.

8.3.3 Reference Consonance

The second quality of IIP concerns the reference function or an integrated information analysis process (i.e., the reference function and the comparator). In the context of simple cybernetic control, the reference function consists of a reference value that is compared with the actual value obtained in the input function (see Carver and Scheier, 2000). In the context of automated information processing, the reference function describes more than a target value: it contains all processes and target values used to analyze the input information. This may be but is not excluded to specific target values (e.g., a specific glucose value), a processing logic (e.g., if-then rules as to whether or not insulin injection is necessary), or a neural network, that processes glucose values. Hence, the reference function consists of objectives and procedures, while the input function contains data and information. In humans, this is conceptually described as a cognitive frame, i.e., 'an explanatory structure that delines entities by describing their relationship to other entities' (see Klein et al., 2007, p. 118). A cognitive frame consists of both goals and cognitive processes for interpreting data. In the context of information analysis, the cognitive frame serves to process the information from the input function and render it available to the output function. Part of a cognitive frame can be, for example, a target value

for controlling the glucose value or a heuristic for estimating the carbohydrates in a meal. In comparison, in the integrated information processing model the term algorithm for machine systems is used. The term 'algorithm' describes a 'predefined set of rules that are followed to solve a task' (see "Algorithm Noun - Definition, Pictures, Pronunciation and Usage Notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.Com", 2024). In this model, this also includes predefined values (e.g., target values from when an AID system stops the insulin supply which represents an internal system logic). Additionally, an algorithm can consist of a complex model (e.g., in the case of language models) or simpler procedures (e.g., a series of logical rules for evaluating a situation).

In the first study presented in this dissertation, the similarity between algorithm and cognitive frame was described as goal congruence, i.e., whether a digital contact tracing was perceived as sharing the same goals and using it was beneficial for one's own goals. Comparable concepts exist in the literature on trust in AI, e.g., as goal alignment by Chiou and Lee (2023) or when looking at an AI system's purpose (Benda et al., 2021). In the context of the IIP model presented in this dissertation, I have decided to use the term *reference consonance* instead. The term *reference* reflects the complexity of an algorithm or cognitive frame better than the term goal, as a cognitive frame can include methods in addition to goals (see Klein et al., 2007). The conceptualization as *consonance* instead of congruence is intended to avoid the impression that the goals and processes of humans and machines must be identical. For example, target values can be represented differently in humans and machines (e.g., a fixed glucose value in mg/dl or a target range in mmol/l) or different processes can be used for the same goal (e.g., a rule-based method and a deep neural network). Based on this information, it could be concluded that there is a lower congruence, although this information processing quality would be fulfilled. The conceptualization as consonance makes it clear that even different descriptions of a reference do not necessarily stand in the way of integrated information processing. However, the different representations of the reference function in machine and human may represent an obstacle to integration if, for example, the human cannot interpret whether the algorithm and cognitive frame are in consonance or not.

Therefore, reference consonance is defined as follows: 'Reference conso-

nance in integrated information processing exists when the objectives and processes used to achieve those objectives by one entity do not contradict the objectives and processes of another entity'. Accordingly, perceived reference consonance describes human 'perception of the extent to which a machine's algorithm does not contradict their cognitive frame'. On the other hand, machines may calculate 'an estimation of the extent to which a human's cognitive frame is free of contradictions to the machine's algorithm'.

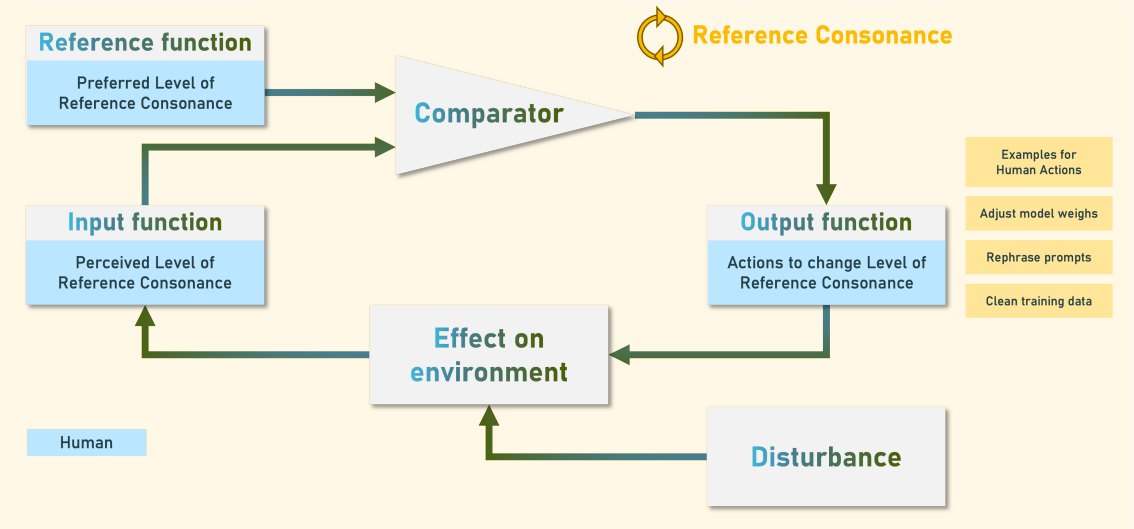


Figure 8.3: Nested loop of human action regulation to reach preferred level of reference consonance

This definition implies that humans and machines can each recognize and respond to a level of reference consonance that diverges from their preferred level. Fig. 8.3 shows the nested loop of the reference consonance, in which a comparison is made with the preferred level of the reference consonance based on knowledge of the other entity's cognitive framework or algorithm. In the event of deviations, a reaction can be generated to correct the reference consonance level. For example, the reference consonance could be lower than the human's preferred level because the system uses information that violates ethical considerations, e.g., medical treatment based on insurance status. To correct this, the human could adapt the machine's algorithm. Table 8.2 shows examples of different situations with a non-preferred level of reference consonance.

For the human-centered development of AI systems (and HCXAI in particular), the

Table 8.2: Examples of low levels of reference consonance with examples from the context of AID systems (referring to study 2)

Perceiving Entity	Target to enhance Reference consonance	Control Action (Examples)	Example
Human	Human Cognitive Frame	Reflect / adopt cognitive frame	Human modifies calculation for insulin demand
	Machine Algorithm	Adapt algorithm (re-training, weight adaption)	Human changes insulin ratio for insulin units / gr carbohydrates in AID calculation of insulin demand
Machine	Human Cognitive Frame	Automatic Feedback, Denial messages	AID system prevents users from dosing too high levels of insulin and presents reasons (why this conflicts with the machine algorithm)
	Machine Algorithm	Adoption of algorithm based on human reactions	After multiple corrections of human user, AID system recalculates insulin ratio used to correct high levels of glucose

ability for humans to determine and correct reference consonance is central.

8.3.4 Output Diagnosticity

The third information processing quality concerns the output of information processing, whereby this initially represents a decision (decision selection) and possibly an action (action implementation). Since the focus of the model is on diagnostic tasks, the implementation of the action (e.g., injecting insulin or pumping insulin into the body) is not described in detail. As depicted at the beginning of this dissertation, making a (diagnostic) decision is based on the formulation and testing of hypotheses. The result of the output function thus represents a tested hypothesis that is considered to be the most accurate. Hypotheses can be, for example, statements about the state of health (e.g., which illness a person has), but also which actions best support existing goals (e.g., which dosage of a medication is the right one). What a hypothesis refers to is therefore closely related to the associated information processing task. In the integrated information processing model presented here, it is assumed that the human output function contains a finite number of hypotheses and that a successfully performed output function is an assessment of these hypotheses (in the sense of a decision in favor of the most appropriate hypothesis and as a basis for possible actions). With the definition of a hypothesis as the result of information processing, human action regulation in the sense of information processing in the frame of the present model is complete. On the other hand, there is the machine output, which is described as inference in the model of integrated information processing.

The inference is created by applying the algorithm (from the reference function) to the content of the input function in the comparator (see Carver and Scheier, 2000). The inference can then be represented visually or verbally, for example, or (in systems with a high level of automation) can also be used directly for the automatic implementation of an action.

The connection between machine inference and human hypotheses lies in the concept of diagnosticity already discussed above. The quality of the integration of human and machine information processing depends on the extent to which the machine-generated inference is diagnostic for the human hypotheses and the extent to which the human creation and evaluation of hypotheses contributes to the machine inference. In other words, this means that there is high output diagnosticity if the relative probabilities of the human hypotheses are changed by the machine-generated inference. For example, the information that the system predicts a drop in blood glucose levels may be sufficiently diagnostic for a person to decide not to inject insulin. The output diagnosticity also depends on the communication of the information, i.e., the extent to which the person can interpret the inference in connection with the hypothesis. That is, varying levels of aggregation and modalities may be represented by different (visual) approaches to communicate inference. For example, the inference of the system (falling glucose level) can be communicated by a downward pointing arrow, a verbal statement or a predicted value.

At the same time, the successful output of integrated information processing also depends on the hypotheses put forward by humans. If the correct hypothesis is not among the hypotheses put forward by the human for testing, the diagnostic accuracy of machine inference may be lower than when evaluating correct hypotheses. For example, when making a diagnosis, the system could provide information on whether a person has an unusually high fever. However, if the person (incorrectly) only considers hypotheses in which fever has no significance, the inference has no diagnostic effect. This process can also be reversed if a machine system only allows a (too small) set of results and human communication cannot be related to it. Support in finding alternative hypotheses (also referred to as option awareness, Pfaff et al., 2013) can improve the diagnosticity of information in the context of interaction with another entity.

In general, output diagnosticity is defined as follows: 'Output diagnosticity in integrated information processing exists when the inference communicated by an entity has sufficient information value to change the relative probabilities of the possible outcomes of the information processing task'. Accordingly, perceived diagnosticity describes 'the extent to which information presented by the machine changes the relative probabilities of existing hypotheses'. In contrast, estimated diagnosticity describes 'the calculated extent to which information that can be communicated to the human can change the relative probabilities of hypotheses currently evaluated by the human'.

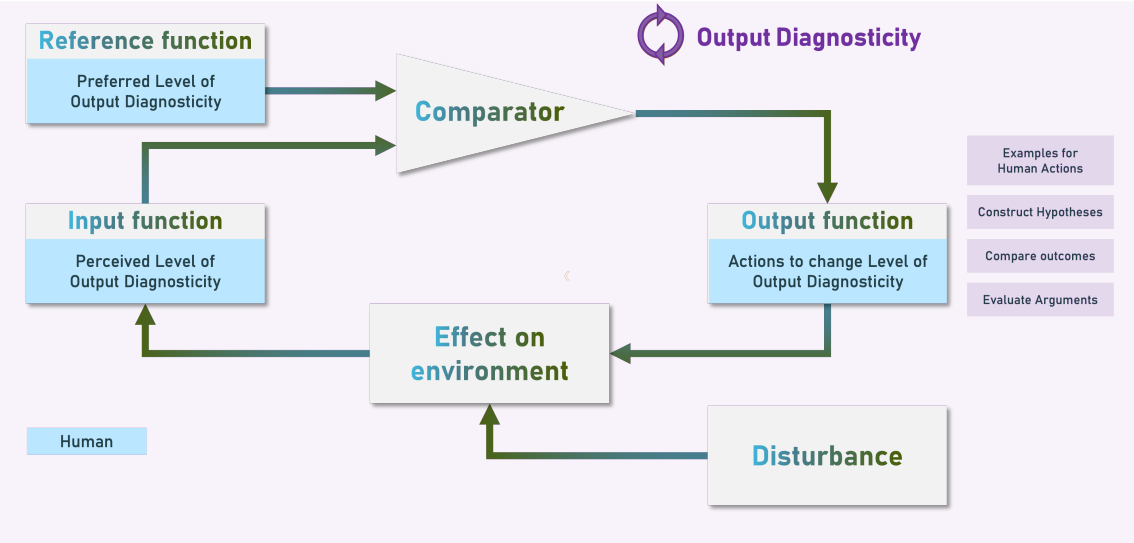


Figure 8.4: Nested loop of human action regulation to reach preferred level of output diagnosticity

This definition emphasizes the need for machine systems to know (or at least be able to estimate) the hypotheses investigated by humans in order to effectively integrate information processing in the output function. Fig. 8.4 shows the nested loop of output diagnosticity, in which the preferred distribution of probabilities is compared to the one achieved within the integrated information processing. In the event of deviations, further hypotheses or communication of inference may be generated. In order to do so, machines require human-awareness (Kambhampati, 2020), which is explained below and discussed in the context of empirical studies. At the same time, however, the given definition of output diagnosticity also points to the need for humans to generate and evaluate hypotheses. Uncalibrated reliance on

Table 8.3: Examples of low levels of output diagnosticity with examples from the context of AID systems (referring to study 2)

Perceiving Entity	Target to improve output diagnosticity	Control Action (Examples)	Example
Human	Human Hypotheses	Generate new hypotheses, adopt hypotheses so inference is applicable	Human considers not only insulin injection but also food intake as possible outcome for glucose development
	Machine Inference	Search machine inference, modify inference, request additional inference details	Human requests machine prediction on different amounts of insulin injection, enables predicted glucose trajectory as visual stimulus
Machine	Human Hypotheses	Suggestion of additional hypotheses (i.e., increasing option awareness)	AID systems suggests adoption of meal time as possible outcome of diabetes therapy review instead of adopting insulin injections
	Machine Inference	Requesting human hypotheses, adoption of inference and inference communication	AID system requests which period of time is concerned in hypothesis and renders glucose level prediction accordingly

AI systems (see Lee and See, 2004), where no engagement with the information to be analyzed has taken place, could be a negative effect of people not forming their own hypotheses (and not using their own information for evaluation). It also follows from the definition that the human impact in the output function can be improved if they are able to form effective hypotheses for the specific task. For example, in the medical process, experts are better at generating and testing hypotheses (P. E. Johnson et al., 1981), while novices in particular can benefit from AI systems as it exceeds their competences (Shen et al., 2019) - further research into how people form and test hypotheses in integrated information processing is needed. Accordingly, how the design of an AI system affects the development and testing of hypotheses (depending on expertise) is a central component of effective systems. Table 8.3 shows examples of different situations with a non-preferred level of output diagnosticity.

