Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support

Pei-Yu Chen

Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support

Pei-Yu CHEN

Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support

Dissertation

for the purpose of attaining the degree of doctor
at Delft University of Technology,
by the authority of Rector Magnificus Prof.dr.ir. H. Bijl
chair of the Board for Doctorates,
to be defended publicly on
Monday, 9 March 2026, at 10:00 o'clock

by

Pei-Yu CHEN

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. C.M. Jonker	Delft University of Technology, *promotor*
Prof. dr. D.K.J. Heylen	University of Twente, *promotor*
Dr. M.L. Tielman	Delft University of Technology, *copromotor*
Dr. M.B. van Riemsdijk	University of Twente, *copromotor*
Independent members:	
Prof. dr. O.E. Scharenborg	Delft University of Technology
Prof. dr. D. Richards	Macquarie University, Australia
Prof. dr. H.B. Verheij	University of Groningen
Dr. S. Wang	University of Twente
Prof. dr. M.A. Neerincx	Delft University of Technology, reserve member

The thesis has been funded by the Hybrid Intelligence Gravitation Programme, which is supported by the Netherlands Organization for Scientific Research (NWO) (project number 024.004.022).

SIKS Dissertation Series No. 2026-01.

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: Alignment dialogue, Human-agent alignment, Behavior support technology, Behavior Change, Conversational Agent, User modeling, Vlues

Printed by: Proefschriftspecialist | https://www.proefschriftspecialist.nl/

Cover by: Peng Lee

Style: TU Delft House Style, with modifications by Moritz Beller
https://github.com/Inventitech/phd-thesis-template

ISBN 978-94-6518-254-4

An electronic version of this dissertation is available at http://repository.tudelft.nl/.

Contents

Summary **ix**

Samenvatting **xi**

中文摘要 **xv**

1 Introduction **1**

1.1 Related Concept and Literature . . . 3

1.1.1 Behavior support technologies . . . 3

1.1.2 User modeling and personalization . . . 7

1.1.3 Shared mental models . . . 7

1.1.4 AI alignment . . . 8

1.1.5 Dialogue systems . . . 11

1.2 Alignment Dialogues: An Interactive Approach to Alignment in Support Agents . . . 13

1.2.1 Conceptual foundations . . . 13

1.2.2 Alignment dialogue as an alternative to other approaches . . . 13

1.2.3 Motivations . . . 14

1.3 Methodologies and Research Questions . . . 16

1.3.1 Research overview . . . 16

1.3.2 Research questions . . . 17

1.4 Thesis Structure . . . 21

2 Acquiring Semantic Knowledge via Human-Agent Alignment Dialogues **23**

2.1 Introduction . . . 23

2.2 Related Work . . . 25

2.2.1 Conversational agents . . . 25

2.2.2 Human-agent teamwork . . . 25

2.3 Human-Agent Alignment Dialogues . . . 26

2.4 User Study . . . 28

2.4.1 Focus group with scenarios . . . 28

2.4.2 Participants . . . 28

2.4.3 Materials . . . 28

2.4.4 Procedure . . . 29

2.4.5 Data analysis method . . . 30

2.5 Results . . . 31

2.5.1 Dialogue variants and how they are perceived differently . . . 31

2.5.2 Tree of codes . . . 32

2.5.3 Connections between the themes . . . 32

2.6 Discussion & Conclusion . . . 35

3 Conversational Strategies for Agents to Understand Users' Reasons 37
3.1 Introduction 38
3.2 Related work 40
3.2.1 Human-Agent Alignment Dialogues and Knowledge-based User Modeling 40
3.2.2 Self-conscious emotions 42
3.3 Dialogue Design 42
3.3.1 Use Case Scenario 42
3.3.2 Dialogue Design Choices 43
3.3.3 Hypotheses 44
3.4 Method. 45
3.4.1 Participants 45
3.4.2 Procedure 45
3.4.3 Measures 46
3.4.4 Text Responses Analysis Method 47
3.5 Results 48
3.5.1 Self-Conscious Emotions Analysis 48
3.5.2 Dialogue Experience Questionnaire (DEQ), Accuracy and Completeness 50
3.5.3 Text responses analysis. 52
3.5.4 Patterns in User Explanations 54
3.6 Discussion 57
3.6.1 Emotions in Different Dialogue Strategies 57
3.6.2 DEQ, Accuracy, and Completeness in Different Dialogue Strategies 57
3.6.3 Text Responses Comparison between Exploratory Dialogues and Focused Dialogues 59
3.6.4 New Topics for User Model. 60
3.6.5 Pitfalls of Alignment Dialogues 61
3.7 Reflection on Alignment Dialogue Design 61
3.7.1 Non-adherence Scenarios 62
3.7.2 Dialogue Paths for Alignment Dialogue 62
3.8 Limitations and Future work 63
3.8.1 Emotions in Alignment Dialogues 63
3.8.2 Alignment Dialogue Design 63
3.9 Conclusion 64

4 Expert Insights on Conversational AI Systems as an Information Intermediary 65
4.1 Introduction 65
4.2 Related Work. 67
4.2.1 User Modeling 67
4.2.2 Ecological Momentary Assessment 67

4.3 Conversational Intermediary AI functionalities. 67
4.4 Methods 68
4.4.1 Participants 68
4.4.2 Material 69
4.4.3 Procedure 70
4.4.4 Data Collection and Analysis. 70
4.5 Results 70
4.5.1 Part I: Insights into Current Practices in Diabetes Care 71
4.5.2 Part II: Experts' Viewpoints on the Envisioned AI Scenario. 72
4.6 Summary and Future Work. 73
4.6.1 Summary. 73
4.6.2 Future Research Directions. 74

5 Presenting User Behavior Information Collected by a Conversational Agent 75
5.1 Introduction 76
5.2 Related Work. 77
5.2.1 User modeling for behavior change support systems 77
5.2.2 Conversational agent for collecting user information. 78
5.2.3 Information presentation and comprehension 79
5.3 User study 80
5.3.1 Participants 80
5.3.2 Experiment Setup 80
5.3.3 Presentation Format Design 81
5.3.4 Measures. 83
5.3.5 Procedure 87
5.4 Results 87
5.4.1 Accuracy of SAGAT questions 88
5.4.2 Response Time of SAGAT questions 90
5.4.3 Usefulness 92
5.5 Discussion 95
5.5.1 Effect of presentation formats 96
5.5.2 Effect of information volume 96
5.5.3 Interaction between presentation formats and information volume. 97
5.5.4 Effect of health and coaching knowledge 97
5.5.5 Learning effect of the 360° tool 97
5.5.6 Variation between scenarios 98
5.5.7 Open questions. 98
5.5.8 Limitations and Future Directions 99
5.6 Conclusion 100

6 Proof of Concept: Intelligent Support Systems for Lifestyle Change 101
6.1 Introduction 101
6.2 System Overview 102
6.2.1 Dialogue 102
6.2.2 Information Extraction 103

6.2.3 User KG . . . 103
6.2.4 Domain KG . . . 103
6.2.5 Reasoning Engine . . . 104
6.3 Demonstration and Use Case . . . 104
6.4 Future Work and Conclusion . . . 104

7 Conclusion 107
7.1 Conclusion . . . 108
7.2 Limitations . . . 114
7.2.1 Methodological Limitations . . . 115
7.2.2 Use Case Limitations . . . 116
7.3 Future work . . . 117
7.4 Contributions . . . 119
7.4.1 Scientific . . . 119
7.4.2 Societal. . . . 120
7.5 Ethical reflection . . . 121
7.5.1 When support becomes “bad” . . . 121
7.5.2 Health is an ethical concept . . . 122
7.5.3 Positionality Statement. . . . 123
7.6 Final remarks . . . 124

A Appendix – Acquiring Semantic Knowledge via Human-Agent Alignment Dialogues 125

B Appendix – Conversational Strategies for Agents to Understand Users’ Reasons 133

C Appendix – Expert Insights on Conversational AI Systems as an Information Intermediary 139

Bibliography 143

SIKS Dissertations 173

Acknowledgments 185

Curriculum Vitae 187

List of Publications 189

Summary

Behavior support systems are often deployed to help individuals manage health-related goals, such as increasing physical activity or adhering to dietary guidelines. However, these systems tend to rely on pre-collected behavioral data or predefined user profiles, which makes it difficult to adapt their support to the nuanced, changing circumstances of people's daily lives. This thesis investigates how conversational interaction, specifically through what is termed *alignment dialogue*, can help behavior support systems better understand and adapt to a user's real-time needs, thereby enabling more personalized, user-centered support.

At the heart of this research lies the idea that behavior support is not just about telling users what to do, but also about understanding why certain advice is adhered to or dismissed in particular contexts. Alignment dialogues are introduced as conversations between an AI system and a user, aimed at eliciting reasons for non-adherence, surfacing underlying values, and revealing contextual barriers. Rather than merely logging behavior, these dialogues help uncover the user's lived experience and evolving perspective. Such interactions offer a way to iteratively build a more accurate and complete user model at run time.

The research begins with an exploratory focus group study to understand how people experience different designs of alignment dialogues. The findings show that probing for deeper knowledge like user values often triggers negative emotional responses, including feelings of being judged or guilty. Participants perceived explicit mentions of values as passive-aggressive or annoying, suggesting that the interpersonal impact of an agent's language is a critical design dimension for successful human-agent alignment dialogues.

These findings inform the next study, a large-scale user experiment that tests different types of alignment dialogue strategies to compare their effectiveness in acquiring user reasons for non-adherence. The results highlight a tension between conversational depth and usability: while open-ended exploratory questions allow for greater perceived completeness, they also increase cognitive effort and negatively impact the dialogue experience. Conversely, multiple-choice formats enhance usability but may limit the depth and variety of gathered insights. These outcomes offer important lessons for designing user-centered alignment dialogues in behavior support systems.

To assess how the outputs of such dialogues could be used by others, the thesis shifts focus to a healthcare context. In particular, it explores how alignment dialogue content could be shared with healthcare professionals as a *Conversational Intermediary AI (CIAI)*. Interviews with medical experts suggest that these dialogue agents could offer valuable support in facilitating early diagnoses, easing sensitive discussions, and preparing patients for consultations, though their effectiveness may depend on the type of healthcare provider involved and the specific phase of the patient's journey. A follow-up experiment examines how best to present this dialogue content to third-party professionals, comparing different dialogue presentation formats. Each format has trade-offs: original dialogues preserve

nuance but take time to read, while summarized lists are faster to interpret but risk oversimplification. The interactive visualization tool could provide a middle ground, offering both overview and detail.