8.3.5 Adequacy, Consonance & Diagnosticity as determinants of automation-related UX

A central assumption in the model of integrated information processing is that the automation-related UX depends on how close to the preferred levels the respective information processing quality is (or is perceived to be). Although the model does not aim to describe processes related to the development of user experience, e.g., the development of perceived trustworthiness (see Schlicker et al., 2022), assumptions

can be made about the effects of system design. For example, a language model could be too complex to recognize reference consonances, resulting in a lower perceived trustworthiness. In the case of an AID system, for example, the possibility of correcting an automatically determined glucose value could be missing. This reduces the perceived input adequacy and thus the perceived usefulness of the system. The positive effects attributed to explanations could also be based on the relationships described in the model: for example, it has been shown that giving reasons for the recommendation of a model can have a positive effect on the acceptance of the recommendation (Shin, 2021). This could be due to the fact that this type of explanation can influence all three information processing qualities compared to a baseline. In the study by Cramer et al. (2008), the explanation reveals, for instance, what information a system has about a recommended artwork (input adequacy), which user profile of interests is used for evaluation (reference consonance), and which aspects of an artwork led to the recommendation (output diagnosticity). In the fourth study of this dissertation, we tried to improve users' ability to recognize the effects of interdependence on integrated information processing and could demonstrate positive effects on diagnostic quality.

At the same time, however, there are factors that are part of the human task, but not directly of the model: for example, there may be a high time pressure, whereby the perceived usefulness of a system depends more on the aggregation of information (see feedback given from participants in a study from Evans et al., 2022) than on the output diagnosticity. Similarly, the risk associated with a missing diagnosis plays an important role in vulnerability and thus the development of trust in the human entity (see Hoesterey and Onnasch, 2023; Jacovi et al., 2021). While these meta-factors (as they affect information processing but are not part of it) are not represented in the model of integrated information processing, the effect is described indirectly: via the preferred levels of information processing qualities. For instance, when determining the appropriate portion size for a meal for a person with diabetes who is not at risk of hypoglycemia, a lower input adequacy may be deemed sufficient than when there is a risk of hypoglycemia. In this case, for example, an outdated glucose value would not result in any action by the person to correct it. However, in the case of an urgent decision regarding the dosage of glucose to be taken in case of hypoglycemia, the simple recommendation of a system may be judged as sufficiently

diagnostic. Conversely, in calm situations, for example, the prediction of glucose progression may be preferred by users.

The development of the SIPA scale represents a central contribution of the present study in the area of the survey of automation-related UX. To what extent can the SIPA scale be used to analyze the qualities of IIP described in the conceptual model? Firstly, it can be stated that the SIPA scale does indeed contain three different facets based on the empirical evidence to date: transparency, understandability and predictability (see Schrills, Zoubir, et al., 2021). Analogous to the concept of situation awareness, there is also a hierarchy between the three levels of situation awareness, which means that perceived transparency is higher than perceived predictability (see e.g., Schrills and Franke, 2023). In general, it can be concluded that the use of the SIPA scale for the assessment of automation-related UX fulfills the requirements set out at the beginning of this dissertation: it is suitable as an instrument for assessing the psychological effects of explanations (on several facets) without investigating the direct experience of the explanations, while it ties in theoretically with existing concepts.

In my view, the three qualities of IIP defined in the previous section are also linked to SIPA as a measure of automation-related UX. However, in my opinion, the three levels of SIPA cannot be directly transferred to the conceptual model of IIP. This means that input adequacy cannot be equated with transparency. This is due to the fact that input adequacy is, for example, a consequence of transparency, but also includes, for example, the ability to correct data or for the AI system to request human information. Consequently, the ability to correct data should not further improve transparency but allow for improvement of input adequacy. Similarly, reference consonance cannot be equated with understanding and predictability cannot be equated with diagnosticity. At the same time, the results of Study 2, for example, show that changes that should have a conceptual effect on input adequacy are also reflected accordingly in the area of perceived transparency. Further empirical studies should investigate how the qualities of IIP are related to SIPA.

8.3.6 XAI for Perceived Input Adequacy

As described above, the effects of XAI can be structured on the basis of IIP qualities. The following sections present examples of explanatory approaches and their impact from the perspective of the IIP model on HAI.

The information disclosure presented in the second study of this dissertation is an example of an XAI approach in the area of input adequacy. Here, the information that the AID system currently possesses was presented (e.g., the current glucose value, the insulin currently acting in the body). There was no possibility of correcting this (and it makes no sense, as these were vignettes and the human participants were not able to acquire information independent from the experiment). As another example D. Wang et al. (2019) conducted a study with an intelligent diagnostic system. Based on a theory-driven development of medical decision support systems, they showed a list of health parameters and the most current value of each parameter to the participants. That is, the system communicates which features are actually observed and currently used for analysis. On top of that, the ability to access this information is also reflected by multiple items in the transparency check questionnaire by Schelenz et al. (2024).

Another example is the design of conversational agents and approaches to make them transparent for users. In the work on explainable conversational user interfaces by Schrills, Schmid, et al. (2021), the presentation of recognized information, e.g., on location, was presented as an example. The aim of this approach was to improve the SIPA by displaying data relevant for information processing by the system. However, this may require additional modalities or communication channels (as can be seen in the example of the mentioned study through visual processing, i.e., recognized intents were demonstrated on a visual screen and not communicated via voice). Study 4 also indicates that instruction may support users in keeping high levels of input adequacy by pointing to information interdependence and the need to correct machine data.

However, XAI approaches usually aim to enable users' understanding which is more than demonstrating unprocessed data of the system. That is, transparency is not given by revealing massive amounts of data to a user but by allowing the user to effectively evaluate input adequacy through interaction. Therefore, most XAI

approaches do not (solely) fall into the category of input adequacy. Nevertheless, guidelines on human-AI interaction (e.g., Amershi et al., 2019) include requirements for the ability to correct the system. One conclusion of this dissertation is therefore that effective approaches for interactively checking and changing input adequacy must be developed and integrated into systems. As large amounts of data can be part of the input function, interdisciplinary cooperation between data visualization and human factors is particularly necessary here.

All in all, revelation of input data, communication of system states and data processed by the system relate essentially to XAI techniques that support the evaluation of input adequacy. Future research should focus on developing strategies that allow systems to effectively communicate machine data without causing information overload and enable users to correct machine data in order to interactively raise input adequacy.

8.3.7 XAI for Perceived Reference Consonance

XAI approaches involve addressing low levels of perceived reference consonance, which occurs when there is a misalignment between the AI's algorithm and the user's cognitive frame. Various XAI techniques such as SHAP (Shapley Additive Explanations, see H. Chen et al., 2021), Partial Dependence Plots (PDPs, see Greenwell et al., 2018), permutation/shuffling methods, and rule-based systems play significant roles in mitigating low levels of perceived reference consonance (in general, refer to Molnar, 2022).

SHAP values, for example, are based on cooperative game theory and offer a robust method to attribute the output of a model to its input features. SHAP values work by assigning an importance value to each feature of a model based on the average marginal contribution of that feature across all possible combinations of features. SHAP ensures fair distribution by considering all possible subsets of features to determine each feature's contribution to the prediction. SHAP values provide both local interpretability, explaining individual predictions, and global interpretability, offering insights into the overall model behavior. Empirical studies have shown that SHAP can effectively help users understand how individual features contribute to a model's predictions, thus aligning the AI's decision-making process with user

expectations and enhancing reference consonance. For instance, Lundberg et al. (2019) demonstrated how SHAP values could make the contributions of each feature in a complex medical diagnosis model comprehensible to clinicians, thereby improving the perceived reference consonance between the model's predictions and the clinicians' expectations. By using SHAP values, the algorithm of more complex AI systems can be compared with the user's own cognitive frame and the use of the system can be adapted accordingly. However, they are not interactive, i.e., this technique does not allow the goals of the system to be changed.

Partial Dependence Plots are another frequently discussed XAI technique affecting the detection of reference consonance. Partial Dependence Plots visualize the relationship between a selected feature (or pair of features) and the predicted outcome of a machine learning model, while averaging out the effects of all other features. This technique helps in understanding the marginal effect of the chosen feature(s) on the prediction by showing how changes in the feature values influence the predicted outcome. PDPs are particularly useful for interpreting complex models like gradient boosting machines and random forests as humans can focus on the contribution of a specific aspect of the model. Studies have shown that PDPs can clarify the marginal effects of features on predictions, which may assist users in verifying the model's behavior against their domain knowledge. For example, in financial modeling, PDPs have been used to show how variables like credit score and income affect loan approval predictions, aligning the AI's decision-making process with financial analysts' expectations and improving perceived reference consonance (see Szepannek and Lübke, 2023).

Permutation or shuffling methods can be considered as XAI techniques and are particularly valuable for assessing the robustness of a model. These approaches involve systematically altering input features to evaluate their impact on model predictions (see Dwivedi et al., 2023). By demonstrating how changes in input features influence outcomes, permutation methods help users understand the sensitivity of the AI system to different inputs, thereby revealing the effects of the AI's algorithms.

Rule-based systems use predefined logical rules to make decisions, which are inherently transparent and easy to interpret. The presentation of the entire set of rules, which rules are applied to a specific input or the concrete and rule-based processing steps represent a direct disclosure of the algorithm. The clarity and simplicity of rule-based

decisions potentially align closely with users' cognitive processes, thereby allowing for the perception and comparison of cognitive frame and algorithm. For example, a rule-based system used in medical diagnostics that follows established clinical guidelines can help clinicians understand and trust the AI's recommendations, supporting reference consonance. In conclusion, various XAI techniques can effectively address unpreferred levels of reference consonance by revealing AI algorithms to users.

The presentation of reference sources can also be used as an element to promote reference consonance. For example, large language models such as ChatGPT sometimes show which files or websites were used to generate an answer (e.g., OpenAI, 2023). Access to databases and the use of the corresponding data can also be demonstrated in the evaluation of nutrition, for example. At this point, however, there can potentially be a mix-up with the input function - especially if the data is used to interpret a user's input (e.g., on a food product).

Future research in human-AI interaction centered on optimizing reference consonance levels should focus on the relationship between XAI approaches and users' ability to effectively support reference consonance. For example, when users cannot change model weights, are there other interactions that enable them to raise reference consonance or do they need to incorporate the perceived level of reference consonance in subsequent output generation? Also, can a system detect deviations of users' cognitive frame and its algorithm and how would it react?

8.3.8 XAI for Perceived Output Diagnosticity

Counterfactual explanations in XAI provide insights by generating hypothetical scenarios that illustrate how changes in input features can lead to different outcomes. This technique can be assumed to be particularly effective in enhancing the user's understanding of the model's decision process by pinpointing the minimal changes required to alter a prediction. For example, if a loan application is rejected, a counterfactual explanation might show that increasing the applicant's income by a specific amount would result in approval or that being physical active lowers the amount of insulin needed to correct high levels of glucose. This approach helps users understand the model's decision boundaries and addresses the perception of

reference consonance by aligning the model's logic with users' cognitive frameworks (see Warren et al., 2022).

Semantic enrichment, such as providing detailed descriptions of heat maps, enhances the interpretability of visual explanations. By adding semantic layers to heat maps, users can better understand the significance of highlighted regions in relation to model predictions. For instance, in medical imaging, enriched heat maps can indicate which specific areas of an image were most influential in diagnosing a condition, thus making the AI's reasoning more transparent and supporting reference consonance. Studies have shown that this method can significantly improve clinicians' trust and understanding of AI diagnostics (see Gianfagna and Di Cecco, 2021; Tonekaboni et al., 2019).

Confidence estimation or ratings provide users with a measure of the model's certainty regarding its predictions. This information is crucial for users to gauge the reliability of AI outputs, especially in high-stakes environments such as healthcare and finance. By presenting confidence scores, AI systems help users make informed decisions and better assess the risks associated with relying on the model's predictions. This approach not only enhances user trust but also aligns the model's behavior with user expectations, thereby supporting reference consonance (Gianfagna and Di Cecco, 2021; T. Le et al., 2023).

Local feature relevance explanations focus on the importance of specific features in making a particular decision. This method helps users understand how individual input variables influence the model's output for a specific instance, providing a detailed view of the decision-making process. For example, in a fraud detection system, local feature relevance might highlight that unusual transaction amounts and locations were key factors in flagging a transaction as suspicious. This granularity allows users to validate the model's reasoning against their knowledge and expectations, thus strengthening reference consonance (Doshi-Velez and Kim, 2017; Lundberg et al., 2019).

In summary, counterfactual explanations, confidence estimation, and local feature relevance are vital XAI techniques that enhance perceived output diagnosticity by making AI systems more transparent and aligned with user expectations. Fu-

ture research should focus on developing adaptive explanation systems, conducting empirical studies on output diagnosticity, investigating long-term impacts of XAI, integrating interdisciplinary approaches, and implementing user feedback mechanisms to continually align AI inference and its explanation with user hypotheses and raise output diagnosticity.

8.3.9 What is new about AI for engineering psychology?

From the perspective of engineering psychology, the question can rightly be asked: what are the unique qualities of AI systems that we have not yet considered in automation research? As already described in chapter 2, existing models of information processing can be transferred to highly automated human-technology interaction through AI - what is the need for new models or a focus on automation-related UX?

The need to use an AI system instead of a non-intelligent system is due to the (current) lack of a sufficient description of a problem to be able to solve it without the use of AI. When using AID systems (generally comparable to the personalized use of medication), the aim is to automate the control of a complex hormonal system with a high degree of uncertainty. At the same time, the system needs a lot of information from the user - for example, about exercise or diet, as described in Study 2. Study 2 shows that interactions with systems that handle tasks of such complexity carry risks for integrated information processing. These risks are similar to those already described in the field of automation: for example, an illusory high level of confidence (see Chromik et al., 2021) in having understood a system is closely related to the idea of complacency and overconfidence (see Harbarth et al., 2024), as individual errors of judgement lead to risky use of the system. In systems defined as AI systems, I believe that this risk is primarily due to a lack of interaction in the input, reference, and output functions, and that the AI-specific challenges of these three steps are therefore also the 'new' challenges for engineering psychology.

For example, in order for humans to achieve the desired level of input adequacy, it is necessary, among other things, to make the machine's data available to humans in such a way that the data can be checked and modified. This is a challenge for engineering psychology if, for example, the way in which the machine stores data

does not match the way in which humans store information. An example of this can be found in Study 1, where people encode information about contacts using contextual information such as location or reason, whereas the contact tracing system defines (and aggregates) encounters solely by time. Regardless of the reasons for the discrepancy in format (technical, privacy, etc.), it is important to understand from an engineering psychology perspective how this affects integrated information processing.

In the context of research on reference consonance, there is a considerable need for research on the control of intelligent systems for engineering psychology. Dang et al. (2022), for example, analyzed the control of generative models using sliders. It was found that more sliders increased cognitive load, but not overall performance - so appropriate forms of control algorithms (especially for multi-parameter models) are key to enabling the widespread use of AI technologies. In contrast to existing research in engineering psychology, the focus is not only on the resources for exercising control (e.g., attention, memory), but also on the ability to mentally simulate which control measures on a model will achieve the desired effect when the algorithm is applied to the input function (cf. mental models, e.g., Carroll and Olson, 1988). While I have already pointed out the need to study mental models, Study 3, for example, shows the risks of models without sufficient controllability - namely the use of inefficient strategies (such as probability matching) when using AI systems.

In my view, the focus of engineering psychology research on diagnostics is the most important and impactful contribution of engineering psychology in the coming period. While the previously presented aspects are also discussed from a design and computer science perspective, the exploration of diagnostic human-AI interaction is based on both the human ability to test hypotheses and the machine provision of inference - making it an excellent example of engineering psychology work. At the time of writing, there are 50 entries on Google Scholar for the terms 'diagnosticity' and 'human-AI interaction'. In contrast, for example, there are 3530 entries for 'reliability' and 'human-ai interaction' or 4960 for 'accuracy' and 'human-ai interaction'. This dissertation shows that explanations (as in Study 2) can lead to an overestimation of system understanding, which I believe is due to a lack of diagnosticity in the explanations. Together with the results of Study 1, which emphasize the importance

of perceived diagnosticity for intelligent automation, this research represents a central point for the engineering psychology agenda in the field of AI.

8.3.10 The Need of Human-aware-systems for Integrated Information Processing

The previous chapters of this dissertation emphasize the need for integrated information processing for effective human-AI interaction. Inspired by concepts such as group awareness (see Bodemer and Dehler, 2011) and situation awareness, the foundation of successful human-AI interaction lies in mutual awareness of information processing. XAI methods, as previously discussed, can enhance a human's understanding of an AI's information processing. Empirical research suggests that care must be taken to avoid creating an illusion of information processing awareness, where users overestimate their understanding of the AI's information processing (see Chromik et al., 2021).

It has to be kept in mind, however, that awareness of information processing needs to be bidirectional in order to achieve successful integration (see M. Johnson et al., 2012 and Klein et al., 2004). AI systems need to incorporate representations of human information processing to function effectively. Human-aware AI systems, as described by Kambhampati (2020), are designed to consider and adapt to the mental models and situation models of human users. These systems aim to synchronize their processes with the user's cognitive frameworks, potentially improving information processing qualities. The IIP model proposed in this dissertation parallels human mental models of information processing (representing human regulatory control and information processing) with the information processing steps of machine systems, aiming to facilitate mutual awareness.

While research on human-aware AI still is in its beginnings, existing approaches include several strategies: For instance, explicability and explanation frameworks are designed to ensure that AI behavior is not only aligned with human expectations but also provides necessary clarifications when deviations occur (Kambhampati, 2020). This involves the AI agent explaining its actions in a way that reconciles differences between its model and the human's mental model, enhancing trust and collaboration.

From my point of view, however, it is important that human-aware AI does not only integrate its information about the user's mental and situational model to justify an output, but in all steps of IIP.

Further research should focus on how human inputs, cognitive frames, and hypotheses can be represented within AI systems. Studies on human-computer interaction must explore which AI actions best support the acquisition and adoption of human aspects of the information processing qualities. This entails investigating how AI can adaptively respond to human cognitive states and intentions, thus supporting a more fluid and intuitive interaction (Sreedharan, 2023). In conclusion, the integration of human-aware systems into AI models is crucial for enhancing mutual information processing awareness, that is, the basis to effectively optimize information processing qualities. AI systems need to be designed with the capability to understand and adapt to human cognitive framework (as discussed in the IIP model), so future research must continue to bridge the gap between human cognition and AI processes by developing methods for machine systems to represent, acquire and act on data regarding human information processing.

8.4 Contributions to Research on Human-Centered XAI

HCXAI research aims to develop AI systems that are not only technically proficient but also considerate of the human users interacting with them. This research emphasizes understanding "who" the human user is, the context of their interaction, and how AI systems can be designed to align with human cognitive processes, values, and social contexts. By integrating both technological advancements and human factors, HCXAI strives to create AI systems that are transparent, trustworthy, and effective in enhancing user experience and decision-making processes (see Ehsan and Riedl, 2020).

The overall objective of this dissertation was to examine how the information displayed by an AI system in diagnostic tasks affects integrated human-AI information processing. This aligns with the HCXAI framework by focusing on the integration

of human and machine and its impact on automation-related user experience and behavior. The dissertation's implications lie in understanding how AI explanations influence user automation-related experience and overall interaction. Specifically, the present research highlights the need for AI systems to provide explanations that are not only accurate but also depend on human states, which is a detrimental aspect of HCXAI.

In particular, the present work contributes to HCXAI research by addressing the divergence between subjective user experience and user intention in AI-assisted diagnostic tasks (see study 1 and 2). It underscores the importance of understanding how different explanation styles and information processing methods affect user perceptions and actions and raises question on the situational variables that enhance the gap between user experience and behavior in information processing. By developing and applying the SIPA scale, I provided empirical evidence on how AI explanations impact user experience and decision-making, especially in the context of predicting machine calculations (see study 2).

Moreover, the research presented here emphasizes the importance of automation-related user experience, advocating for a holistic approach that considers both technical and human factors in the design of AI systems. The findings suggest that explanations must be tailored to the users' needs and contexts to foster effective interaction and avoid risks like over-reliance and illusions of competence (Sieker et al., 2024).

In greater detail, study 1 demonstrated the relevance of information processing qualities for use intention and qualitatively identified potential improvements for integrated information processing. In particular, the importance of diagnosticity was demonstrated and possible solutions (e.g., by letting users include the mask status in digital contact tracing) were discussed. Understanding why people choose to use AI systems is crucial, as this knowledge can guide the development of more intuitive and effective AI tools. Burton et al. (2020) emphasized that user acceptance and trust are significantly influenced by the alignment of AI's decision-making processes with users' cognitive models and expectations. Thus, future research should focus on identifying and enhancing the factors that drive user intention and acceptance, ensuring that AI systems provide clear, comprehensible explanations of their operations and decisions

(Burton et al., 2020; D. Wang et al., 2019).

In addition, study 2 underscored the risks associated with the illusion of information processing awareness and the necessity to study automation-related user experience and performance as two different constructs. In previous research, explainability-accuracy trade-off has been discussed to demonstrate drawbacks of explainable systems (Crook et al., 2023): higher levels of explainability are possibly, but not necessarily connected to lower levels of accuracy. However, a concept that could be termed the transparency-complacency trade-off needs to be focused even more when discussing human factors and XAI. A trade-off between transparency and complacency may describe situations in which users receive excessive information (potentially causing an information overload, see Sewnath and Crijnen, 2021), leading them to overestimate their understanding and fail to question the accuracy of the provided information. Chromik et al. (2021) identified this risk, demonstrating that users may become complacent due to an illusion of understanding. Research must continue to explore the trade-off between transparency and complacency and identify contexts in which it is most likely to occur, focusing on how different presentation methods and levels of detail impact user vigilance and decision-making accuracy. Interestingly, study 2 also demonstrated that explanations might be utilized differently depending on the task a human user has (users spent more time checking information when they were tasked to make a prediction compared to an observation). That highlights the importance to focus on human tasks and human information processing when designing XAI systems.

Furthermore, study 3 did not confirm the presence of a question-behavior effect in trust studies related to AI, which softens potential methodological concerns but also raises questions about the connection between automation-related UX and (observable) behavior. The lack of predictive power of trust assessment on reliance underscores the importance of understanding human strategies in integrating AI (outputs) into their decision-making processes. Chiou and Lee (2023) discussed that grasping these human information processing strategies is crucial for developing AI systems that effectively support human decision-making. Future research should investigate the dynamic interactions between human strategies and AI assistance, particularly in complex and high-stakes environments, to design systems that provide

diagnostic information (and not only information perceived as diagnostic).

Finally, I demonstrated in study 4 that approaching integrated information processing includes the communication of interdependence in diagnostic tasks. While explanations usually aim to improve mental models of users about how an AI systems works, the observed improvement in diagnostic quality in study 4 demonstrates that improving the mental model of the overall information processing may be more desirable. Furthermore, the design of instruction based on seamful design constitutes one of the first application of seamful design in human-AI interaction and can be seen as approach for further AI-based DSS, in and outside of the medical field. Further research can utilize my approach to analyse informational interdependence and use instructions to let users experience the effects of integrated information processing rather than trying to explain them.

In summary, the research presented in this dissertation introduced theoretical concepts of human-AI interaction, a psychologically oriented conceptual model of integrated information processing, multiple research instruments (notably the SIPA scale), and the use of Kandinsky patterns as a paradigm to study integrated information processing in human-AI interaction. The designed paradigm supports theory-driven research in examining how AI-displayed information in diagnostic tasks affects automation-related user experience and behavior. The findings presented within this dissertation highlight the necessity for further research into the cognitive and psychological aspects of human-AI interaction and the experimental examination of interaction that allows for integrated information processing. That is, research in HCXAI should aim to develop AI systems that do not only demonstrate high levels of accuracy but are designed in the context of human cognitive processes and expectations, thereby enhancing adequacy, supporting consonance and providing diagnosticity to allow users the integration of AI into diagnostic decision-making processes.

8.5 Practical Implications for Diagnostic AI in High Risk Areas

The timing of this dissertation coincides with various international efforts to achieve effective legal regulation of AI technology. In addition to activities in the US (see White House Office of Science and Technology Policy, 2022), UK (see Roberts et al., 2022) and Canada (see Scassa, 2023), the work of the EU High Level Expert Group (see High-Level Expert Group on AI, 2019) and the EU Regulation on Artificial Intelligence (see European Union, 2024) approved of in April 2024 should be mentioned in particular. In practice, the regulations still raise open questions regarding the concrete implementation, testing and safeguarding of the requirements.

The results of this dissertation can, for example, contribute to the implementation of the still abstract requirements of the EU AI Act. A key example is human oversight. Human oversight can be defined on different levels (Sterz et al., 2024) and it is currently unclear which design approaches meet the requirements for human oversight. For example, the technical requirements for human oversight may include control options for human users. However, sufficient training and motivation to oversee an AI system may also constitute human oversight (Sterz et al., 2024), and the results of Study 2 suggest that simply disclosing more information is not sufficient to ensure better oversight. Publications to which I contributed and which followed Study 2 (see Schrills et al., 2023) also support the claim that oversight is more demanding than transparency. The model of integrated information processing presented in this dissertation can describe both technical and human requirements for effective human oversight. For instance, for AID systems as an example of high-risk systems in the sense of the EU AI Act, requirements could be derived with respect to the input, reference and output functions: the correction of input adequacy as a component of effective human oversight requires the possibility of adding and correcting relevant data on exercise or food intake. It should also be possible to edit target values, such as when insulin should be given, to improve reference compliance. Finally, it must be ensured that an action (e.g., eating or stopping exercise) can be derived clearly and with sufficient diagnostic accuracy from the information provided by the system, e.g., in the case of hypoglycemia.