To demonstrate the technical feasibility of these concepts, the final part of the thesis introduces a system prototype. This prototype includes modules for alignment dialogue, semantic information extraction, knowledge graph construction, and reasoning. It transforms user responses into structured knowledge using RDF triples and stores them in a user knowledge graph, which is continuously updated throughout the dialogue. This user knowledge graph can be cross-referenced with a domain knowledge graph containing medical and behavioral expertise. The reasoning module utilizes both graphs to recommend actions that are contextually relevant and clinically sound. For instance, if exercise is deemed beneficial, the system can choose between recommending a gym visit or a walk with friends, depending on the user's preferences, schedule, and motivations.

What makes this system stand out is not just its technical integration, but its philosophical shift: it treats alignment not as a static match between user and system, but as a co-constructed, evolving understanding of the user. The approach emphasizes not only whether users adhere to their goal behaviors, but also the reasons behind why they do not, bringing greater transparency and interpretability to the system. This perspective redefines personalization from categorization or prediction to interpretation, where systems must learn to ask the right questions, make sense of the user's situation, and provide support that reflects their evolving goals and constraints.

The findings across these studies show that the effectiveness of alignment is determined not only by the accuracy of the information gathered but by the interpersonal and emotional quality of the interaction. Furthermore, the transition of this information to third parties highlights that comprehension and usefulness are sensitive to the presentation format and length of the dialogue. Ultimately, these results indicate that the contents of a user model are not merely technical requirements but are value-dependent and context-sensitive, requiring a flexible approach to both data collection and presentation.

Altogether, the thesis contributes to both the theory and practice of behavior support systems. It proposes a novel dialogue-based approach to AI alignment through alignment dialogues, empirically tests different dialogue and presentation strategies, and presents a proof-of-concept system that operationalizes the alignment dialogue approach. The research suggests that personalization grounded in dialogue opens new paths for hybrid collaboration between human and machine. Rather than optimizing only for efficiency or adherence, it advocates for behavior support systems that adapt by conversing, reflecting, and aligning with users.

Samenvatting

Behavior support systems worden vaak ingezet om individuen te helpen bij het behalen van gezondheidsdoelen, zoals het verhogen van fysieke activiteit of het volgen van voedingsrichtlijnen. Deze systemen leunen echter vaak op vooraf verzamelde gedragsgegevens of vooraf gedefinieerde gebruikersprofielen, waardoor het lastig is om de ondersteuning aan te passen aan de genuanceerde, veranderlijke omstandigheden van het dagelijks leven. Dit proefschrift onderzoekt hoe conversatie-interactie, specifiek via zogenaamde *alignment dialogue*, behavior support systems kan helpen om de actuele behoeften van een gebruiker beter te begrijpen en zich daaraan aan te passen, om zo meer gepersonaliseerde, gebruikersgerichte ondersteuning mogelijk te maken.

De kern van dit onderzoek is het idee dat behavior support niet alleen gaat over het vertellen wat gebruikers moeten doen, maar ook over het begrijpen waarom bepaald advies in specifieke contexten wordt opgevolgd of genegeerd. Alignment-dialogen worden geïntroduceerd als gesprekken tussen een AI-systeem en een gebruiker, gericht op het achterhalen van redenen voor het niet-naleven van adviezen, het blootleggen van onderliggende waarden en het zichtbaar maken van contextuele barrières. In plaats van alleen gedrag te registreren, geven deze dialogen inzicht in de beleving en het veranderende perspectief van de gebruiker. Dergelijke interacties bieden een manier om tijdens het gebruik iteratief een nauwkeuriger en vollediger gebruikersmodel op te bouwen.

Het onderzoek begint met een verkennende focusgroepstudie om te begrijpen hoe mensen verschillende ontwerpen van alignment-dialogen ervaren. De bevindingen laten zien dat het doorvragen naar diepere kennis, zoals waarden van een gebruiker, vaak negatieve emotionele reacties oproept, waaronder het gevoel veroordeeld te worden of schuldig te zijn. Deelnemers ervoeren het expliciet vermelden van waarden als passief-agressief of irritant, wat suggereert dat de interpersoonlijke impact van het taalgebruik van een agent een cruciale ontwerpfactor is voor een succesvolle afstemming tussen mens en agent.

Deze bevindingen vormen de basis voor de volgende studie, een grootschalig gebruikersexperiment waarin verschillende strategieën voor alignment-dialogen met elkaar worden vergeleken op hun effectiviteit in het verkrijgen van redenen voor niet-naleving. De resultaten wijzen op een spanningsveld tussen gespreksdiepgang en bruikbaarheid: hoewel open, verkennende vragen zorgen voor een grotere waargenomen volledigheid, verhogen ze ook de cognitieve inspanning en hebben ze een negatieve invloed op de gesprekservaring. Meerkeuzevragen daarentegen verbeteren de bruikbaarheid, maar beperken mogelijk de diepgang en variëteit van de verzamelde inzichten. Deze uitkomsten bieden belangrijke lessen voor het ontwerpen van gebruikersgerichte alignment-dialogen in behavior support systems.

Om te beoordelen hoe de uitkomsten van dergelijke dialogen door anderen gebruikt kunnen worden, verschuift het proefschrift de focus naar de zorgpraktijk. In het bijzonder wordt onderzocht hoe de inhoud van een alignment-dialoog kan worden gedeeld met zorgprofessionals als een *Conversational Intermediary AI (CIAI)*. Interviews met medische

experts wijzen erop dat deze gespreksagenten waardevolle ondersteuning kunnen bieden bij het faciliteren van vroege diagnoses, het bespreekbaar maken van gevoelige onderwerpen, en het voorbereiden van patiënten op consulten, hoewel hun effectiviteit afhankelijk kan zijn van het type zorgverlener en de specifieke fase in het traject van de patiënt. In een vervolgexperiment wordt onderzocht hoe deze dialooginformatie het beste kan worden gepresenteerd aan externe professionals, waarbij verschillende presentatievormen worden vergeleken. Elke vorm kent voors en tegens: originele dialogen behouden nuance maar kosten tijd om te lezen, terwijl samenvattende lijsten sneller te begrijpen zijn maar het risico te simplistisch te zijn met zich meebrengen. Een interactieve visualisatietool biedt mogelijk een middenweg door zowel overzicht als detail te kunnen tonen.

Om de technische haalbaarheid van dergelijke systemen aan te tonen, introduceert het laatste deel van het proefschrift een prototype. Dit prototype bevat modules voor alignment-dialoog, semantische informatie-extractie, kennisgrafen en redenering. Het zet gebruikersreacties om in gestructureerde kennis met behulp van RDF-triples en slaat deze op in een gebruikers-kennisgraaf, die tijdens de dialoog voortdurend wordt bijgewerkt. Deze gebruikers-kennisgraaf kan worden gekoppeld aan een domein-kennisgraaf met medische en gedragsexpertise. De redeneermodule gebruikt beide grafen om acties aan te bevelen die zowel contextueel relevant als klinisch onderbouwd zijn. Als bijvoorbeeld lichaamsbeweging wordt aanbevolen, kan het systeem kiezen tussen het aanbevelen van een sportschoolbezoek of een wandeling met vrienden, afhankelijk van de voorkeuren, agenda en motivaties van de gebruiker.

Wat dit systeem bijzonder maakt, is niet alleen de technische integratie maar de filosofische verschuiving: *alignment* wordt niet gezien als een statische match tussen gebruiker en systeem, maar als een gezamenlijk geconstrueerd, zich ontwikkelend begrip van de gebruiker. De benadering legt de nadruk niet alleen op de vraag óf gebruikers hun gedragsdoelen naleven, maar ook op de redenen waarom zij dit niet doen, wat zorgt voor grotere transparantie en interpreteerbaarheid van het systeem. Dit perspectief geeft een nieuwe definitie aan personalisatie: van categorisatie of voorspelling naar interpretatie, waarbij systemen moeten leren de juiste vragen te stellen, de situatie van de gebruiker te begrijpen en ondersteuning te bieden die aansluit bij diens veranderende doelen en omstandigheden.

De bevindingen uit deze studies tonen aan dat de effectiviteit van *alignment* niet alleen wordt bepaald door de nauwkeurigheid van de verzamelde informatie, maar ook door de interpersoonlijke en emotionele kwaliteit van de interactie. Bovendien laat de overdracht van deze informatie naar derden zien dat begrip en bruikbaarheid afhankelijk zijn van de presentatievorm en de lengte van de dialoog. Uiteindelijk wijzen deze resultaten erop dat de inhoud van een gebruikersmodel niet louter technische vereisten zijn, maar waarde-afhankelijk en contextgevoelig, wat een flexibele benadering vereist voor zowel de verzameling van gegevens als de presentatie daarvan.

Al met al draagt dit proefschrift bij aan zowel de theorie als de praktijk van behavior support systems. Het introduceert een nieuwe, op dialoog gebaseerde aanpak voor AI-alignment door middel van alignment-dialogen, onderzoekt empirisch verschillende dialoog- en presentatievormen, en presenteert een proof-of-concept-systeem waarin de alignment-dialoog-aanpak concreet gemaakt is. Het onderzoek laat zien dat personalisatie via dialoog nieuwe mogelijkheden biedt voor hybride samenwerking tussen mens en machine. In plaats van alleen te optimaliseren voor efficiëntie of naleving, pleit het voor behavior support systems die zich aanpassen door te praten, te reflecteren en zich af te stemmen op de gebruiker.

中文摘要

當今社會越來越依賴數位科技，許多人也開始透過各種健康應用程式來幫助自己達成像是多運動、健康飲食等目標。不過，這些系統通常是根據使用者過去的行為或一開始設定的個人資料來提供建議，很難即時理解人在日常生活中所面臨的各種突發情況或心理變化。因此，本論文探討了一種新的互動方式，稱之為 alignment dialogue，希望透過對話的方式，讓智慧系統更了解人在當下的真實狀態，從而提供更貼近個人需求的支持。

這項研究的核心想法是：幫助行為改變不該只是一直告訴他們「該做什麼」，而更應該去了解「為什麼某些建議有時候無法被接受」。Alignment dialogue 是系統和使用者之間的一種對話，其內容主要是為了了解使用者為什麼沒有照目標去做，是什麼價值觀、情緒或生活情境影響了他們的選擇。這樣的對話能讓系統不只是記錄使用者「有沒有做」，而是一步步建立一個對人的更深入、更有彈性的理解。

研究一開始，我透過焦點團體訪談，觀察人們對不同對話風格的反應。有些人覺得開放式問題能讓他們多表達，但也會覺得有點麻煩或壓力大。這些結果幫助我設計了下一個大規模的實驗，比較不同對話策略的效果，例如讓人自由回答、一步步引導或提供選項。結果顯示，有效的 alignment dilaogue 需要在「深入」與「好回答」之間取得平衡。

接下來，我把焦點放在這些 alignment dialogue 的內容能不能也對主要使用者外的另一個利害關係人有幫助，比如生活教練或醫療人員。我訪談了一些醫療專家，他們認為這樣的對話系統有潛力幫助病人在看診前釐清問題、處理敏感議題，甚至可能促進早期診斷。不過，他們也提醒，不同階段的病人、不同角色的醫護人員，可能需要不同的資訊呈現方式。因此我設計了一項實驗，比較把對話原文、摘要清單、或互動式圖表提供給專業人員的效果。結果顯示，各種方式都有利弊，互動工具則有可能兼顧細節與效率。

最後，我與同事們實作了一個 prototype，把整個 alignment dialogue 的流程具體呈現出來。這個系統能把使用者說的話轉換成結構化的資訊，儲存在一個「使用者知識圖譜」中，同時搭配一個內建的醫療與行為知識庫。系統會根據這些資料，推薦符合使用者情境的建議。例如：在已知運動對於使用者是有益的前提下，系統便會根據個人偏好、行程與心情，建議是去健身房，還是和朋友去散步。

這套系統的獨特之處，不只是技術上的創新，更是它在理念上的改變：它不把「理解一個人」當作一開始就能設定好的事情，而是透過一段段對話，和使用者一起

慢慢摸索、調整。它不只是判斷「你有沒有做到」，更關心「你為什麼沒有做到」。這樣一來，系統的建議會更透明、更能讓人理解，也更符合人的真實需要。

整體來說，這篇論文對設計更貼近人心的行為支持系統提出了新的方法與實作成果。它強調，真正有效的個人化支持，不只是靠分析數據或分類人群，更需要透過對話與理解，建立人與智慧系統之間更深層的合作關係。未來的健康科技不只是要幫助人「照做」，而是能真正走進人們的生活與掙扎，陪伴他們一起前進。

1

Introduction

Consider Sarah, a professional balancing a demanding job and a commitment to improving her fitness. She uses an app designed to keep her on track with her exercise goals, and she also consults with a health coach who provides personalized advice during their occasional check-ins. However, her coach is busy and cannot monitor her progress daily or directly understand why she might ignore certain advice from the app.

One afternoon, while Sarah is in the middle of an important meeting, her app sends a notification suggesting that she go for a run. Another day, it recommends a workout when the weather is awful, with heavy rain pouring outside. Sometimes, the app's advice fits perfectly with Sarah's schedule and preferences, but on certain days, even the most reasonable suggestion just doesn't appeal to her mood or energy levels.

This scenario highlights a problem: the support provided by the app does not always align with what Sarah needs or wants in that moment. The advice, while well-intentioned, fails to account for the immediate context or her changing circumstances. As a result, Sarah often dismisses the recommendations.

The root of this issue lies in the information the app uses to derive its support. Such a behavior support app (which we refer to as an agent below) bases its recommendations on the data it has gathered about the user's habits, preferences, and goals. But this data is limited and cannot always fully capture the nuances of the user's daily life or why she might reject the advice in specific situations.

One common approach to addressing this problem is to improve the data itself, using more sophisticated data about behavior or algorithms to refine the app's understanding of the user. However, only relying on behavior data can be problematic. It can be difficult to infer why a user rejects the advice from data alone because observable behavior rarely reveals the underlying reasoning, motivations, or situational constraints, for example whether it was due to external factors like weather or personal reasons such as stress or fatigue. Moreover, from a transparency and trust perspective, users should understand why the agent engages with them in a particular situation, and the agent should, in turn, be able to grasp their reasons for not following through.

To address these limitations, we propose the concept of alignment dialogues: a dialogue in which the agent explicitly interacts about the situation with the user. This approach allows the support agent to directly engage with the user by asking questions to understand why a recommendation was not followed and to gather information that can improve future support. At the same time, it helps the user better understand how the system works and why it responds the way it does.

Consider the example in the beginning: after Sarah dismisses a recommendation to go for a run, the agent initiates an alignment dialogue:

Agent: Could you tell me the reasons you didn't go for your run this afternoon?

Sarah: I was too stressed after work and preferred something lighter instead.

The agent records this feedback, updates the information it stores about the user, and uses this updated information to suggest more appropriate activities in the future. With this interaction, the support agent can provide support that is more personalized, transparent, and responsive to the user's circumstances.

This alignment dialogue not only could improve the app's recommendations for Sarah but the insights gathered from these alignment dialogues can be shared with Sarah's health coach, offering a clearer picture of her preferences, challenges, and patterns over time. In this way the alignment dialogue becomes a bridge, fostering better collaboration between Sarah, the app, and her health coach to achieve more aligned support.

However, how to design and implement such alignment dialogues remains an open question. This leads to the central question of the research:

> *Main RQ: How can behavior support agents use dialogues to acquire user information that could be used to personalize support that aligns with users' needs?*

1.1 Related Concept and Literature

This section discusses the literature on concepts related to the idea of alignment dialogue, particularly within the context of updating user models to better support user behavior. Since the ultimate goal of alignment dialogues is to facilitate behavior support, we begin by examining what supporting behavior entails.

1.1.1 Behavior support technologies

Behavior support technologies are often researched within the broader context of behavior *change* rather than behavior *support* (see e.g., [10, 107, 326]). In the transtheoretical model (TTM) [258, 259], which includes the stages of Pre-contemplation, Contemplation, Preparation, Action, and Maintenance, behavior support, as used in this thesis, refers broadly to assisting individuals not only in enacting change during stages such as Action, but also in maintaining behavior change. In this sense, behavior support is not limited to sustaining change, but includes helping users navigate and implement change as it unfolds. To capture the full process of changing and maintaining behavior, we adopt the term Behavior Change Support Systems (BCSS), as defined by Oinas-Kukkonen [236]: "an information system designed to form, alter or reinforce attitudes, behaviors or an act of complying without using deception, coercion or inducements."

BCSS can take various forms, including SMS messaging, email, telephone, or videoconferencing [326], as well as web or mobile applications [237]. These systems often implement behavior change techniques (BCTs), such as providing information about health consequences, goal setting, self-monitoring of behavior, and action planning [8].

Research in the field of behavior change support typically focuses on three key aspects: (1) the theoretical basis of the intervention, (2) the BCTs employed, and (3) the mode of delivery [326]. Studies generally investigate how these characteristics influence behavior outcomes [138, 272].

Theoretical basis

Theoretical research about behavior change support focuses on developing and refining theories such as:

Social Cognitive Theory Bandura et al. [27]'s Social Cognitive Theory (SCT) is a foundational framework emphasizing the dynamic interplay of personal factors, behavior, and environment. In practice, SCT has been widely applied to health behavior change; for example, a recent review of SCT-based interventions found that virtually all studies targeted self-efficacy and reported positive behavior outcomes (e.g., increased physical activity, healthier diet) [149]. This demonstrates how SCT's core constructs are used to design effective behavior change programs in contemporary research.

Self-Determination Theory Self-Determination Theory (SDT) posits that human motivation and well-being are driven by the fulfillment of three innate needs: competence, autonomy, and relatedness [281]. In terms of empirical support, a meta-analysis of 184 studies in health domains confirmed that SDT principles translate to better behavior change outcomes: interventions that support autonomous motivation and need satisfaction lead to improved health behaviors and outcomes [231]. This broad evidence base solidifies SDT as both a theoretical and applied cornerstone in behavior change research.

Theory of Planned Behavior Theory of Planned Behavior (TPB) is a prominent theory of reasoned action that links beliefs to behavior [13]. TPB explains that an individual's intention to perform a behavior is the central predictor of that behavior, and intentions themselves are shaped by three constructs: attitude toward the behavior, subjective norm, and perceived behavioral control [14]. Decades of research have applied TPB across behaviors and populations; comprehensive meta-analyses of hundreds of studies affirm that TPB's components reliably predict intentions and behavior in health and other domains (see, e.g., [22, 208]). These findings underscore TPB's value in both understanding and intervening on health-related behaviors.

Constructs contributing to behavior change Two examples of psychological constructs that are central to many behavior change theories are attitude and self-efficacy. Attitude refers to an individual's overall evaluation of a behavior, shaped by beliefs about its outcomes. It is a key determinant in models such as the Theory of Planned Behavior, where positive attitudes toward a behavior increase the likelihood of forming the intention to perform it. Foundational work defines attitude as a "psychological tendency expressed by evaluating a particular entity with some degree of favor or disfavor" [103], while empirical meta-analyses confirm that attitudes, though context-sensitive, are moderately predictive of behavior [175]. Self-efficacy, introduced by Bandura [26], refers to an individual's belief in their capability to perform a specific behavior. It plays a central role in Social Cognitive Theory and has been widely adopted across other models due to its strong predictive power. Higher self-efficacy has been consistently linked to greater persistence, effort, and success in behavior change attempts. For example, nutrition intervention studies show that participants with higher self-efficacy are more likely to adopt and maintain healthier dietary habits [21]. Both constructs are frequently targeted in interventions aiming to promote lasting behavioral change.

Behavior change techniques (BCTs)

Research on BCTs examines which specific techniques are most effective for facilitating behavior change.

Systematic reviews linking BCTs to effectiveness One major line of behavior change research focuses on identifying which specific BCTs are most effective in practice. Meta analyses have consistently highlighted self-regulatory techniques such as self-monitoring, goal setting, and feedback as especially impactful. For example, Michie et al. [213] conducted a meta-regression across 122 interventions promoting healthy eating and physical activity and found that interventions combining self-monitoring with goal setting and feedback yielded significantly greater behavioral change than those without. Similarly, Samdal et al. [285] reviewed interventions targeting overweight and obese adults and found that programs using these techniques, especially when paired with motivational interviewing, produced better outcomes. Together, these studies suggest that the effectiveness of behavior change interventions is strongly tied to the presence and quality of particular BCTs.