Since the conceptual model of integrated information processing can be used to establish requirements for human-AI interaction in high-risk systems, it can also be used to structure user training and system design. For example, the skills to be learnt by a person using an AID system could be trained based on input, reference and output functions. In terms of input adequacy, for example, people would need to be able to recognize what information they need to give to the system, as the system has no sensors. In the case of AID systems, this could be food intake, physical activity or health status. In terms of reference congruence, people should be trained to communicate their goals to the system: for example, to aim for a higher value when driving in order to avoid unexpected hypoglycemia. In some systems, this may allow the use of a temporary sport or ZEN mode (e.g., "Diabeloop DBLG1 System Overview", 2024), while in others the glucose target may need to be changed. The interpretation of system cues also needs to be trained. For example, people should be able to recognize invalid glucose patterns and, for example, test the hypothesis of a sensor defect.

Finally, research on SIPA and the illusion of explanatory depth also shows that requirements for high-risk systems cannot be defined as process standards alone. This means that the defined requirements must include empirical analyses of risk measures (e.g., XAI). A mere description of which XAI methods have been selected and applied does not do justice to integrated information processing.

Overall, this work provides a basis for the operationalization of ethical and legal requirements and at the same time underlines the need for empirical testing of high-risk applications of AI systems.

8.6 Limitations

The studies in this dissertation are conceptually diverse: a survey relying on qualitative and descriptive data focusing on real-world experience with automated technology, an experimental study with a specific target group and system that emulates a specific application, and a study with an abstract, foundational task that addresses general cognitive mechanisms and methodological questions of HAI research. While such

broad and diverse approaches allow to deduct general gaps in the research of HAI, compared to works incorporating only studies focusing on a single thematic area, the empirical results in this work need replication to prove their stability. In addition, the IIP model as conceptual model needs to be based on further experimental research that allows for causal deductions. For instance, there was no experimental investigation into the effects of specific properties of digital contact tracing, but only the results of a descriptive study within a pandemic situation.

The research presented employs both self-developed and established scales, as well as newly developed items to capture facets of user experience in Human-AI interaction. That is, a key assumption of this work is that traditional variables of UX (e.g., acceptance, usefulness, hedonic or pragmatic quality) may be not sufficient to describe how users experience integrated information processing. The development, validation and usefulness of variables able to capture UX in automated diagnosis does not, however, draw on the same mountain of literature - and the results drawn from it must therefore also be considered in the light of these limitations of the measurement tools used. Although the SIPA scale is a central contribution across all studies (and has been validated, see Schrills et al., 2024), the results of this dissertation do not cover affective or emotional components of user experience, e.g., as dimensions of trust, as these were not included, e.g., in study 1. Instead, the dissertation primarily focuses on cognitive elements of Human-AI interaction and does not address topics such as anthropomorphic representations, user needs, or individual user characteristics (i.e., user diversity variables) beyond the Affinity for Technology Interaction. After completing and evaluating the studies in this dissertation, it is clear that the effects of the integration of machine and human information processing on humans should be investigated more thoroughly: in particular, the experience of autonomy and one's own competence as central elements in self-determination theory play an important role here (cf. Moradbakhti et al., 2024). The own satisfaction with the use of AI systems was less considered in this dissertation. The use of suitable scales (as the Psychological Needs Scale for Technology Use from Moradbakhti et al., 2024) should be investigated more intensively alongside the cognitive questions in HAI. While these were not the primary goals of this research, emotional responses also play a critical role in everyday diagnostic contexts and warrant further exploration.

The conceptual model of integrated information processing, detailed in this dissertation, integrates existing theoretical models from the broader domains of automation and general human technology interaction and action regulation with technical XAI approaches, responding to the findings of the presented studies. However, in the present stage of development the IIP model cannot be considered a theory in its own right, as the relationships identified do not yield empirically falsifiable hypotheses. However, considering how Poulton (1966) discusses engineering psychology, the proposed model allows to derive hypotheses about human-centered design and the effects on human users, therefore, constituting a significant contribution of engineering psychology to research on HAI. Yet, future research needs to empirically test the relationships among the presented concepts, considering context factors such as time pressure and cognitive load, which are central to action regulation but only indirectly addressed in the conceptual model.

This dissertation explicitly addresses diagnostic tasks, deliberately applying a broad definition of diagnosis to include decision-making processes beyond medical inquiries, technical trouble-shooting or even legal classifications. The most recent discussions about generative AI (see for example Zamfirescu-Pereira et al., 2023), especially since the release of large language models by OpenAI (OpenAI, 2023), are less addressed within this dissertation although they may constitute for an important part of information processing in human-AI teams. While the conceptual model proposed in this dissertation can be applied to large language models involved in integrated information processing for diagnostic tasks, the studies and model development did not consider generative tasks, where AI (and users) create tangible content. In my view, it would be useful to test the applicability of the model to generative systems on the basis of case studies and, if necessary, derive design implications.

Overall, this work identified empirical gaps in existing research and conceptualizations of Human-AI interaction and offers theoretical considerations for bridging these gaps. Here, the focus of this work lied on the conceptual integration of machine and human information processing. However, further empirical research is necessary to concretize the concepts underlying the integrated information processing model in applied contexts and to derive actionable recommendations for high-risk systems, such as those in medical domains.

8.7 Future Research

This dissertation can lay the groundwork for extensive empirical research into the integration of humans and machines in information processing. Future research should specifically address the following areas:

Construction of Scales for Perceived Information Processing Qualities: Developing reliable and valid scales to assess how humans perceive information processing qualities is essential. This can build on existing studies, such as the scales in study 1, the transparency check scales discussed from Schelenz et al. (2024), and measures of group awareness and human-AI teaming (Attig et al., 2024).

Development of Methods for Automatic Calculation of Information Processing Qualities: Methods for automatically calculating information processing qualities by systems should be developed. These can draw from procedures in human-aware AI or cognitive tracing methods, such as those used in interactive teaching and learning systems. That is, computational models need to be able to represent the mental state of a user in terms of input, reference and output function and adopt their information processing accordingly. Accordingly, cognitive modeling of users (see Aleven, 2010), where knowledge and strategies are described, may need to be involved in the design of XAI approaches. All in all, further development of student and user models is necessary to calculate information processing qualities based on user status and develop interaction schemes to address user states.

Standardized Assessment Procedures: Procedures for the standardized assessment of information processing qualities (e.g., to assess compliance with regulation) are crucial. This includes the assessment of human supervision, as discussed by Sterz et al. (2024). For instance, empirical assessment techniques based on situation awareness (e.g., freeze probes, see Salmon et al., 2009) may be employed to analyze the state of information for both entities at specific times, and to evaluate how well humans recognize system goals that differ from their own. Emphasis should be placed on developing explicitly behavior-based measures rather than questionnaires predicting behavior.

Design of Interactions to Improve Information Processing Qualities: Inter-

actions between humans and AI systems should be designed to enhance information processing qualities. This should be informed by existing automation research, focusing on human action regulation and interventions to modify human information processing. Such interactions must promote desirable system characteristics, such as human autonomy and the experience of control. Interdisciplinary research into the visualization of machine status (including data, algorithms, and inference processes) is particularly critical.

Experimental Investigation of Automation-Related User Experience (UX): The impact of interactions on automation-related user experience should be explored through carefully designed experiments. These experiments should contribute to theoretical models while providing practical recommendations. It is essential to clearly define and theoretically assign the developed interventions prior to empirical investigation. The expected effects must be described at both the behavioral level and the level of automation-related UX.

In summary, advancing the field of diagnostic Human-AI interaction requires a multifaceted approach encompassing scale construction, method development, standardized assessment, interaction design, and experimental investigation. Each of these areas is integral to enhancing our understanding and optimization of human-machine information processing integration.

In summary, this dissertation underscores the profound need for a deeper integration of interdisciplinary and transdisciplinary research strategies. The integrated nature of human-AI interaction and the quest for human-centered, explainable AI require insights from diverse fields, including psychology, computer science, ethics, and sociology. This call for 'integrated research for integrated information processing' is not just a strategic consideration, but a necessity to ensure the development of AI systems that are both effective and ensure meaningful human contribution. Future research should prioritize the creation of frameworks that facilitate such collaborations, drawing on the strengths of different disciplines to foster innovation and societal benefit. In doing so, we can pave the way for AI technologies that are not only technically robust, but also aligned with human values and needs (Calero Valdez et al., 2024).

9 Conclusion

The research presented in this dissertation significantly contributes to advancing the understanding of human-centered explainable AI (XAI) systems, particularly within diagnostic tasks from an engineering psychology perspective. Herein, this work addresses to contribute filling empirical and theoretical gaps concerning end-users and their interactions with AI, as opposed to developers or professionals, thereby providing a more comprehensive view of how XAI can be optimized for diverse user groups.

Two contributions are, from my point of view, particularly important: first, the refinement and application of the SIPA scale. The SIPA scale allows for differentiated and detailed assessments of the effects of AI explanations on user experience in automated systems and highlights the need for theory-driven concepts and operationalization in automation-related UX research.

The second key element developed through the present research is the model of integrated human-AI information processing. This model emphasizes the necessity of thinking about human and machine information processing at the same time when discussing diagnostic process, rather than position AI as merely a tool or humans as supervisors without their own information processing. It highlights how AI systems can enhance human decision-making by providing explanations that align with human cognitive processes. This model is crucial for developing AI systems that are both effective and human-centered, as it ensures that AI support is seamlessly integrated into human cognition.

Looking somewhat more into the future, my findings underscore the necessity of human-aware systems that integrate automated information processing in a manner that supports user autonomy and control. Future research should continue to explore

the dynamics of human-AI interaction, particularly how users develop and test hypotheses within AI-supported environments. This involves investigating how AI explanations can be dynamically adapted based on contextual cues and user feedback to enhance interpretability and trustworthiness.

In conclusion, this dissertation contributes to the foundational understanding of human-centered XAI by providing empirical evidence and theoretical insights that guide the design of more effective and user-friendly XAI systems. The refined SIPA scale, experimental findings, and focus on user experience and performance offer a comprehensive framework for future research and development in the field of explainable AI. The ultimate goal remains to empower users through transparent, reliable, and safe AI interactions, fostering a collaborative and informed integration between humans and intelligent systems (Shneiderman, 2020a).

On a final note: from my point of view, the way how humans process information is an essential part of being human, especially in its diversity. We differ in what information we bring into diagnostic processes, we differ in our goals and in the way we build hypotheses. We differ in the diagnoses we pursue in our everyday lives, in the resources we bring to reach them. This diversity is the basis of our human progress and individual development. As in the quote mentioned at the beginning, the development of AI systems is a technology that interferes significantly with this essential way of being human. When Anders (1980) says that we need to make the "consequences of not working bearable" (p.80), it is not just about how we organize our free time. Rather, it is about how our needs to contribute as human beings can be met in a digitalized information-processing world and society - fueled by AI. To ensure that AI technology is not just a catalyst for efficiency and automation, but a milestone of human development, AI needs to be integrated into human information processing and not the other way around.

10 Bibliography

Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., & Lapuschkin, S. (2023). From attribution maps to human-understandable explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, *5*(9), 1006–1019. https://doi.org/10.1038/s42256-023-00711-8 (cit. on p. 26).

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052 (cit. on pp. 20, 38).

Adams, D. A., Nelson, R. R., & Todd, P. A. (1992). Perceived Usefulness, Ease of Use, and Usage of Information Technology: A Replication. *MIS Quarterly*, *16*(2), 227. https://doi.org/10.2307/249577 (cit. on p. 29).

Ahmad, S. F., Han, H., Alam, M. M., Rehmat, M. K., Irshad, M., Arraño-Muñoz, M., & Ariza-Montes, A. (2023). Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications*, *10*(1), 311. https://doi.org/10.1057/s41599-023-01787-8 (cit. on p. 2).

Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making*, *21*(1), 178. https://doi.org/10.1186/s12911-021-01542-6 (cit. on p. 34).

Aleven, V. (2010). Rule-Based Cognitive Modeling for Intelligent Tutoring Systems. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.). J. Kacprzyk (Ed.), *Advances in Intelligent Tutoring Systems* (pp. 33–62, Vol. 308). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14363-2_3 (cit. on p. 196).

Algorithm noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com. (2024).

Retrieved August 16, 2024, from https://www.oxfordlearnersdictionaries.com/definition/english/algorithm# (cit. on p. 172).

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does Explainable Artificial Intelligence Improve Human Decision-Making? *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(8), 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819 (cit. on p. 35).

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233 (cit. on pp. 23, 181).

Amparore, E., Perotti, A., & Bajardi, P. (2021). To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods. *PeerJ Computer Science*, *7*, e479. https://doi.org/10.7717/peerj-cs.479 (cit. on p. 26).

Anders, G. (1980). *Die Antiquiertheit des Menschen Bd. II: Über die zerstörung des lebens im zeitalter der dritten industriellen revolution* (Vol. 2). Verlag C.H.BECK oHG. (Cit. on pp. 1, 199).

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2019, September 14). *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*. arXiv: 1909.03012 [cs, stat]. https://doi.org/10.48550/arXiv.1909.03012 (cit. on p. 25).

Attig, C., Wollstadt, P., Schrills, T., Franke, T., & Wiebel-Herboth, C. B. (2024). More than Task Performance: Developing New Criteria for Successful Human-AI Teaming Using the Cooperative Card Game Hanabi. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3613905.3650853 (cit. on pp. 3, 196).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140 (cit. on p. 27).

Bae, S., Lee, Y. K., & Hahn, S. (2023). Friendly-Bot: The Impact of Chatbot Appearance and Relationship Style on User Trust. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). Retrieved August 16, 2024, from https://escholarship.org/uc/item/0gr051sj (cit. on p. 40).