BCTs clustering and ontologies To support systematic design and evaluation, BCTs have been formally classified into taxonomies and ontologies. The Behavior Change Technique Taxonomy v1 (BCTTv1) by Michie et al. [214] is a foundational framework that identifies 93 distinct BCTs organized into clusters like "Goals and Planning" or "Feedback and Monitoring." This taxonomy has become a widely adopted standard for coding and comparing interventions. Building on this, more recent work has linked BCTs to psychological mechanisms of action. For instance, showing how "instruction on how to perform the behavior" often targets self-efficacy or beliefs about capabilities [73]. These efforts have culminated in the development of the Behavior Change Technique Ontology (BCTO), which expands the taxonomy to over 250 BCTs in a hierarchical structure [200]. Ontologies like the BCTO aim to enable better integration of BCTs with computational systems and consistent annotation of intervention content.

BCTs in digital interventions As behavior change technologies shift toward digital formats, researchers have explored how BCTs are implemented in mobile apps, web platforms, and conversational agents. Reviews of digital health interventions have shown that successful programs often include a greater number of BCTs, especially those targeting self-regulation. For example, a review of technology-based diabetes prevention programs found that effective interventions consistently included goal setting, self-monitoring, feedback on behavior, and social support [319]. In the domain of conversational agents, a scoping review found that chatbots for health and lifestyle change frequently used instruction, emotional support, and problem solving as core BCTs, although few systems were explicitly grounded in behavior change theory [201]. These findings highlight the importance of not only including BCTs in digital tools but also designing them with a clear theoretical foundation to ensure effectiveness and user engagement.

Delivery mode

Research on delivery mode focuses on both the medium through which support is delivered and the manner in which it is communicated.

Platforms for behavior change support Research on delivery modes has explored a range of technological platforms for delivering behavior change support, including web-based systems, mobile applications, embodied conversational agents, and ubiquitous computing technologies. Web-based interventions have long been a focus of behavior change research due to their accessibility and scalability, with meta-analyses showing moderate effectiveness for promoting physical activity, weight loss, and mental health outcomes [326]. Mobile apps have since taken center stage, offering portability, sensor integration, and real-time support; reviews such as Milne-Ives et al. [219]'s systematic review of mobile health apps for behavior change highlight both their promise and variability in user engagement and long-term effectiveness. Embodied agents, i.e., digital avatars with social presence, have been studied for their ability to enhance motivation and relational connection, particularly in interventions requiring long-term adherence [242]. Ubiquitous and context-aware technologies, such as wearables and smart environments, integrate behavior support more seamlessly into daily life, although they introduce new challenges in terms of usability, ethical design, and data handling.

User perspectives on behavior change technologies Another important line of work investigates how users experience, accept, and evaluate BCSS. In a qualitative study, Dennison et al. [87] identified personalization, usability, and perceived usefulness as key factors influencing engagement with smartphone health apps. Later research has reinforced the importance of designing systems that support autonomy, accommodate diverse goals, and adapt over time to individual needs [253]. Emotional connection, perceived intelligence, and low interaction burden have also emerged as critical to sustained user acceptance, especially in systems that deliver long-term support. These findings underscore that the success of behavior change technologies depends as much on how users perceive and interact with them as on their technical functionality.

Foundational models and research agendas The study of BCSS is guided by foundational models that define their goals and structures. Oinas-Kukkonen and Harjumaa [238]'s persuasive systems design (PSD) framework established BCSS as systems intended to form, alter, or reinforce users' attitudes and behaviors, drawing from both system design and behavioral science. This framework has influenced later work outlining ethical guidelines, user modeling, and adaptive intervention strategies. Complementary perspectives, such as Fogg's Behavior Model, highlight the interplay of motivation, ability, and triggers in shaping behavior, offering conceptual guidance for persuasive system design [116]. These foundational models have catalyzed research agendas emphasizing the importance of contextual adaptation, ethical responsibility, and long-term engagement in the development of digital behavior support systems.

Research focuses on developing and evaluating different platforms for BCSS, including web-based applications [159], embodied avatars [161], and ubiquitous technologies [67]. Other areas of research include exploring user perspectives on BCSS [87], systematically reviewing the effectiveness of these systems [219], and proposing foundational studies, research models, and agendas for BCSS [236, 237].

Regardless of the theoretical basis, techniques employed, or delivery modes used, implicitly, all BCSS have some form of information about the user, such as the user's goals or preferences. This information constitutes a type of user model.

1.1.2 User modeling and personalization

User modeling aims to provide personalized interventions by capturing various aspects of users to predict behavior and tailor strategies accordingly. Below, we summarize common approaches to user modeling and personalization.

User models aim to capture and represent key aspects of users. User models can be created using a user-guided approach, where models are created directly from information provided by users [119]. These models often incorporate factors such as personal values and social practices [222], cognitive abilities and working memory [310], as well as personality traits and personal values [205]. Alternatively, user models can be created through an automatic approach, where user information is inferred from interactions with the system [345, 346]. Regardless of the approach, user models rely on validated questionnaires or structured data to quantify abstract constructs and represent relationships between them, often using statistical analysis. By structuring and formalizing user information, these models lay the foundation for personalization.

Personalization builds on user modeling to deliver tailored interventions. A common approach involves categorizing users into profiles based on, e.g., personality [196, 325], gamer types [18], or cognitive styles [265], using tools like questionnaires or physiological data (e.g., eye tracking and physical activity) [173]. These profiles guide strategies or designs deemed suitable for users with similar characteristics.

Dynamic personalization can also be achieved through reinforcement learning, which adapts recommendations to maximize long-term outcomes. For example, reinforcement learning algorithms suggest future activities based on users' recent physical activity data in weight loss interventions to optimize engagement and effectiveness [129].

Another method involves reasoning between user models and knowledge bases to deliver context-specific and tailored advice. This approach combines patient models, which capture user-specific data, with expert-derived knowledge organized into structured frameworks or rules. By applying reasoning mechanisms, these systems dynamically filter and present relevant information, ensuring the advice is both personalized and actionable [68].

1.1.3 Shared mental models

If a behavior change support system (BCSS) consistently maintains a user model that perfectly reflects the model the user has of themselves, it is more likely to provide support that is aligned with the user's needs. This aspiration is closely related to the theory of shared mental models (SMM).

Converse et al. [74] defined SMM as "knowledge structures held by members of a team that enable them to form accurate explanations and expectations for the task, and, in turn, coordinate their actions and adapt their behavior to demands of the task and other team members." The basic idea is that mental model sharedness, defined as "overlapping mental representations by team members that reflect how the group members as a collectivity think or characterize phenomena" [165], improves team performance (see, e.g., [204, 316, 328]). Shared mental model theory has been translated to the context of (human-)agent teams, arguing that sharedness is also important when artificial agents are involved [153].

Since the agent cannot directly inspect the content of the model the user has of themselves, alignment dialogues offer a means to acquire relevant information and update the user model.

However, while SMM provides a framework for cognitive synchronization, researchers in the behavior change space often emphasize the Working Alliance (WA) as a construct of greater relevance for individual change. SMM appears more relevant to teams performing tasks together where both parties have active roles, focusing on and measuring shared knowledge about the team and the task itself. In contrast, WA involves shared goals, agreement on tasks, and a sense of bond specifically aimed at getting one party to change their behavior [41]. Crucially, WA has been found to be a major predictor of adherence to treatment advice, which is of critical importance in behavior change compared to purely task-oriented models [144].

While the theory of SMM emphasizes the benefits of alignment between team members' internal representations, much work has also examined how such sharedness can be achieved. In human-human teams, SMM are often developed through communication, joint training, and coordination practices that encourage mutual understanding of roles, goals, and task strategies [75, 204]. In human-agent teams, research suggests that transparency and explainability are critical for building alignment, as they help users understand the agent's reasoning, capabilities, and goals [60, 151]. Techniques such as shared task representations, model visualization, and interactive plan negotiation have also been proposed to support mental model convergence [45, 190]. These approaches provide insights into how behavior support agents could gradually construct SMM with users, not only by updating internal user models through dialogue, but also by exposing aspects of their own reasoning to foster trust and alignment.

1.1.4 AI alignment

Broadening the scope of alignment from the specific relationship between an agent and a user to the larger context of AI and humanity, we turn to the concept of the AI alignment. This refers to the challenge of ensuring that AI systems align with human values so that it does not pose a threat to humans and society [232, 278–280]. The most simple definition of the AI alignment problem is "AI do what we want them to do" [69]. The authors of IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [292] propose that autonomous systems should "always be subordinate to human judgment and control. If machines engage in human communities as autonomous agents, then those agents will be expected to follow the community's social and moral norms." These researchers advocate that AI agents that are well-aligned with human values and preferences could bring immense benefits to humanity. In these frameworks, the AI alignment problem is discussed from an ethical perspective, focusing on how to ensure AI is properly aligned with societal values. It is assumed that the community norms and social norms are made explicit, or can be learned from human behavior, as Russell [279] proposes.

Below we highlight major technical strategies and socio-technical frameworks for addressing the alignment problem. Technical strategies for AI alignment focus on enabling AI systems to infer and act on human preferences through data-driven or learning-based methods. Researchers have used Inverse Reinforcement Learning (IRL) to explore how AI systems can infer and act upon human values (e.g., [134, 139]). Abbeel and Ng [2] trained agents to recover human objectives by observing expert behavior rather than receiving explicit rewards. This line of work has evolved into Cooperative Inverse Reinforcement Learning [134], which frames alignment as a cooperative game in which the agent must infer human preferences through interaction. Reinforcement learning from human feedback further refines this idea by having agents learn directly from human comparisons over behaviors, allowing the system to align with nuanced and context-specific judgments [70]. These approaches are technically powerful but often limited by challenges in reward specification, data efficiency, and transparency.

Socio-technical approaches to alignment emphasize the importance of involving human stakeholders, not only as data sources but as collaborators in the design and oversight of AI behavior. The Society-in-the-Loop framework proposes embedding societal values through deliberative processes and human oversight mechanisms that mirror the idea of an algorithmic social contract [262]. Participatory design frameworks like WeBuildAI go further by allowing diverse community members to co-create AI policies, ensuring that system behavior reflects the values of affected populations [188]. Together, these approaches highlight that alignment is not just a technical problem but a social one, requiring intentional design to respect human values, norms, and institutional constraints.

Human values At the center of AI alignment – whether in the context of this thesis (between an agent and a user) or the broader societal challenge (between superintelligent AI and humanity) – lies the core concept of human values: what individuals and societies consider important in life, guiding our decisions and actions.

Values are conceptualized as criteria people use to choose what to do and to evaluate others and events [122, 315]. Values are relatively stable in adults and play a central role in motivating behavior [273, 289]. Several theories have been developed to describe and classify human values.

One widely recognized framework is Schwartz's Theory of Basic Values [289]. This theory originally identified ten basic values and was subsequently refined into a model of 19 values arranged in a circular structure [290, 291]. In this model, tensions between opposing values are visualized at opposite ends of the circumplex. Schwartz's values are broad and abstract; Rokeach's Value Survey [273] distinguishes between 18 Terminal Values (desirable end-states of existence, such as *a sense of accomplishment*) and 18 Instrumental Values (preferred modes of behavior that serve as means to achieve terminal values, such as *honesty or ambition*); Moral Foundations Theory [132] emphasizes constructs related to morality; Kahle's List of Values (LOV) [154] classify people on Maslow's hierarchy, and they relate more closely to the values of life's major roles (i.e., marriage, parenting, work, leisure, daily consumption) than do the values in the Rokeach [155].

These theories and their associated instruments not only provide ways to measure values but also help explain differences in attitudes and behaviors at both personal and societal levels [290]. Understanding and incorporating human values is critical for designing AI systems that align with individual and societal expectations, ensuring they act in ways that respect and support what people care about most.

Incorporating values in AI Incorporating human values in AI aligns well with Value Sensitive Design (VSD), a theoretically grounded approach to systematically integrating human values into technology design [122]. VSD advocates for an iterative methodology that combines conceptual, empirical, and technical investigations to ensure that technology accounts for the values of its stakeholders.

A review of two decades of VSD work confirms the importance of these tripartite activities and underscores the need for rigorous technical methods that operationalize abstract values [332]. Recent extensions of VSD into AI contexts have led to Value-Sensitive AI (VSAI) frameworks, which adapt these methods to address ethical concerns and systemic risks posed by data-driven technologies [283]. These include hybrid approaches like value-based engineering, which formalize value elicitation and system design through ISO/IEEE 7000 standards [297]. Together, these approaches provide robust guidance for systematically integrating values into AI systems, from requirements gathering to algorithm design and evaluation.

In addition to design frameworks like VSD and VSAI, argumentation-based approaches have also been proposed to embed human values into AI reasoning processes. For example, Verheij [322]'s notion of value-guided argumentation formalizes how the persuasiveness of an argument depends on the values it promotes or demotes. By integrating values into the logic of argumentation, this approach contributes to designing intelligent systems guided by embedded human values, a step towards more ethical AI in an age of pervasive autonomous systems [322].

Conceptual investigations identify and analyze the values of the direct and indirect stakeholders, addressing trade-offs between competing values, such as autonomy versus security. *Empirical investigations* explore how stakeholders perceive and prioritize values in practical, interactive contexts, investigating discrepancies between stated preferences and actual behaviors. *Technical investigations* assess how existing technological features and mechanisms support or hinder human values. Together, these investigations ensure VSD.

While both VSD and alignment dialogues aim to incorporate human values into technology, they differ in timing and scope. VSD is primarily concerned with embedding values during the design phase and often targets stakeholder groups, shaping the foundational assumptions and architecture of a system. In contrast, alignment dialogues operate during runtime and focus on the evolving needs of individuals. They aim to surface values dynamically, through interaction, enabling personalized and context-sensitive adaptation. Ideally, the design of alignment dialogues should be informed by VSD principles to ensure value-aware behavior from the outset. At the same time, insights gathered through alignment dialogues during deployment can feed back into future design processes, helping to refine which values matter most in lived use.

Value elicitation Value elicitation is a critical step in aligning AI systems with human values. In the context of VSD, this involves "identifying the benefits and harms for each stakeholder group" [122]. Similarly, value elicitation is a key process in software engineering, particularly during the requirements engineering (RE) phase. For example, tools have been developed to embed values in the RE process by collecting value-based user stories [89]. Similarly, chatbots have used laddering techniques to uncover users' values [270]. These values are then translated into norms and design requirements with a hierarchical structure [315], where design requirements represent the most concrete layer of a hierarchy beneath overarching values and norms.

Values have been elicited explicitly through questionnaires featuring predefined value lists, such as Schwartz's framework [290]. However, directly asking individuals about their values can yield incomplete or hypothetical responses that may not reflect their actual behavior in real-life contexts [42]. Additionally, these methods often lack grounding in specific contexts, making it challenging to capture nuanced value preferences [193, 257].

Alignment dialogues may present a promising avenue for eliciting values in a more interactive and personalized manner. Rather than relying on predefined questionnaires or inferring preferences through methods like inverse reinforcement learning, this approach envisions direct dialogue as a means for agents to engage users in meaningful exchanges about their goals, preferences, and values. By grounding the process in real-time, context-sensitive interaction, alignment dialogues could offer a more nuanced and flexible way of capturing what matters to users. However, how to effectively structure such dialogues remains an open research question. This thesis aims to lay the foundation for the realization of what this kind of dialogue could and should look like.

1.1.5 Dialogue systems

Alignment dialogues can be seen as a form of task-oriented dialogue [312], focusing on completing specific operational tasks. While traditional task-oriented dialogues often handle activities such as booking facilities [133], alignment dialogues target the more nuanced task of ensuring that the user model accurately reflects the user's needs, values, and preferences.

In the healthcare domain, conversational agents have been widely explored [182]. They are often used to support self-care or provide information to users, employing approaches that are typically finite state-based (e.g., [255, 303]) or frame-based (e.g., [113, 148]). Finite state-based systems rely on predefined dialogue paths and follow strict, system-directed interactions, where the system directs users by presenting a limited number of choices at each turn [210]. In contrast, frame-based systems use a frame structure, where each slot represents a piece of information the system needs to gather [209]. These systems allow users to provide information in any order, capturing multiple data points efficiently and facilitating task completion [210]. Both approaches are handcrafted and heavily rely on domain knowledge to define states or slots. However, alignment dialogues, as a novel concept, to the best of our knowledge, lack standardized tools for capturing the reasons behind user non-adherence or misalignment.

More recently, large language models (LLMs) have significantly reshaped the landscape of conversational agents. These models enable more open-ended interactions without the need for heavily handcrafted dialogue structures, allowing systems to sustain multi-turn context and produce highly fluent responses [52]. However, while LLMs offer powerful expressive capabilities, they also introduce new alignment challenges, particularly in terms of controllability, transparency, and behavioral consistency. LLMs can be difficult to steer reliably, often behave opaquely due to their black-box nature, and may generate contradictory or misleading outputs across similar prompts, raising concerns about their dependability in sensitive domains [39, 327].

Research around alignment dialogues in this thesis, by contrast, aim to establish a deliberate, structured process through which an AI agent interacts with a user to uncover intentions, values, or contextual barriers. While LLMs may serve as underlying engines to facilitate such dialogues, the core design principles explored in this thesis remain essential: fostering transparency, trust, and adaptability through meaningful interaction. This project was initiated prior to the widespread deployment of LLMs, but its contributions, especially the emphasis on interactive alignment mechanisms, continue to be highly relevant. As AI systems become increasingly general and powerful, designing purposeful alignment dialogues remains a crucial pathway for ensuring these systems behave in ways that are comprehensible and truly responsive to human needs [20, 39].

These developments in language models have expanded the possibilities for conversational agents across many domains, including health. In the context of e-health applications, conversational agents are increasingly used to assist users throughout different stages of behavior change by offering coaching, encouragement, and advice [95, 147, 254, 329]. Existing techniques from behavioral change practices, such as goal setting [38], motivational interviewing [156, 247], and questionnaires [104, 198], have also been integrated into conversational agents.

Alignment dialogues, however, have a distinct focus: rather than directly supporting behavior change, they aim to ensure that the user model accurately reflects the user's needs and preferences in a specific circumstance.

Dialogue alignment It is important to mention that in the area of dialogue systems, a different but related alignment dialogue concept is studied, namely *dialogue alignment*. This concerns alignment processes *in* dialogues, as opposed to the use of dialogues *for* human-machine alignment.

The concept of alignment in dialogue was first proposed by Pickering and Garrod [256] in response to the need for an account of the basic processing mechanisms that are employed during natural dialogues, instead of isolated words, sentences and texts alone. Their fundamental claim was that interlocutors align their linguistic representations in a dialogue, and it is such convergence and alignment that underlies communication success. For example, interlocutors tend to develop the same set of referring expressions to refer to specific objects [49]. Besides the linguistic representation, at nonlinguistic levels interlocutors imitate each other in many respects [46, 128], such as postures and facial expressions. Moreover, Garrod and Pickering [127] identified several sources of alignment, where they discussed alignment via beliefs about one's interlocutor, imitation, agreements between interlocutors, feedback, and physical co-presence.

Dialogue alignment in the conversational agent community thus focuses on how people use linguistic and non-linguistics aspects in dialogues to align with each other and achieve communication success. A dialogue in this sense can be seen as "a joint action in which the participants' goal is to establish aligned representations of what they are talking about" [128]. This has some overlap with our notion of alignment dialogue in the sense that an alignment dialogue is also a joint action with a goal, however, in this setting with the aim to establish aligned representations of user needs. Dialogue alignment processes may be part of an alignment dialogue to achieve successful human-machine alignment and provide the proper support to the user.

1.2 Alignment Dialogues: An Interactive Approach to Alignment in Support Agents

In the context of behavior support agents, effective support relies on the agent's ability to accurately capture and interpret the user's current situation and needs. However, it is often unrealistic to expect the agent to always have complete knowledge or understanding of either. In many cases, it becomes necessary to involve the user in the loop to create or refine the required knowledge. Dialogue is the most natural and intuitive way for humans to interact and share information [256], which suggests it may be a useful approach for addressing these gaps.