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, *11* (cit. on p. 164).

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001 (cit. on p. 15).

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8 (cit. on p. 21).

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445717 (cit. on pp. 25, 32, 35, 41).

Barney, J. B., & Hansen, M. H. (1994). Trustworthiness as a Source of Competitive Advantage. *Strategic Management Journal*, *15*(S1), 175–190. https://doi.org/10.1002/smj.4250150912 (cit. on p. 29).

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning. *Organizational Behavior and Human Decision Processes*, *42*(1), 88–110. https://doi.org/10.1016/0749-5978(88)90021-0 (cit. on p. 1).

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 (cit. on p. 20).

Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking Aided Decision Making in a Signal Detection Task. *Human Factors*, *59*(6), 881–900. https://doi.org/10.1177/0018720817700258 (cit. on pp. 21, 163).

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-

off in Machine Learning for Public Policy. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 248–266. https://doi.org/10.1145/3531146.3533090 (cit. on p. 164).

Benda, N. C., Reale, C., Ancker, J. S., Ribeiro, J., Walsh, C. G., & Novak, L. L. (2021). Purpose, process, performance: Designing for appropriate trust of AI in healthcare. *Proceedings of the CHI Conference on Human Factors in Computing Systems, Yokohama*, 1–5 (cit. on p. 172).

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 78–91. https://doi.org/10.1145/3514094.3534164 (cit. on pp. 20, 34).

Bertrand, A., Viard, T., Belloum, R., Eagan, J. R., & Maxwell, W. (2023). On Selective, Mutable and Dialogic XAI: A Review of What Users Say about Different Types of Interactive Explanations. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3544548.3581314 (cit. on pp. 21, 165).

Birmingham, H., & Taylor, F. (1954). A Design Philosophy for Man-Machine Control Systems. *Proceedings of the IRE*, *42*(12), 1748–1758. https://doi.org/10.1109/JRPROC.1954.274775 (cit. on p. 17).

Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management*, *13*(3), 271–286. https://doi.org/10.1016/S0956-5221(97)84644-1 (cit. on p. 29).

Bodemer, D., & Dehler, J. (2011). Group awareness in CSCL environments. *Computers in Human Behavior*, *27*(3), 1043–1045. https://doi.org/10.1016/j.chb.2010.07.014 (cit. on p. 187).

Bolton, M. L. (2022). Trust is Not a Virtue: Why We Should Not Trust Trust. *Ergonomics in Design*, 10648046221130171. https://doi.org/10.1177/10648046221130171 (cit. on p. 31).

Boreak, N. (2020). Effectiveness of Artificial Intelligence Applications Designed for Endodontic Diagnosis, Decision-making, and Prediction of Prognosis: A Systematic Review. *The Journal of Contemporary Dental Practice*, *21*(8), 926–934. https://doi.org/10.5005/jp-journals-10024-2894 (cit. on p. 8).

Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., & Nensa, F. (2023). Explainable AI in medical imaging: An overview for

clinical practitioners – Saliency-based XAI approaches. *European Journal of Radiology*, *162*, 110787. https://doi.org/10.1016/j.ejrad.2023.110787 (cit. on p. 2).

Bosch, J., Olsson, H. H., & Crnkovic, I. (2021). Engineering AI Systems: A Research Agenda. In A. K. Luhach & A. Elçi (Eds.), *Advances in Systems Analysis, Software Engineering, and High Performance Computing* (pp. 1–19). IGI Global. https://doi.org/10.4018/978-1-7998-5101-1.ch001 (cit. on p. 5).

Bowen, J. L. (2006). Educational Strategies to Promote Clinical Diagnostic Reasoning (M. Cox & D. M. Irby, Eds.). *New England Journal of Medicine*, *355*(21), 2217–2225. https://doi.org/10.1056/NEJMra054782 (cit. on p. 1).

Broadbent, D. E. (1965). Information Processing in the Nervous System: The nervous system, limited in its ability to process sensory data, must operate selectively and economically. *Science*, *150*(3695), 457–462. https://doi.org/10.1126/science.150.3695.457 (cit. on p. 8).

Broadbent, D. E. (1958). *Perception and communication*. Pergamon Press. (Cit. on p. 8).

Broadbent, D. E. (1982). Task combination and selective intake of information. *Acta Psychologica*, *50*(3), 253–290. https://doi.org/10.1016/0001-6918(82)90043-9 (cit. on p. 8).

BS ISO/IEC 42001:2023: Information technology — artificial intelligence — management system. (2023). (Cit. on p. 13).

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1–21. https://doi.org/10.1145/3449287 (cit. on pp. 20, 21).

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155 (cit. on pp. 189, 190).

Cabral, S., Restrepo, D., Kanjee, Z., Wilson, P., Crowe, B., Abdulnour, R.-E., & Rodman, A. (2024). Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Internal Medicine*, *184*(5), 581. https://doi.org/10.1001/jamainternmed.2024.0295 (cit. on p. 20).

Calero Valdez, A., Heine, M., Franke, T., Jochems, N., Jetter, H.-C., & Schrills, T. (2024). The European commitment to human-centered technology: The integral role of HCI in the EU AI Act's success. *i-com*, *0*(0). https://doi.org/10.1515/icom-2024-0014 (cit. on pp. 3, 197).

Carroll, J. M. (1997). Human–computer interaction: Psychology as a science of design. *International Journal of Human-Computer Studies*, *46*(4), 501–522. https://doi.org/10.1006/ijhc.1996.0101 (cit. on p. 5).

Carroll, J. M., & Olson, J. R. (1988). Mental Models in Human-Computer Interaction1. In M. A. R. T. I. N. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 45–65). North-Holland. https://doi.org/10.1016/B978-0-444-70536-5.50007-5 (cit. on p. 186).

Carver, C. S., & Scheier, M. F. (2000). On the Structure of Behavioral Self-Regulation. In *Handbook of Self-Regulation* (pp. 41–84). Elsevier. https://doi.org/10.1016/B978-012109890-2/50032-9 (cit. on pp. 3, 41, 166, 167, 171, 175).

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics*, *3*(1), 241–255. https://doi.org/10.1007/s43681-022-00167-3 (cit. on p. 41).

Chanda, T., Hauser, K., Hobelsberger, S., Bucher, T.-C., Garcia, C. N., Wies, C., Kittler, H., Tschandl, P., Navarrete-Dechent, C., Podlipnik, S., Chousakos, E., Crnaric, I., Majstorovic, J., Alhajwan, L., Foreman, T., Peternel, S., Sarap, S., Özdemir, İ., Barnhill, R. L., ... Brinker, T. J. (2024). Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications*, *15*(1), 524. https://doi.org/10.1038/s41467-023-43095-4 (cit. on p. 34).

Chapanis, A. (1965). On the allocation of functions between men and machines. *Occupational Psychology*, *39*(1), 1–11 (cit. on p. 17).

Chen, H., Lundberg, S., & Lee, S.-I. (2021). Explaining Models by Propagating Shapley Values of Local Components. In A. Shaban-Nejad, M. Michalowski, & D. L. Buckeridge (Eds.), *Explainable AI in Healthcare and Medicine* (pp. 261–270, Vol. 914). Springer International Publishing. https://doi.org/10.1007/978-3-030-53352-6_24 (cit. on pp. 2, 181).

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282. https://doi.org/10.1080/1463922X.2017.1315750 (cit. on p. 18).

Chignell, M., Wang, L., Zare, A., & Li, J. (2023). The Evolution of HCI and Human Factors: Integrating Human and Artificial Intelligence. *ACM Transactions on Computer-Human Interaction*, *30*(2), 1–30. https://doi.org/10.1145/3557891 (cit. on p. 7).

Chin-Parker, S., & Bradner, A. (2010). Background shifts affect explanatory style: How a pragmatic theory of explanation accounts for background effects in the generation of explanations. *Cognitive Processing*, *11*(3), 227–249. https://doi.org/10.1007/s10339-009-0341-4 (cit. on p. 164).

Chiou, E. K., & Lee, J. D. (2023). Trusting Automation: Designing for Responsivity and Resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995 (cit. on pp. 172, 190).

Chrisley, R. (2003). Embodied artificial intelligence. *Artificial Intelligence*, *149*(1), 131–150. https://doi.org/10.1016/S0004-3702(03)00055-9 (cit. on p. 6).

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 307–317. https://doi.org/10.1145/3397481.3450644 (cit. on pp. 41, 162, 185, 187, 190).

Chromik, M., Eiband, M., Völkel, S., & Buschek, D. (2019). Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. Retrieved August 16, 2024, from https://www.semanticscholar.org/paper/Dark-Patterns-of-Explainability%2C-Transparency%2C-and-Chromik-Eiband/b289d835b8d8d29bbdc8529530a8919a98139e10 (cit. on p. 36).

Colin, J., Fel, T., Cadene, R., & Serre, T. (2022). What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. *Advances in Neural Information Processing Systems*, *35*, 2832–2845. Retrieved August 16, 2024, from https://proceedings.neurips.cc/paper_files/paper/2022/hash/13113e938f2957891c0c5e8df811dd01-Abstract-Conference.html (cit. on p. 39).

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496. https://doi.org/10.1007/s11257-008-9051-3 (cit. on p. 178).

Crook, B., Schlüter, M., & Speith, T. (2023). Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI). *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, 316–324. https://doi.org/10.1109/REW57809.2023.00060 (cit. on pp. 27, 190).

Dang, H., Mecke, L., & Buschek, D. (2022). GANSlider: How Users Control Generative Models for Images using Multiple Sliders with and without Feedforward Information. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3491102.3502141 (cit. on p. 186).

Davis, F. D., et al. (1989). Technology acceptance model: TAM. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption*, *205*, 219 (cit. on p. 33).

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008 (cit. on pp. 29, 33).

Deci, E. L., & Ryan, R. M. (1987). The support of autonomy and the control of behavior. *Journal of personality and social psychology*, *53*(6), 1024 (cit. on p. 14).

Dekker, S. W. A., & Woods, D. D. (2002). MABA-MABA or Abracadabra? Progress on Human-Automation Co-ordination. *Cognition, Technology & Work*, *4*(4), 240–244. https://doi.org/10.1007/s101110200022 (cit. on p. 17).

Diabeloop DBLG1 System Overview. (2024). dbl-diabetes. Retrieved August 16, 2024, from https://www.dbl-diabetes.com/dblg1-system (cit. on p. 193).

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, *99*, 101896. https://doi.org/10.1016/j.inffus.2023.101896 (cit. on p. 31).

DIN EN ISO 9241-11:2018-11, Ergonomie der Mensch-System-Interaktion_ - Teil_ 11: Gebrauchstauglichkeit: Begriffe und Konzepte (ISO_ 9241-11:2018); Deutsche Fassung EN_ ISO_ 9241-11:2018. (2018). https://doi.org/10.31030/2757945 (cit. on p. 33).

DIN SPEC 92001-3:2023-08, Künstliche Intelligenz_ - Life Cycle Prozesse und Qualitätsanforderungen_ - Teil_ 3: Erklärbarkeit; Text Englisch. (2023). https://doi.org/10.31030/3446955 (cit. on pp. 24, 25).

Doran, D., Schulz, S., & Besold, T. R. (2017). *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives*. arXiv: 1710.00794 [cs]. Retrieved August 13, 2024, from http://arxiv.org/abs/1710.00794 (cit. on p. 24).

Doshi-Velez, F., & Kim, B. (2017, March 2). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv: 1702.08608 [cs, stat]. Retrieved August 10, 2024, from http://arxiv.org/abs/1702.08608 (cit. on pp. 11, 25, 184).

Duch, W., Setiono, R., & Zurada, J. (2004). Computational intelligence methods for rule-based data understanding. *Proceedings of the IEEE*, *92*(5), 771–805. https://doi.org/10.1109/JPROC.2004.826605 (cit. on p. 11).

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.*, *55*(9), 194:1–194:33. https://doi.org/10.1145/3561048 (cit. on p. 182).

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7 (cit. on p. 30).

Edgar, G. K., Catherwood, D., Baker, S., Sallis, G., Bertels, M., Edgar, H. E., Nikolla, D., Buckle, S., Goodwin, C., & Whelan, A. (2018). Quantitative Analysis of Situation Awareness (QASA): Modelling and measuring situation awareness using signal detection theory. *Ergonomics*, *61*(6), 762–777. https://doi.org/10.1080/00140139.2017.1420238 (cit. on p. 32).

Ehsan, U., & Riedl, M. O. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In C. Stephanidis, M. Kurosu, H. Degen, & L. Reinerman-Jones (Eds.), *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence* (pp. 449–466). Springer International

Publishing. https://doi.org/10.1007/978-3-030-60117-1_33 (cit. on pp. 36, 188).

Ehsan, U., & Riedl, M. O. (2024). Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns*, *5*(6). https://doi.org/10.1016/j.patter.2024.100971 (cit. on p. 36).

Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., Riener, A., & Riedl, M. O. (2021). Operationalizing Human-Centered Perspectives in Explainable AI. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3411763.3441342 (cit. on p. 2).

Endsley, M., Sollenberger, R., & Stein, E. (2000). Situation awareness: A comparison of measures. *Proceedings of the Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium, Savannah, GA* (cit. on p. 31).

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543 (cit. on pp. 3, 18, 32, 37, 41).

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned From Human–Automation Research. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350 (cit. on pp. 17, 32, 37).