1.2.1 Conceptual foundations

Consider the example introduced earlier: the agent would have been unlikely to know that Sarah was stressed and preferred a different form of exercise at that moment without receiving explicit input from her. Dialogue enables such critical information to be surfaced and incorporated into the agent's user model. In this section, we formally introduce the concept of alignment dialogue and outline the main challenges involved in crafting and designing such dialogues, which form the core focus of this thesis.

To begin conceptualizing alignment dialogues, it is useful to reflect on what it means for a support agent to be aligned or misaligned with a user. What constitutes alignment or misalignment between a support agent and a user, and under what circumstances do these situations occur? Understanding when and why these mismatches occur provides the groundwork for defining what alignment dialogue seeks to address.

In this context, it is crucial to distinguish between a user's *behavior goal* and their underlying *purpose*. In this thesis, we frequently refer to concrete activities as behavior goals, such as reducing sugar or attending the gym. However, we recognize that from a behavioral science perspective, these are often *tasks* rather than final objectives. A user may attend the gym to lose weight, build muscle, improve health, or seek social interaction.

Another important consideration is how alignment dialogue unfolds in practice. That is, how might the interaction between user and agent unfold to identify and repair misalignment? What does such a dialogue look like when it is working well? Central to these questions is the notion of *user information*. Defining what information is relevant and necessary is crucial, as it determines the content of the alignment dialogue. Finally, how should an alignment dialogue be crafted and designed? Here, we examine the process of alignment dialogue: what design considerations, conversational structures, and interaction strategies are important to enable effective knowledge exchange between user and agent.

By exploring these conceptual foundations, this thesis aims to lay the groundwork for alignment dialogues that improve the ability of support agents to provide personalized, adaptive, and trustworthy assistance.

1.2.2 Alignment dialogue as an alternative to other approaches

The alignment dialogue agent we envision differs from traditional applications in conversational agent research in several important aspects. Most commonly, task-oriented dialogues focus on completing specific operational tasks, such as booking a facility [133].

While some research has begun to explore higher-level concepts—for example, Abdulrahman [3] and Abdulrahman et al. [5] examine agents that elicit user goals and beliefs to tailor behavior change explanations. Capturing situation-dependent concepts like personal values through runtime interaction remains largely unexplored.

An alternative approach to achieving alignment could be through methods such as inverse reinforcement learning (see, e.g., [134, 139]), which aims to infer human preferences from behavioral data. While these are worthwhile technical projects, data-driven approaches have their limitations. First, the existing data only reflects past or current user behaviors. It does not account for the desired behavior that the support agent aims to provide. Furthermore, data-driven approaches usually lack transparency because of the complex relation between input data and a model's output [92]. This makes it difficult for users to understand how the system works. This lack of transparency can undermine user trust, especially in contexts where personal values and sensitive goals are at stake.

To address these challenges, this thesis proposes an alternative approach: using direct conversation to bridge the gap between the agent's model and the user's evolving needs. We introduce the concept of *alignment dialogue*: a dialogue in which the agent and the user collaboratively aim to achieve or maintain alignment. Through alignment dialogues, users can directly communicate higher-level concepts, such as their values, goals, or desires that are otherwise difficult for the agent to infer from behavioral data alone.

Moreover, alignment dialogues offer additional benefits for transparency and trust. By allowing users to articulate the reasons behind their decisions or preferences, the dialogue implicitly reveals how the agent works and why it offers certain types of support. This can strengthen the user's confidence in the system. In cases of misalignment, engaging the user in dialogue also helps surface the underlying reasons (the *why* behind their behavior) which in turn can enable the agent to adapt more effectively when similar situations arise in the future.

Figure 1.1 provides an overview of how alignment dialogues could be incorporated when employed in a support agent. The process begins with an initial conversation between the support agent and the user, during which the agent asks about user's goal and values or other motivational attitudes. However, the initial user model derived from this conversation is unlikely to fully capture the user's needs and the contextual factors that may shift over time. To address this, alignment dialogues are employed to maintain or restore alignment between the user and the agent throughout the system's use.

1.2.3 Motivations

The need for alignment dialogues arises from the dynamic and ever-changing contexts in which supportive technologies are employed. Misalignment between the agent and the user is inevitable due to the changes in a user's circumstances and needs (discussed further in Chapter 2). In this section, we outline several key reasons for opting for human-agent alignment dialogues, as opposed to relying on purely automatic or data-driven approaches.

Explaining the model to the user requires explicit representation One motivation for using dialogues is to make the agent's internal model understandable and transparent to the user. This is important because users are more likely to trust and engage with a system when they can see how its decisions are made and influence how it represents

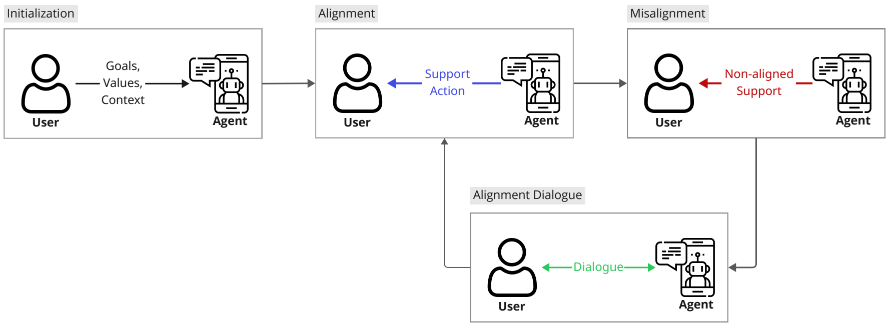


Figure 1.1: The alignment process cycle. Following initialization, the system enters a state of alignment where it provides a support action (blue). If misalignment is detected (non-aligned support, red), an alignment dialogue is triggered to reconcile the agent's model with the user's goals and context, eventually returning the system to an aligned state.

their needs. Data-driven methods, by contrast, often rely on opaque models and implicit inferences, making it difficult for users to understand or intervene in how they are represented. This requirement places certain constraints on the design of the model: it must be explicit, interpretable, and capable of being translated through interaction. A central challenge lies in balancing the expressivity of the model's internal representations with the user's ability to comprehend and interact with them. Dialogue provides a mechanism for surfacing and refining these representations in collaboration with the user, allowing them to directly inspect, update, or challenge the assumptions guiding the agent's behavior. This fosters not only better personalization but also greater user trust and control over how the system operates.

Human-machine interactions differ from human-human interactions It may seem reasonable to design support agents by mimicking how human coaches interact with their clients in similar situations. However, we argue that people may perceive and engage with machines differently from how they interact with humans. While human-human interactions can serve as templates for designing agent dialogue, direct replication may not be effective because machines are perceived as fundamentally different entities. Additionally, translating human-human interactions into machine behaviors is not straightforward [227]. It may be difficult for human coaches to explicitly convey how they tailor their advice to individual clients or patients in diverse circumstances. These differences underscore the need for a distinct approach such as alignment dialogue, that is intentionally designed for human-agent interaction, rather than relying solely on analogies to human-human dialogue. Alignment dialogue offers a structured way to bridge the interaction gap, allowing users to explicitly communicate their reasoning, values, and situational needs to the agent in a way that accommodates the unique dynamics of interacting with a machine.

Machines represent human states differently from humans themselves A related challenge stems from how machines and humans represent user states/mental models differently. A human coach's understanding (mental model) of a client is often shaped by narratives, memory, and lived experience [135]. Machines, on the other hand, typically employ structured models, such as user profiles or knowledge graphs, to represent and reason about the user's needs and preferences [101]. Because of this fundamental difference in representation, behavior data alone often fails to capture the richness of human reasoning and context. This makes it difficult to infer user needs purely from observation, especially in dynamic or nuanced situations. Alignment dialogues are therefore needed not just to bridge the gap between human and machine representations, but to actively involve users in shaping how they are represented. By enabling users to explain their situation in their own terms, dialogue provides a direct, interpretable channel for updating the user model in ways that data-driven methods cannot achieve on their own.

1.3 Methodologies and Research Questions

This section outlines the overall methodological approach of the thesis and presents the research questions it addresses. We describe how each study contributes to understanding the role of alignment dialogue in behavior support systems, and explain how different methods were selected and sequenced. Figure 1.2 provides a visual overview of the research structure and the relationships between the key actors and study components.

1.3.1 Research overview

This thesis considers a three-actor setting, , as illustrated in Figure 1.2, involving: the human user, the human supporter (e.g., a coach or clinician), and the support agent. The support agent functions as an intermediary, interacting with both the user and the human supporter. The figure also highlights the main system components involved and the flow of information between actors. Each numbered box corresponds to a research question addressed in this thesis.

On the left side of the diagram, the focus is on support for the user. The DIALOGUE AGENT engages in conversations with the user to gather information, which is then stored in a USER MODEL (RQ1, RQ2, RQ3).

On the right side, the focus shifts to supporting the human supporter. A different DIALOGUE AGENT is involved in collecting user insights, which are then translated into an appropriate INFORMATION PRESENTATION format (RQ4, RQ5). This component ensures that relevant user information is conveyed to the human supporter in a way that is comprehensible and useful for decision-making. Finally, as a proof of concept, we developed a prototype that implements alignment dialogue in practice, demonstrating how a support agent can acquire user information and use it for personalization (RQ6). Note that RQ6 is not explicitly marked in Figure 1.2, as it refers to a technical implementation of the overall pipeline rather than a specific component.

This figure provides a visual summary of the system's architecture and the two core interaction pathways examined in this thesis: (1) from agent to user, and (2) from agent to human supporter. Our methodological approach follows a sequential logic—beginning with qualitative methods to explore user needs and expert perspectives, and progressing to controlled user experiments to evaluate design alternatives and assess alignment in practice.

These methodologies were selected based on the exploratory nature of the research

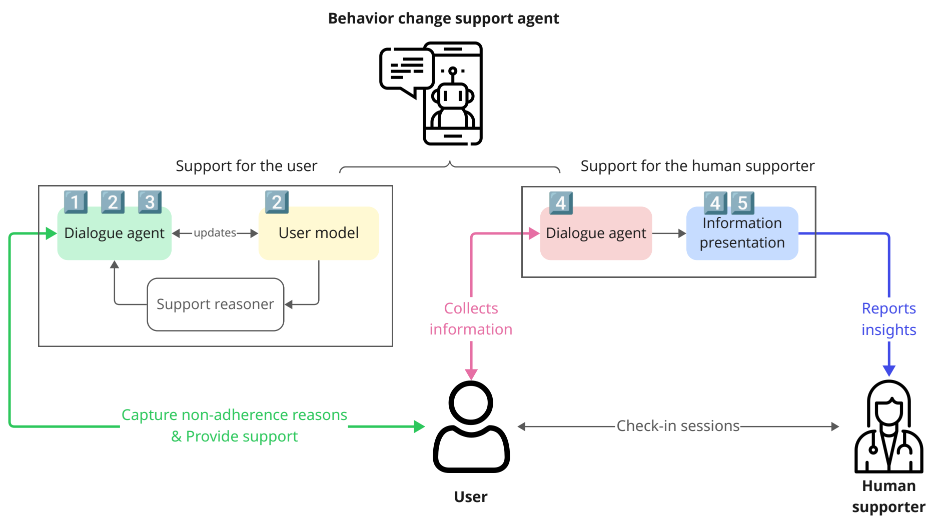


Figure 1.2: Overview of the alignment dialogue system architecture and corresponding research questions. The left side illustrates the components involved in collecting user information and updating the user model through alignment dialogue (RQ1–RQ3). The right side shows how this information is presented to third parties (RQ4–RQ5). RQ6, concerning the technical implementation of the full pipeline, is not depicted but spans the entire architecture.

topic. Because the concept of alignment dialogue is relatively novel in the context of behavior support systems, qualitative methods such as focus groups and expert interviews offer a way to investigate how people interpret, respond to, and make sense of these interactions. Building on insights gained from these qualitative studies, we then conducted controlled user experiments to investigate emerging themes more systematically and evaluate specific aspects of alignment dialogue in practice.

1.3.2 Research questions

To address this overarching question, we break it down into several sub-questions that correspond to different stages of the research.

> *RQ1: Which dimensions are important for designing good alignment dialogues?*

Methodology RQ1 concerns the DIALOGUE AGENT in Figure 1.2: how alignment dialogues should be designed. We began with an exploratory focus group study to identify important dimensions from the perspective of potential users. The study was conducted in the context of behavior support agents that use alignment dialogues to update their user model in different scenarios.

A key methodological choice was to use focus groups (as opposed to interviews) because they allow participants to build on each other's perspectives and to generate richer discussions [162, 178]. Participants read each dialogue as if they were the user in the scenario and discussed which aspects they found helpful or problematic. This included reflections on the agent's depth of reasoning, the timing and initiator of the dialogue, how the agent responded to non-adherence, and how emotions such as guilt and shame arise as a result. The sessions were transcribed and analyzed using inductive thematic analysis [252], incorporating methods from grounded theory to identify relevant themes and dimensions [47, 178, 301] .

Contributions This study served as a first step toward conceptualizing alignment dialogues as an interactive approach to acquiring and updating user information. It sharpened the definition of alignment dialogue and identified design-sensitive dimensions, such as the timing of dialogue initiation and the balance between abstract reasoning and user comfort, that must be carefully considered when crafting such interactions. Together, these insights lay the groundwork for designing alignment dialogues that are not only informative but also emotionally attuned and contextually adaptive.

> *RQ2: What constitutes* user information *in alignment dialogues?*
> *RQ3: How do different strategies for realizing alignment dialogues affect the user's dialogue experience?*

Methodology RQ2 concerns both the DIALOGUE AGENT and the USER MODEL, and RQ3 concerns only the DIALOGUE AGENT, as shown in Figure 1.2. For these two RQs, we conducted a large-scale online user experiment (N = 234). Participants each provided a goal for healthy behavior and identified three situations in which they might fail to adhere to this goal. In this study, we addressed RQ3 by comparing three different dialogue strategies that were designed based on prior literature, and RQ2 through analyzing user responses.

The methodology combined quantitative and qualitative measures. Quantitatively, participants rated self-conscious emotions (e.g., guilt, shame, pride), perceived accuracy and completeness of the dialogue, and dialogue experience. Qualitatively, we analyzed open-text responses using a hybrid of deductive and inductive coding. Deductive coding measured the presence of social, contextual, and value-related information, while inductive coding surfaced emergent topics (e.g., personal challenges, norms, emotional outcomes) not explicitly asked for in the dialogues.

Contributions This study refines the concept of *user information* in alignment dialogues, showing that it encompasses not only values, contextual factors, and social influences, but also emergent elements such as emotional outcomes, personal norms, and behavioral trade-offs. It also demonstrates how different dialogue strategies shape the type and quality of information acquired, as well as the user's emotional response. The findings highlight a key design tension between completeness and cognitive demand, suggesting that hybrid strategies, balancing open-ended flexibility with targeted structure, are well-suited for real-world alignment dialogue systems.

RQ4: What are the concerns and benefits of using an alignment dialogue that collects patient information and presents it to healthcare professionals in diabetes care?

Methodology RQ4 concerns the DIALOGUE AGENT (but on the agent-to-human supporter interaction side) and the INFORMATION PRESENTATION. To explore the practical and ethical implications of using alignment dialogue as a communication bridge between patients and healthcare professionals, we conducted an expert group interview study with five healthcare domain experts. The study centered on a proposed Conversational *Intermediary* AI (CIAI) system designed to gather user information through dialogue and summarize it for healthcare providers in between consultations, particularly in the context of Type 2 diabetes care.

Participants were presented with a storyboard illustrating six scenes that showcased the envisioned functionalities of the CIAI system, ranging from eliciting patients' healthy lifestyle goals, capturing reasons for non-adherence, to summarizing relevant insights for healthcare professionals. Data was collected through a combination of note-taking, post-it annotations, and digital worksheets, then thematically coded and analyzed.

Contributions This study offers an initial assessment of alignment dialogue as an AI-mediated intermediary between patients and healthcare providers in lifestyle-related care. Experts highlighted its potential to streamline consultations by capturing context-rich patient insights and summarizing them for clinical use. These benefits are especially salient in settings where provider time is limited and understanding behavior-related barriers is essential. At the same time, the study surfaces critical concerns regarding privacy, the transparency of data handling, the rigidity of dialogue structures, and the clinical boundaries of AI-generated suggestions. Together, these findings point to both the promise and the design challenges of deploying alignment dialogues in healthcare, underscoring the need for careful consideration of ethical, contextual, and professional norms.

RQ5: Which format is most effective for presenting alignment dialogue content to third parties, ensuring comprehension and accuracy?

Methodology RQ5 focuses on the INFORMATION PRESENTATION. To examine how alignment dialogue content can best be presented to third parties, such as lifestyle coaches, we conducted a controlled user experiment comparing different formats for presenting conversational agent data. These formats represent increasing levels of abstraction: from the raw dialogue, to a categorized list of summaries, to a visual interactive interface. Participants were asked to assess the user's situation based on the presented format and answer comprehension questions inspired by the SAGAT (Situation Awareness Global Assessment Technique) framework [106], and usability of the assigned presentation format.

Real-world behavior support systems often face a bottleneck where users struggle to access or comprehend their own mental state representations or user models. To bypass this challenge, this study deliberately assumed that the dialogue data provided to participants—representing another person's information—was *accurate* and *complete*. By first investigating how third-party observers interpret such data, this research establishes a foundation for future work on how users might directly engage with and resolve conflicts within their own models.

Contributions This study demonstrates how alignment dialogue content can be transformed and presented to *third parties* in ways that balance comprehension, efficiency, and usability. It reveals a fundamental trade-off between contextual richness and processing speed: while original dialogues support the most accurate understanding, structured and visual formats improve speed and usability. The interactive interface emerged as a promising compromise, providing overview and structure while retaining some narrative context. These findings underscore the importance of aligning presentation format with communicative purpose and stakeholder needs, offering guidance for designing alignment-aware interfaces that support both clarity and practical decision-making. Importantly, understanding how alignment dialogue content is best presented, whether to third parties or to the user themselves, can inform how this information is represented within the user model, including what structure or level of granularity supports downstream use. This has implications for dialogue design: knowing how information will need to be interpreted later shapes what kinds of content should be elicited and how it should be framed to ensure clarity, relevance, and interpretability. Moreover, the findings also point to future applications where such information may be presented back to users, raising considerations around transparency, user trust, and privacy. In this sense, the formats studied here provide a conceptual link between information presentation, model representation, and alignment dialogue design, supporting the development of systems that are not only adaptive and expressive, but also user-facing and ethically aware by design.

> *RQ6: How can we technically realize an alignment dialogue pipeline that supports personalized behavior change?*

Methodology To explore how alignment dialogues can be implemented in real-world systems, we developed and demonstrated a technical prototype that integrates dialogue, information extraction, knowledge representation, and reasoning. This system illustrates a full pipeline for acquiring user information through dialogue and transforming it into structured knowledge that can be used for personalized support.

The dialogue component engages the user in a rule-based alignment dialogue, aimed at surfacing high-level information such as reasons for behavior, contextual influences, and personal values. The information provided by the user is automatically extracted into machine-readable RDF triples using an information extraction module.

This structured information is stored in a User Knowledge Graph (User KG), which includes not only lifestyle and contextual information but also health data such as blood sugar or weight. The User KG is supported by an OWL-based ontology and continuously updated through ongoing dialogue. In parallel, a Domain Knowledge Graph (Domain KG) captures general medical knowledge relevant to diabetes care, such as treatment methods and condition-specific guidelines.

To decide on appropriate actions or interventions, a reasoning module compares user-specific data (from the User KG) with medical knowledge (from the Domain KG). It selects interventions that align both with clinical recommendations and the user's preferences, context, and values. For example, if physical activity is recommended, the system may choose between advising a gym session or a walk with family based on the patient's known preferences.

Contributions This integrated architecture enables the system to (1) elicit relevant and personalized user information via alignment dialogue, (2) represent it explicitly and transparently in structured form, and (3) reason over this information to tailor support to the individual. In doing so, it demonstrates how alignment dialogue can serve not only as an interface for acquiring user data but also as a foundation for transparent and adaptive behavior change support.

1.4 Thesis Structure

This thesis investigates how alignment dialogue can be used to acquire information in behavior support systems, and how that information can be shared with human supporters to improve collaborative care. The chapters are structured to follow the conceptual and empirical development of this idea across two main stages: the first focusing on agent-to-user interaction, and the second on agent-to-human supporter communication. The final chapter integrates these findings and reflects on the implications for the design of behavior support systems more broadly.