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). A Comparative Analysis of Sagat and Sart for Evaluations of Situation Awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(1), 82–86. https://doi.org/10.1177/154193129804200119 (cit. on p. 31).

Estes, W. K. (Ed.). (2014). *Handbook of learning and cognitive processes. Volume 1, Introduction to concepts and issues*. Psychology Press. (Cit. on p. 5). OCLC: 881840355.

European Union. (2024). Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 on the harmonisation of laws, regulations, and administrative provisions of the member states as regards artificial intelligence (ai act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689 (cit. on pp. 2, 7, 21, 36, 192).

Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., & Holzinger, A. (2022). The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems*, *133*, 281–296. https://doi.org/10.1016/j.future.2022.03.009 (cit. on p. 178).

Ferrell, W. R., & Sheridan, T. B. (1967). Supervisory control of remote manipulation. *IEEE Spectrum*, *4*(10), 81–88. https://doi.org/10.1109/MSPEC.1967.5217126 (cit. on p. 12).

Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's Law in action. *GigaScience*, *8*(6), giz053. https://doi.org/10.1093/gigascience/giz053 (cit. on p. 24).

Fitts, P. M. (Ed.). (1951). *Human engineering for an effective air-navigation and traffic-control system*. National Research Council, Div. of. (Cit. on p. 17).

Franke, T., Attig, C., & Wessel, D. (2019). A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, *35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150 (cit. on p. 39).

Gajos, K. Z., & Mamykina, L. (2022). Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 794–806. https://doi.org/10.1145/3490099.3511138 (cit. on p. 21).

Gianfagna, L., & Di Cecco, A. (2021). Explainable AI: Needs, Opportunities, and Challenges. In L. Gianfagna & A. Di Cecco (Eds.), *Explainable AI with Python* (pp. 27–46). Springer International Publishing. https://doi.org/10.1007/978-3-030-68640-6_2 (cit. on p. 184).

Gil, M., Albert, M., Fons, J., & Pelechano, V. (2019). Designing human-in-the-loop autonomous Cyber-Physical Systems. *International Journal of Human-Computer Studies*, *130*, 21–39. https://doi.org/10.1016/j.ijhcs.2019.04.006 (cit. on p. 8).

Gilbert, S. (2024). The EU passes the AI Act and its implications for digital medicine are unclear. *npj Digital Medicine*, *7*(1), 135. https://doi.org/10.1038/s41746-024-01116-6 (cit. on p. 2).

Godin, G., Germain, M., Conner, M., Delage, G., & Sheeran, P. (2014). Promoting the return of lapsed blood donors: A seven-arm randomized controlled trial

of the question–behavior effect. *Health Psychology*, *33*(7), 646–655. https://doi.org/10.1037/a0033505 (cit. on p. 31).

Göndöcs, D., & Dörfler, V. (2024). AI in medical diagnosis: AI prediction & human judgment. *Artificial Intelligence in Medicine*, *149*, 102769. https://doi.org/10.1016/j.artmed.2024.102769 (cit. on p. 2).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT press. (Cit. on p. 11).

Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). *A Simple and Effective Model-Based Variable Importance Measure*. arXiv: 1805.04755 [cs, stat]. Retrieved August 16, 2024, from http://arxiv.org/abs/1805.04755 (cit. on p. 181).

Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence Program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850 (cit. on pp. 2, 24).

Guttmann, M., & Ge, M. (2024). Research Agenda of Ethical Recommender Systems based on Explainable AI. *Procedia Computer Science*, *238*, 328–335. https://doi.org/10.1016/j.procs.2024.06.032 (cit. on p. 29).

Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, *186*, 122120. https://doi.org/10.1016/j.techfore.2022.122120 (cit. on p. 34).

Harbarth, L., Gößwein, E., Bodemer, D., & Schnaubert, L. (2024). (Over)Trusting AI Recommendations: How System and Person Variables Affect Dimensions of Complacency. *International Journal of Human–Computer Interaction*, *0*(0), 1–20. https://doi.org/10.1080/10447318.2023.2301250 (cit. on p. 185).

Hardy, M., & Harvey, H. (2020). Artificial intelligence in diagnostic imaging: Impact on the radiography profession. *The British Journal of Radiology*, *93*(1108), 20190840. https://doi.org/10.1259/bjr.20190840 (cit. on p. 7).

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909 (cit. on p. 32).

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2022). *On the Effect of Information Asymmetry in Human-AI Teams*. arXiv: 2205.01467

[cs]. Retrieved July 9, 2024, from http://arxiv.org/abs/2205.01467 (cit. on p. 170).

High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI* (Report). European Commission. Brussels. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (cit. on p. 192).

Ho, C., Soon, D., Caals, K., & Kapur, J. (2019). Governance of automated image analysis and artificial intelligence analytics in healthcare. *Clinical Radiology*, *74*(5), 329–337. https://doi.org/10.1016/j.crad.2019.02.005 (cit. on p. 2).

Hoesterey, S., & Onnasch, L. (2023). The effect of risk on trust attitude and trust behavior in interaction with information and decision automation. *Cognition, Technology & Work*, *25*(1), 15–29. https://doi.org/10.1007/s10111-022-00718-y (cit. on pp. 19, 178).

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570 (cit. on pp. 30, 31).

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1096257 (cit. on pp. 3, 27, 28, 33, 36).

Holzinger, A. (2018). From Machine Learning to Explainable AI. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 55–66. https://doi.org/10.1109/DISA.2018.8490530 (cit. on p. 24).

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the Quality of Explanations: The System Causability Scale (SCS). *KI - Künstliche Intelligenz*, *34*(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z (cit. on pp. 27, 36).

Hopkins, D., & Schwanen, T. (2021). Talking about automated vehicles: What do levels of automation do? *Technology in Society*, *64*, 101488. https://doi.org/10.1016/j.techsoc.2020.101488 (cit. on p. 13).

Hornbæk, K., & Oulasvirta, A. (2017). What Is Interaction? *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5040–5052. https://doi.org/10.1145/3025453.3025765 (cit. on p. 12).

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI.

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 624–635. https://doi.org/10.1145/3442188.3445923 (cit. on p. 178).

Jensen III, C. J., McElreath, D. H., & Graves, M. (2022). *Introduction to intelligence studies*. Routledge. (Cit. on p. 6).

Johnson, M., & Bradshaw, J. M. (2021). How Interdependence Explains the World of Teamwork. In W. F. Lawless, J. Llinas, D. A. Sofge, & R. Mittu (Eds.), *Engineering Artificially Intelligent Systems: A Systems Engineering Approach to Realizing Synergistic Capabilities* (pp. 122–146). Springer International Publishing. https://doi.org/10.1007/978-3-030-89385-9_8 (cit. on p. 22).

Johnson, M., Bradshaw, J. M., Feltovich, P., Jonker, C., van Riemsdijk, B., & Sierhuis, M. (2012). Autonomy and interdependence in human-agent-robot teams. *IEEE Intelligent Systems*, *27*(2), 43–51. https://doi.org/10.1109/MIS.2012.1 (cit. on p. 187).

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of Human-Robot Interaction*, *3*(1), 43. https://doi.org/10.5898/JHRI.3.1.Johnson (cit. on pp. 16, 22).

Johnson, P. E., Duran, A. S., Hassebrock, F., Moller, J., Prietula, M., Feltovich, P. J., & Swanson, D. B. (1981). Expertise and Error in Diagnostic Reasoning*. *Cognitive Science*, *5*(3), 235–283. https://doi.org/10.1207/s15516709cog0503_3 (cit. on p. 177).

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research*, *32*(3), 713–735. https://doi.org/10.1287/isre.2020.0980 (cit. on p. 8).

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113–153. https://doi.org/10.1080/1463922021000054335 (cit. on pp. 15, 32).

Kadir, M. A., Mosavi, A., & Sonntag, D. (2023). Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, 000111–000124. https://doi.org/10.1109/INES59282.2023.10297629 (cit. on p. 27).

Kambhampati, S. (2020). Challenges of Human-Aware AI Systems: AAAI Presidential Address. *AI Magazine*, *41*(3), 3–17. https://doi.org/10.1609/aimag.v41i3.5257 (cit. on pp. 26, 176, 187).

Kern, L., & Doherty, M. E. (1982). 'Pseudodiagnosticity' in an idealized medical problem-solving environment: *Academic Medicine*, *57*(2), 100–4. https://doi.org/10.1097/00001888-198202000-00004 (cit. on p. 162).

Klein, G., Woods, D., Bradshaw, J., Hoffman, R., & Feltovich, P. (2004). Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems*, *19*(06), 91–95. https://doi.org/10.1109/MIS.2004.74 (cit. on p. 187).

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A Data–Frame Theory of Sensemaking. In *Expertise Out of Context*. Psychology Press. (Cit. on pp. 171, 172).

Kochkach, A., Kacem, S. B., Elkosantini, S., Lee, S. M., & Suh, W. (2024). On the Different Concepts and Taxonomies of eXplainable Artificial Intelligence. In A. Bennour, A. Bouridane, & L. Chaari (Eds.), *Intelligent Systems and Pattern Recognition* (pp. 75–85). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-46338-9_6 (cit. on p. 25).

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.604977 (cit. on pp. 30, 31).

Kühl, N., Goutier, M., Hirt, R., & Satzger, G. (2020). *Machine Learning in Artificial Intelligence: Towards a Common Understanding* (1). https://doi.org/10.48550/ARXIV.2004.04686 (cit. on p. 7).

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473. https://doi.org/10.1016/j.artint.2021.103473 (cit. on p. 31).

Le, P. Q., Nauta, M., Nguyen, V. B., Pathak, S., Schlötterer, J., & Seifert, C. (2023). Benchmarking eXplainable AI - A Survey on Available Toolkits and Open Challenges. *Proceedings of the Thirty-Second International Joint Conference*

on Artificial Intelligence, 6665–6673. https://doi.org/10.24963/ijcai.2023/747 (cit. on p. 34).

Le, T., Miller, T., Singh, R., & Sonenberg, L. (2023). Explaining Model Confidence Using Counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(10), 11856–11864. https://doi.org/10.1609/aaai.v37i10.26399 (cit. on p. 184).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 (cit. on pp. 30, 177).

Levine, A. B., Schlosser, C., Grewal, J., Coope, R., Jones, S. J., & Yip, S. (2019). Rise of the Machines: Advances in Deep Learning for Cancer Diagnosis. *Trends in Cancer*, *5*(3), 157–169. https://doi.org/10.1016/j.trecan.2019.02.002 (cit. on p. 1).

Lewicki, R. J., & Bunker, B. B. (1995). Trust in relationships: A model of development and decline. In *Conflict, cooperation, and justice: Essays inspired by the work of Morton Deutsch* (pp. 133–173). Jossey-Bass/Wiley. (Cit. on p. 29).

Lin, H., Li, R., Liu, Z., Chen, J., Yang, Y., Chen, H., Lin, Z., Lai, W., Long, E., Wu, X., Lin, D., Zhu, Y., Chen, C., Wu, D., Yu, T., Cao, Q., Li, X., Li, J., Li, W., ... Liu, Y. (2019). Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. *EClinicalMedicine*, *9*, 52–59. https://doi.org/10.1016/j.eclinm.2019.03.001 (cit. on p. 8).

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health*, *1*(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2 (cit. on p. 9).

Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human Mental Workload: A Survey and a Novel Inclusive Definition. *Frontiers in Psychology*, *13*, 883321. https://doi.org/10.3389/fpsyg.2022.883321 (cit. on p. 32).

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). *Explainable AI for Trees:*

From Local Explanations to Global Understanding. arXiv: 1905.04610 [cs, stat]. Retrieved August 16, 2024, from http://arxiv.org/abs/1905.04610 (cit. on pp. 182, 184).

Lyell, D., Coiera, E., Chen, J., Shah, P., & Magrabi, F. (2021). How machine learning is embedded to support clinician decision making: An analysis of FDA-approved medical devices. *BMJ Health & Care Informatics*, *28*(1), e100301. https://doi.org/10.1136/bmjhci-2020-100301 (cit. on p. 22).

Lyratzopoulos, G., Vedsted, P., & Singh, H. (2015). Understanding missed opportunities for more timely diagnosis of cancer in symptomatic patients after presentation. *British Journal of Cancer*, *112*(S1), S84–S91. https://doi.org/10.1038/bjc.2015.47 (cit. on p. 9).

Mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. (2002). The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, *24*(1), 51–70. https://doi.org/10.1016/S0933-3657(01)00092-6 (cit. on p. 7).

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th australasian conference on information systems*, *53*, 6–8 (cit. on pp. 30, 33, 40).

Malle, B. F. (2022). Attribution Theories. In *Theories in Social Psychology, Second Edition* (pp. 93–120). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781394266616.ch4 (cit. on p. 164).

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709. https://doi.org/10.2307/258792 (cit. on p. 29).

McCarthy, J. (2007). From here to human-level AI. *Artificial Intelligence*, *171*(18), 1174–1182. https://doi.org/10.1016/j.artint.2007.10.009 (cit. on p. 5).

McLellan, J., Heneghan, C., Roberts, N., & Pluddemann, A. (2023). Accuracy of self-diagnosis in conditions commonly managed in primary care: Diagnostic accuracy systematic review and meta-analysis. *BMJ Open*, *13*(1), e065748. https://doi.org/10.1136/bmjopen-2022-065748 (cit. on p. 15).

Meder, B., & Mayrhofer, R. (2017). Diagnostic reasoning. *Oxford handbook of causal reasoning*, 433–458 (cit. on p. 1).

Merritt, S. M., Sinha, R., Curran, P. G., & Ilgen, D. R. (2015). Attitudinal predictors of relative reliance on human vs. automated advisors. *International Journal*

of Human Factors and Ergonomics, *3*(3-4), 327–345. https://doi.org/10.1504/IJHFE.2015.072982 (cit. on p. 31).

Miller, E. M. (2020). Using Continuous Glucose Monitoring in Clinical Practice. *Clinical Diabetes*, *38*(5), 429–438. https://doi.org/10.2337/cd20-0043 (cit. on p. 15).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007 (cit. on pp. 24, 26, 161, 165).

Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–342. https://doi.org/10.1145/3593013.3594001 (cit. on pp. 21, 22, 26).

Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE*, *49*(1), 8–30. https://doi.org/10.1109/JRPROC.1961.287775 (cit. on p. 6).

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044 (cit. on p. 6).

Molloy, R., & Parasuraman, R. (1996). Monitoring an Automated System for a Single Failure: Vigilance and Task Complexity Effects. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *38*(2), 311–322. https://doi.org/10.1177/001872089606380211 (cit. on p. 14).

Molnar, C. (2022). *Interpretable machine learning. A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book (cit. on p. 181).

Moradbakhti, L., Leichtmann, B., & Mara, M. (2024). Development and Validation of a Basic Psychological Needs Scale for Technology Use. *Psychological Test Adaptation and Development*, *5*, 26–45. https://doi.org/10.1027/2698-1866/a000062 (cit. on p. 194).

Müller, H., & Holzinger, A. (2021). Kandinsky Patterns. *Artificial Intelligence*, *300*, 103546. https://doi.org/10.1016/j.artint.2021.103546 (cit. on p. 159).

Naugler, C., & Church, D. L. (2019). Automation and artificial intelligence in the clinical laboratory. *Critical Reviews in Clinical Laboratory Sciences*, *56*(2), 98–110. https://doi.org/10.1080/10408363.2018.1561640 (cit. on p. 2).

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *Acm Computing Surveys*. https://doi.org/10.1145/3583558 (cit. on p. 27).

Nelson, J. D. (2005). Finding Useful Questions: On Bayesian Diagnosticity, Probability, Impact, and Information Gain. *Psychological Review*, *112*(4), 979–999. https://doi.org/10.1037/0033-295X.112.4.979 (cit. on pp. 160, 162).

Nendaz, M., & Perrier, A. (2012). Diagnostic errors and flaws in clinical reasoning: Mechanisms and prevention in practice. *Swiss Medical Weekly*. https://doi.org/10.4414/smw.2012.13706 (cit. on p. 1).

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning: Diagnostic error and reasoning. *Medical Education*, *44*(1), 94–100. https://doi.org/10.1111/j.1365-2923.2009.03507.x (cit. on p. 1).

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(3), 476–488. https://doi.org/10.1177/0018720813501549 (cit. on pp. 14, 15, 19, 21, 22).

OpenAI. (2023). GPT-4: OpenAI's large language model. https://openai.com/gpt-4 (cit. on pp. 183, 195).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730–27744 (cit. on p. 19).

Palacio, S., Lucieri, A., Munir, M., Ahmed, S., Hees, J., & Dengel, A. (2021). XAI handbook: Towards a unified framework for explainable AI. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 3766–3775 (cit. on p. 24).

Paleja, R., Ghuy, M., Arachchige, N. R., Jensen, R., & Gombolay, M. (2022). *The Utility of Explainable AI in Ad Hoc Human-Machine Teaming*. arXiv: 2209.03943 [cs]. https://doi.org/10.48550/arXiv.2209.03943 (cit. on p. 33).

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations

in AI. *ACM Transactions on Computer-Human Interaction*, *29*(4), 1–33. https://doi.org/10.1145/3495013 (cit. on pp. 19, 35).

Parasuraman, R., Sheridan, T., & Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354 (cit. on pp. 3, 7, 9, 13, 14, 166).

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160. https://doi.org/10.1518/155534308X284417 (cit. on p. 29).

Pfaff, M. S., Klein, G. L., Drury, J. L., Moon, S. P., Liu, Y., & Entezari, S. O. (2013). Supporting Complex Decision Making Through Option Awareness. *Journal of Cognitive Engineering and Decision Making*, *7*(2), 155–178. https://doi.org/10.1177/1555343412455799 (cit. on pp. 22, 175).

Pineau, J., Montemerlo, M., Pollack, M., Roy, N., & Thrun, S. (2003). Towards robotic assistants in nursing homes: Challenges and results. *Robotics and Autonomous Systems*, *42*(3-4), 271–281. https://doi.org/10.1016/S0921-8890(02)00381-0 (cit. on p. 8).

Popper, K. (2002). *The Logic of Scientific Discovery* (2nd ed.). Routledge. https://doi.org/10.4324/9780203994627 (cit. on p. 23).

Poulton, E. C. (1966). Engineering Psychology. *Annual Review of Psychology*, *17*(1), 177–200. https://doi.org/10.1146/annurev.ps.17.020166.001141 (cit. on pp. 6, 195).

Raisamo, R., Rakkolainen, I., Majaranta, P., Salminen, K., Rantala, J., & Farooq, A. (2019). Human augmentation: Past, present and future. *International Journal of Human-Computer Studies*, *131*, 131–143. https://doi.org/10.1016/j.ijhcs.2019.05.008 (cit. on p. 7).

Ramkumar, A., Stappers, P. J., Niessen, W. J., Adebahr, S., Schimek-Jasch, T., Nestle, U., & Song, Y. (2017). Using GOMS and NASA-TLX to Evaluate Human–Computer Interaction Process in Interactive Segmentation. *International Journal of Human–Computer Interaction*, *33*(2), 123–134. https://doi.org/10.1080/10447318.2016.1220729 (cit. on p. 32).

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., GI Genius CADx Study Group, Antonelli, G., Awadie, H., Bernhofer, S., Carballal, S., Dinis-Ribeiro, M., Fernández-Clotett, A., Esparrach, G. F., Gralnek, I., Higasa, Y., Hirabayashi, T., Hirai, T., Iwatate, M., Kawano, M., ... Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, *12*(1), 14952. https://doi.org/10.1038/s41598-022-18751-2 (cit. on p. 8).

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study. *Proceedings of the 34th International Conference on Machine Learning*, 2940–2949. Retrieved August 16, 2024, from https://proceedings.mlr.press/v70/ritter17a.html (cit. on pp. 165, 166).

Roberts, H., Babuta, A., Morley, J., Thomas, C., Taddeo, M., & Floridi, L. (2022). Artificial Intelligence Regulation in the United Kingdom: A Path to Global Leadership? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4209504 (cit. on p. 192).

Rong, Y., Leemann, T., Borisov, V., Kasneci, G., & Kasneci, E. (2022). *A Consistent and Efficient Evaluation Strategy for Attribution Methods*. arXiv: 2202.00449 [cs]. https://doi.org/10.48550/arXiv.2202.00449 (cit. on p. 27).

Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). AI and XAI second opinion: The danger of false confirmation in human–AI collaboration. *Journal of Medical Ethics*, jme-2024–110074. https://doi.org/10.1136/jme-2024-110074 (cit. on p. 20).

Roth, E. M., Sushereba, C., Militello, L. G., Diiulio, J., & Ernst, K. (2019). Function Allocation Considerations in the Era of Human Autonomy Teaming. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 199–220. https://doi.org/10.1177/1555343419878038 (cit. on p. 16).

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1 (cit. on p. 41).

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson. (Cit. on p. 5).

Saffer, D. (2010). *Designing for interaction: Creating innovative applications and devices* (2nd ed). New Riders. (Cit. on p. 12).
OCLC: ocn403417500.

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics*, *39*(3), 490–500. https://doi.org/10.1016/j.ergon.2008.10.010 (cit. on p. 196).

Sarkar, A., Vijaykeerthy, D., Sarkar, A., & Balasubramanian, V. N. (2022). A Framework for Learning Ante-Hoc Explainable Models via Concepts, 10286–10295. Retrieved August 13, 2024, from https://openaccess.thecvf.com/content/CVPR2022/html/Sarkar_A_Framework_for_Learning_Ante-Hoc_Explainable_Models_via_Concepts_CVPR_2022_paper.html (cit. on p. 26).

Scassa, T. (2023). Regulating AI in Canada: A Critical Look at the Proposed Artificial Intelligence and Data Act. *Canadian Bar Review*, *101*, 1. https://heinonline.org/HOL/Page?handle=hein.journals/canbarev101&id=1&div=&collection= (cit. on p. 192).

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228 (cit. on p. 14).

Schelenz, L., Segal, A., Adelio, O., & Gal, K. (2024). Transparency-Check: An Instrument for the Study and Design of Transparency in AI-based Personalization Systems. *ACM Journal on Responsible Computing*, *1*(1), 1–18. https://doi.org/10.1145/3636508 (cit. on pp. 180, 196).

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617–626. https://doi.org/10.1145/3514094.3534128 (cit. on p. 35).

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., & Langer, M. (2022). *A Micro and Macro Perspective on Trustworthiness: Theoretical Underpinnings*

of the Trustworthiness Assessment Model (TrAM). https://doi.org/10.31234/osf.io/qhwvx (cit. on pp. 40, 41, 168, 177).

Schmidt, A. (2020). Interactive Human Centered Artificial Intelligence: A Definition and Research Challenges. *Proceedings of the International Conference on Advanced Visual Interfaces*, 1–4. https://doi.org/10.1145/3399715.3400873 (cit. on p. 12).

Schrills, T., & Franke, T. (2020). Color for Characters - Effects of Visual Explanations of AI on Trust and Observability. In H. Degen & L. Reinerman-Jones (Eds.), *Artificial Intelligence in HCI* (pp. 121–135, Vol. 12217). Springer International Publishing. https://doi.org/10.1007/978-3-030-50334-5_8 (cit. on p. 3).

Schrills, T., & Franke, T. (2023). How Do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems. *ACM Trans. Interact. Intell. Syst.*, *13*(4), 25:1–25:34. https://doi.org/10.1145/3588594 (cit. on pp. 37, 38, 179).

Schrills, T., Gruner, M., Peuscher, H., & Franke, T. (2023). Safe Environments to Understand Medical AI - Designing a Diabetes Simulation Interface for Users of Automated Insulin Delivery. In V. G. Duffy (Ed.), *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management* (pp. 306–328, Vol. 14029). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35748-0_23 (cit. on pp. 3, 192).

Schrills, T., Schmid, L., Jetter, H.-C., & Franke, T. (2021). An Explainability Case-Study for Conversational User Interfaces in Walk-Up-And-Use Contexts, 10.18420/muc2021. Retrieved August 16, 2024, from https://dl.gi.de/handle/20.500.12116/37352 (cit. on p. 180).

Schrills, T., Sieger, M., Gruner, M., & Franke, T. (2024). Evaluation of a Scale to Assess Subjective Information Processing Awareness of Humans in Interaction with Automation & Artificial Intelligence. https://doi.org/10.54941/ahfe1004640 (cit. on pp. 37, 39, 194).

Schrills, T., Zoubir, M., Bickel, M., Kargl, S., & Franke, T. (2021). Are users in the loop? Development of the subjective information processing awareness scale to assess XAI. *Proceedings of the ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI (HCXAI 2021), Upol Ehsan, Q. Vera Liao, Martina Mara, Mark Riedl, Andreas Riener, Marc Streit, Sandra*

Wachter, and Philipp Wintersberger (Eds.). The Internet (cit. on pp. 3, 37, 179).

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-022-00867-8 (cit. on p. 25).

Sewnath, G., & Crijnen, J. (2021). *How much is too much? Levels of AI Explainability within Decision Support Systems' User Interfaces for improved decision-making performance*. (Cit. on pp. 32, 190).

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The new system technology*. Springer. (Cit. on p. 1).
OCLC: 1367988287.

Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S. Y., Fang, P.-H., & Ming, W.-K. (2019). Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR Medical Informatics*, *7*(3), e10010. https://doi.org/10.2196/10010 (cit. on p. 177).

Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. MIT Press. (Cit. on pp. 12, 13).

Sheridan, T. B., & Ferrell, W. R. (1974). *Man-machine systems; Information, control, and decision models of human performance.* the MIT press. (Cit. on pp. 13, 14).

Sheridan, T. B., Verplank, W. L., & Brooks, T. L. (1978). Human/computer control of undersea teleoperators. *NASA. Ames Res. Center the 14th Ann. Conf. on Manual Control* (cit. on p. 12).

Sherr, J. L., Heinemann, L., Fleming, G. A., Bergenstal, R. M., Bruttomesso, D., Hanaire, H., Holl, R. W., Petrie, J. R., Peters, A. L., & Evans, M. (2023). Automated insulin delivery: Benefits, challenges, and recommendations. A Consensus Report of the Joint Diabetes Technology Working Group of the European Association for the Study of Diabetes and the American Diabetes Association. *Diabetologia*, *66*(1), 3–22. https://doi.org/10.1007/s00125-022-05744-z (cit. on p. 15).

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-*

Computer Studies, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551 (cit. on pp. 35, 178).

Shneiderman, B. (2020a). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118 (cit. on pp. 3, 161, 199).

Shneiderman, B. (2020b). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, *10*(4), 1–31. https://doi.org/10.1145/3419764 (cit. on p. 2).

Shokri, R., Strobel, M., & Zick, Y. (2019). Privacy Risks of Explaining Machine Learning Models. *ArXiv*. Retrieved August 13, 2024, from https://www.semanticscholar.org/paper/Privacy-Risks-of-Explaining-Machine-Learning-Models-Shokri-Strobel/a1fc6049caeb7ffd6ecfaaf990cacb42e7c01538 (cit. on p. 27).

Sieker, J., Junker, S., Utescher, R., Attari, N., Wersing, H., Buschmeier, H., & Zarrieß, S. (2024). *The Illusion of Competence: Evaluating the Effect of Explanations on Users' Mental Models of Visual Question Answering Systems*. arXiv: 2406.19170 [cs]. https://doi.org/10.48550/arXiv.2406.19170 (cit. on p. 189).

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2023). Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *International Journal of Human–Computer Interaction*, *39*(7), 1390–1404. https://doi.org/10.1080/10447318.2022.2101698 (cit. on p. 35).

Simon, H. A., & Newell, A. (1964). Information processing in computer and man. *American Scientist*, *52*(3), 281–300. Retrieved July 28, 2024, from http://www.jstor.org/stable/27839071 (cit. on p. 8).

Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. https://doi.org/10.1145/3531146.3534639 (cit. on pp. 25, 26).

Speith, T., Crook, B., Mann, S., Schomäcker, A., & Langer, M. (2024). Conceptualizing understanding in explainable artificial intelligence (XAI): An abilities-

based approach. *Ethics and Information Technology*, *26*(2), 40. https://doi.org/10.1007/s10676-024-09769-3 (cit. on p. 162).

Sprott, D. E., Spangenberg, E. R., Block, L. G., Fitzsimons, G. J., Morwitz, V. G., & Williams, P. (2006). The question–behavior effect: What we know and where we go from here. *Social Influence*, *1*(2), 128–137. https://doi.org/10.1080/15534510600685409 (cit. on p. 31).

Sreedharan, S. (2023). Human-aware AI —A foundational framework for human–AI interaction. *AI Magazine*, *44*(4), 460–466. https://doi.org/10.1002/aaai.12142 (cit. on p. 188).

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Farina, M. (2021). A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, *9*, 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315 (cit. on p. 2).

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2495–2507. https://doi.org/10.1145/3630106.3659051 (cit. on pp. 28, 192, 196).

Sutton, S. G., Arnold, V., & Holt, M. (2018). How Much Automation Is Too Much? Keeping the Human Relevant in Knowledge Work. *Journal of Emerging Technologies in Accounting*, *15*(2), 15–25. https://doi.org/10.2308/jeta-52311 (cit. on p. 16).

Szepannek, G., & Lübke, K. (2023). How much do we see? On the explainability of partial dependence plots for credit risk scoring. *Argumenta Oeconomica*, (1 (50)). Retrieved August 16, 2024, from https://journals.ue.wroc.pl/aoe/article/view/1047 (cit. on p. 182).

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, *28*(2), 454–475. https://doi.org/10.3758/s13423-020-01825-5 (cit. on p. 165).

Taylor, R. M. (2017). Situational Awareness Rating Technique (Sart): The Development of a Tool for Aircrew Systems Design. In E. Salas (Ed.), *Situational Awareness* (1st ed., pp. 111–128). Routledge. https://doi.org/10.4324/9781315087924-8 (cit. on pp. 32, 38).

Tedeschi, L. O. (2006). Assessment of the adequacy of mathematical models. *Agricultural Systems*, *89*(2-3), 225–247. https://doi.org/10.1016/j.agsy.2005.11.004 (cit. on p. 169).

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8 (cit. on p. 5).

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (Eds.), *Proceedings of the 4th machine learning for healthcare conference* (pp. 359–380, Vol. 106). PMLR. https://proceedings.mlr.press/v106/tonekaboni19a.html (cit. on pp. 34, 184).

Tsai, C.-H., You, Y., Gui, X., Kou, Y., & Carroll, J. M. (2021). Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3411764.3445101 (cit. on p. 32).

Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., Azizi, S., Singhal, K., Cheng, Y., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., Gottweis, J., … Natarajan, V. (2024). *Towards Conversational Diagnostic AI*. arXiv: 2401.05654 [cs]. Retrieved August 13, 2024, from http://arxiv.org/abs/2401.05654 (cit. on p. 20).

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *J. of Math*, *58*(345-363), 5 (cit. on p. 11).

Turner, R. M. (1992). A view of diagnostic reasoning as a memory-directed task. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (cit. on p. 1).

Ueno, T., Kim, Y., Oura, H., & Seaborn, K. (2023). Trust and Reliance in Consensus-Based Explanations from an Anti-Misinformation Agent. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3544549.3585713 (cit. on pp. 19, 40).

Valdez, A. C., Heine, M., Franke, T., Jochems, N., Jetter, H.-C., & Schrills, T. (2024). The European commitment to human-centered technology: The integral role

of HCI in the EU AI Act's success. *i-com*, *23*(2), 249–261. https://doi.org/10.1515/icom-2024-0014 (cit. on p. 36).

Van De Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *66*(1), 180–208. https://doi.org/10.1177/00187208221077804 (cit. on p. 28).

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404. https://doi.org/10.1016/j.artint.2020.103404 (cit. on p. 34).

Vashishth, T. K., Kumar, B., Sharma, V., Chaudhary, S., Kumar, S., & Sharma, K. K. (2023). The Evolution of AI and Its Transformative Effects on Computing: A Comparative Analysis. In B. K. Mishra (Ed.), *Advances in Civil and Industrial Engineering* (pp. 425–442). IGI Global. https://doi.org/10.4018/979-8-3693-0044-2.ch022 (cit. on p. 6).

Vidulich, M. A., & Tsang, P. S. (2012). Mental Workload and Situation Awareness. In *Handbook of Human Factors and Ergonomics* (pp. 243–273). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118131350.ch8 (cit. on p. 32).

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300831 (cit. on pp. 180, 190).

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002 (cit. on pp. 5, 11).

Wang, X., & Yin, M. (2021). Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 318–328. https://doi.org/10.1145/3397481.3450650 (cit. on p. 35).

Warren, G., Smyth, B., & Keane, M. T. (2022). "Better" Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI). In M. T. Keane & N. Wiratunga (Eds.), *Case-Based Reasoning Research and Development*

(pp. 63–78). Springer International Publishing. https://doi.org/10.1007/978-3-031-14923-8_5 (cit. on p. 184).

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models*. arXiv: 2206.07682 [cs]. Retrieved August 13, 2024, from http://arxiv.org/abs/2206.07682 (cit. on p. 19).

Weitz, K. (2022). Towards Human-Centered AI: Psychological concepts as foundation for empirical XAI research. *it - Information Technology*, *64*(1-2), 71–75. https://doi.org/10.1515/itit-2021-0047 (cit. on p. 3).

Wester, J., Schrills, T., Pohl, H., & Van Berkel, N. (2024). "As an AI language model, I cannot": Investigating LLM Denials of User Requests. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3613904.3642135 (cit. on p. 3).

White House Office of Science and Technology Policy. (2022). Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/news-updates/2022/10/04/fact-sheet-biden-harris-administration-announces-key-actions-to-advance-tech-accountability-and-protect-the-rights-of-the-american-public/ (cit. on p. 192).

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2021). *Engineering psychology and human performance*. Routledge. (Cit. on pp. 1, 6).

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and Levels of Automation: An Integrated Meta-analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 389–393. https://doi.org/10.1177/154193121005400425 (cit. on p. 14).

Wickens, C. D., & Scott, B. D. (1983). A comparison of verbal and graphical information presentation in a complex information integration decision task. In *Tech. Rep. EPL-83-1/ONR-83-1*. Engineering-Psychology Research Laboratory, University of Illinois Urbana ... (Cit. on p. 160).

Wilkenfeld, D. A., & Lombrozo, T. (2015). Inference to the Best Explanation (IBE) Versus Explaining for the Best Inference (EBI). *Science & Education*, *24*(9), 1059–1077. https://doi.org/10.1007/s11191-015-9784-4 (cit. on p. 164).

Xu, Z., Jain, S., & Kankanhalli, M. (2024, January 22). *Hallucination is Inevitable: An Innate Limitation of Large Language Models*. arXiv: 2401.11817 [cs]. Retrieved 2024, from http://arxiv.org/abs/2401.11817 (cit. on p. 19).

Yacoby, Y., Green, B., Jr, C. L. G., & Doshi-Velez, F. (2022). "If it didn't happen, why would I change my decision?": How Judges Respond to Counterfactual Explanations for the Public Safety Assessment. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *10*, 219–230. https://doi.org/10.1609/hcomp.v10i1.22001 (cit. on p. 25).

Yang, S. C.-H., Folke, N. E. T., & Shafto, P. (2022). A Psychological Theory of Explainability. *Proceedings of the 39th International Conference on Machine Learning*, 25007–25021. Retrieved August 16, 2024, from https://proceedings.mlr.press/v162/yang22c.html (cit. on p. 164).

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3544548.3581388 (cit. on p. 195).

Zhao, X., Zhang, W., Xiao, X., & Lim, B. (2021). Exploiting Explanations for Model Inversion Attacks, 682–692. Retrieved August 13, 2024, from https://openaccess.thecvf.com/content/ICCV2021/html/Zhao_Exploiting_Explanations_for_Model_Inversion_Attacks_ICCV_2021_paper.html (cit. on p. 27).

A List of Figures

2.1 Different levels of Situation Awareness, their corresponding levels of SIPA and system properties connected to SIPA levels 38

8.1 The conceptual model of integrated information processing 166
8.2 Nested loop of human action regulation to reach preferred level of input adequacy 170
8.3 Nested loop of human action regulation to reach preferred level of reference consonance 173
8.4 Nested loop of human action regulation to reach preferred level of output diagnosticity 176

B List of Tables

2.1 Items of the Subjective Information Processing Awareness (SIPA) Scale and the Corresponding Instruction 39

8.1 Examples of low levels of input adequacy with examples from the context of AID systems (referring to study 2) 171
8.2 Examples of low levels of reference consonance with examples from the context of AID systems (referring to study 2) 174
8.3 Examples of low levels of output diagnosticity with examples from the context of AID systems (referring to study 2) 177

Acronyms

AI Artificial Intelligence.

AIA European Union Artificial Intelligence Act.

AID Automated Insulin Delivery.

DL Deep Learning.

DSS Decision Support System.

ESS Explanation Satisfaction Scale.

HAI Human-AI Interaction.

HASO Human-Autonomy System Oversight.

HCI Human-Computer Interaction.

HCXAI Human-centered Explainable Artificial Intelligence.

IIP Integrated Information Processing Model.

IPA Information Processing Awareness.

LOA Levels of Automation.

NASA-TLX NASA Task Load Index.

SAGAT Situation Awareness Global Assessment Technique.

SAT situation-awareness-based-transparency.

SCS System Causability Scale.

SIPA Subjective Information Processing Awareness.

XAI Explainable Artificial Intelligence.

Personal Information

Name and Title	Tim Philipp Peter Schrills, B.Sc., M.Sc.
Birth Details	born on 26.11.1994
Family Status	married
Nationality	german
Adress	Universität zu Lübeck Institut Multimediale und Interaktive Systeme Ingenieurpsychologie & Kognitive Ergonomie Ratzeburger Alle 160 23562 Lübeck
Phone	045131015135
E-Mail	tim.schrills@uni-luebeck.de

Education

10/2013 - 07/2017	**Bachelor of Science, Psychology (1.6)** Heinrich-Heine University, Düsseldorf
10/2017 - 10/2019	**Master of Science, Psychology (1.1)** University Duisburg-Essen
From 10/2019	**PhD Candidate** University of Lübeck

Projects

08/2021 - 06/2022	**Development of an Intervention to Achieve Optimal Trust in Automated Insulin Delivery** Project Acquisition, Project Coordination
02/2019 - 02/2021	**Nutzerzentrierte Potenzialsteigerung von E-CarSharing in Schleswig-Holstein** Study Coordination, Project Member

01/2020 - 12/2022	**Reallabor Nutzerzentriertes Bidirektionales Laden** Project Acquisition, Project Member
03/2020 - 09/2020	**Climate Crafting** Project Acquisition, Project Member
03/2020 - 01/2023	**Cooperative and Communicative AI for Medical Image-Guided Diagnostics** Project Acquisition, Project Coordination
01/2021 - 12/2021	**eGov-Campus: Lernplattform für E-Government** Project Member, Book Publication
06/2021 - 05/2024	**Multi-Agent-Simulation of Intelligent Resource Regulation in Integrated Energy and Mobility** Project Acquisition, Project Member
08/2021 - 07/2024	**Mensch-Maschine-Schnittstellen für kooperative Ressourcenregulation bei der intelligenten Elektromobilität** Project Member
From 04/2020	**Nutzer:innenzentrierte Integration von On-Demand-Ridepooling in den ÖPNV (3 Phases)** Project Acquisition, Project Coordination
From 11/2022	**KI-gestützte Theranostik von Frakturen im Kindes- und Jugendalter für alle Ärztinnen und Ärzte** Project Coordination
From 01/2023	**Wegweiser UX für KI: Online-Kompetenzaufbau: UX für gemeinwohlorientierte KI** Project Member
From 05/2024	**Adaptive KI-Unterstützung für effiziente Makleranwendungen** Project Acquisition, Project Coordination

Teaching

WS21 & WS22	**Engineering Psychology** Seminar
	Specialization profile in media informatics
SS23	**Emotion and Motivation psychology** Seminar