Chapter 2 explores the foundational concept of alignment dialogue, focusing on how behavior support agents can engage users in meaningful interaction to understand their changing needs. Through a qualitative focus group study, this chapter identifies the key design dimensions that shape the effectiveness and user experience of alignment dialogues (RQ1).

Chapter 3 continues this exploration by examining what types of user information can be meaningfully acquired through such dialogue, and how different dialogue strategies affect the quality and emotional tone of the interaction. A large-scale user experiment investigates how users respond to open-ended, focused, or structured alignment dialogues, offering insight into the trade-offs between information richness and user comfort (RQ2, RQ3).

Chapter 4 shifts the focus to the agent's role as an intermediary between users and human supporters. Based on an expert focus group study in the context of Type 2 diabetes care, it explores how alignment dialogue outputs could be used to support healthcare professionals. It identifies both the perceived benefits and critical concerns related to privacy, naturalness, and clinical relevance (RQ4).

Chapter 5 investigates how to present alignment dialogue content to third parties in a way that supports accurate comprehension and efficient use. A controlled user experiment compares three formats (original dialogue, structured list, and interactive 360° tool), shedding light on the trade-offs between fidelity, usability, and stakeholder needs (RQ5).

Chapter 6 presents a technical proof of concept that brings together the components of an alignment dialogue system. It demonstrates how dialogue-based information can be extracted, represented in knowledge graphs, and used in reasoning to personalize support. The system integrates dialogue management, semantic information extraction, user modeling, and context-sensitive decision-making (RQ6).

Chapter 7 concludes the thesis by synthesizing findings across the six research questions. It reflects on the broader implications for behavior support technology, the ethical and design challenges ahead, and directions for future research.

Throughout this thesis, we position alignment dialogue as a human-centered approach to personalization, complementary to data-driven modeling, and contribute to the broader Hybrid Intelligence research program. While this thesis focuses specifically on alignment within behavior support, related efforts have investigated adjacent questions such as how to foster appropriate trust in AI systems through formal modeling and integrity-based explanation strategies [211], or how to optimize algorithmic support for behavior change through reinforcement learning [17]. Other work has proposed the use of default logic to create adaptable user models that can handle exceptions in behavior patterns [335], or explored how argumentation can address knowledge base inconsistencies arising from shifting user preferences and environmental changes [141]. Together, these contributions aim to build more adaptive, transparent, and collaborative intelligent systems.

2

Acquiring Semantic Knowledge for User Model Updates via Human-Agent Alignment Dialogues: An exploratory focus group study

For personal assistive technologies to effectively support users, they need a user model that records information about the user, such as their goals, values, and context. Knowledge-based techniques can model the relationships between these concepts, enabling the support agent to act in accordance with the user's values. However, user models require updating over time to accommodate changes and continuously align with what the user deems important. In our work, we propose and investigate the use of human-agent alignment dialogues *for establishing whether user model updates are needed and acquiring the necessary information for these updates. In this paper, we perform an exploratory qualitative focus group study in which we investigate participants' opinions about written examples of alignment dialogues, as a foundation for their design. Transcripts were analyzed using thematic analysis. A main theme that emerged concerns the potential impact of agent utterances on the user's feelings about themselves and about the agent.*

2.1 Introduction

Behavior support technology is being developed to perform tasks on our behalf or to guide our actions. In the area of health and well-being, for example, there are support agents that remind us to take our medicine [216], to help us eat healthier [288], and to coach us

Pei-Yu Chen, Myrthe L. Tielman, Dirk KJ Heylen, Catholijn M. Jonker, and M. Birna van Riemsdijk. "Acquiring semantic knowledge for user model updates via human-agent alignment dialogues." In *HHAI 2023: Augmenting Human Intellect*, pp. 93-107, IOS Press, 2023.

on well-being [324]. As support agents become more and more integrated with our daily lives, it becomes even more important that they provide support that is in line with the goals, norms, values, capabilities and context of users [320].

Existing work has proposed computational representations for capturing such human notions as a user model in the agent [12, 79, 164, 309]. These Semantic User Models employ knowledge-based techniques, comparable to a representation of ontologies in a semantic web context, through which the user's motivational attitudes and their relations with user actions are modelled explicitly [32]. For example, a user model can describe the daily activities and sub-activities of the user, the user's capabilities in performing these activities, and which values are promoted or demoted by which activities [309]. If the user's goals and value preferences are also modeled, the agent can select support actions that are in alignment with what the user finds important.

A challenge is how to acquire the information that is to be captured in a Semantic User Model. This is particularly challenging because such a model not only records individual pieces of information about the user but also relations between concepts. Some of this information, e.g., regarding the current habits of a user, may be obtained through analysis of behavioral user data. However, doing so for high-level concepts such as values or goals can be challenging [23]. Moreover, behavioral data reflects people's past behavior rather than their future desired behavior. Capturing the latter is particularly important for agents intended to support a user in changing their behavior. In addition, data-driven approaches can lack transparency because of the complex relationship between input data and a model's output [94]. This makes it difficult for users to understand how the system works and also to adapt the system to their preferences.

In our research, we explore a complementary approach for acquiring user model information, namely via interaction between user and support agent, specifically via a user-agent *dialogue*. The idea is that the support agent will have a conversation with the user where the agent asks the user about the activities they need support with and the underlying values [32]. An initial version of such a user model, however, is unlikely to provide a complete and fully accurate picture of the user's needs and contextual factors throughout the period of use of the agent. There can be situations where the user model needs to be updated, for example, because of changes in user needs or context.

In this paper, we take a first step in designing human-machine *alignment dialogues* (introduced in Section 2.3) that aim to identify and repair such misalignments between user and agent, or prevent future misalignments. We perform an exploratory qualitative focus group user study, in which we show participants different written variants of what such human-agent alignment dialogues might look like and discuss their opinions (Section 2.4). With this first study, we aim to identify dimensions that are important for designing good alignment dialogues. We analyze the focus group transcripts using thematic analysis, through which we identify main themes, concepts, and their relations. Based on this we highlight several considerations that need to be taken into account in the next step in developing alignment dialogue models (Section 2.5). We discuss related work in Section 2.2 and discuss our findings and conclude the paper in Section 2.6.

2.2 Related Work

The concept of human-agent alignment dialogues can be positioned at the intersection of research on conversational agents and human-agent teamwork.

2.2.1 Conversational agents

Conversational agents have already been extensively investigated in the context of healthcare [181], for example, to support users in self-care, retrieving information, or non-task related interactions. These dialogue approaches are typically frame-based, where users are asked to fill in slots in a template, or they take place through a series of pre-determined steps. Elicitation and use of richer Semantic User Models is a novel approach that facilitates more comprehensive and personalized support. This richness means that more nuanced and context-aware information can be integrated into the model, which requires updates as changes occur. Dialogues to facilitate such updates have, to the best of our knowledge, not yet been investigated.

Moreover, in the area of conversational agents, a concept related to alignment dialogues is studied, namely *dialogue alignment* [256]. This concerns alignment processes *in* dialogues, as opposed to the use of dialogues *for* human-machine alignment as we introduce in this paper. For example, interlocutors in a conversation tend to develop the same set of referring expressions to refer to specific objects [49]. Dialogue alignment processes may be part of an alignment dialogue in order to achieve successful human-machine alignment and provide the proper support to the user.

2.2.2 Human-agent teamwork

Furthermore, in our approach, we take inspiration from research in human-agent teamwork, since the user and support agent can be viewed as a team working together to ensure the user is supported appropriately.

From shared mental model theory, we know that mental model sharedness – defined as "overlapping mental representations by team members that reflect how the group members as a collectivity think or characterize phenomena" [165] – improves team performance (see, e.g., [204, 316, 328]). Shared mental model theory has been translated to the context of (human-)agent teams, arguing that sharedness is also important when artificial agents are involved [153]. Sharedness of the mental model that the agent has of the user's goals, and the user's own mental model of their goals, can improve the agent's support and alignment with the user's needs. Since the agent cannot directly inspect the content of the user's mental model, alignment dialogues can be a way to elicit the relevant information and update its user model.

Moreover, the coactive design approach to human-agent teamwork argues for the importance of designing for human-machine interdependence for realizing resilient human-machine systems [150]. Alignment dialogues can be viewed as a way of designing for interdependence in the context of support agents, ensuring that human-machine misalignments can be identified and repaired.

2.3 Human-Agent Alignment Dialogues

In this section, we outline the concept of alignment dialogues. We will start with an illustrating example of how misalignment between users and support agents could happen and what a corresponding alignment dialogue could look like.

Alignment dialogue example:

Scenario: Upon initialization, John has told the agent about his ideal exercise schedule. However, since then, his opinion about this has changed. When the agent asks him to stick to the original schedule, a misalignment situation occurs.

John: I don't like the exercise schedule.

Agent: Are you no longer motivated to exercise?

John: I am, but I just want more variety.

Agent: Okay. Anything in particular that you would like to include?

John: Could you add swimming to my schedule and make suggestions more randomly?

Agent: Yes. Anything else I can do for you?

John: No, thank you.

We identify three types of alignment (Figure 2.1), referring to the desired support as the *purpose* of alignment, the user model as the *means* for aligned support, and behavioral non-compliance as the *trigger* indicating a possible need for starting an alignment dialogue: (1) alignment between *agent's support actions* and the support *users need/want* which is reflected in their actual behavior, (2) alignment between agent's *user model* and user's *self-model*, and (3) alignment between user's *actual behavior* and their *desired behavior*.

Alignment types The first type of alignment is the broadest and encompasses the second and third types. As the agent can never fully grasp the user's true self-model, it cannot be certain about the second and third types of alignments. Therefore, the focus is on the first type of alignment, which pertains to the match between the agent's support actions and the user's needs/wants. We define the user's needs as *the means for them to achieve a specific goal that the agent facilitates and promotes.* A corresponding notion of misalignment can be described conversely as *a situation where the provided support does* not *match what the user needs or wants.*

It is important to note that what the user wants at a specific moment may not be what they need with regard to their goals. Similarly, there might be a conflict between short-term and long-term goals [124]. This then gives rise to what may be seen as a moral challenge of whether the agent should align with what the user wants now, or what they need long term. In this paper, we focus on *how* the user wants to talk about what they need or want in an alignment dialogue, and how to resolve this via an alignment dialogue. The outcome of this dialogue can then be a way for the agent to determine the most appropriate support.

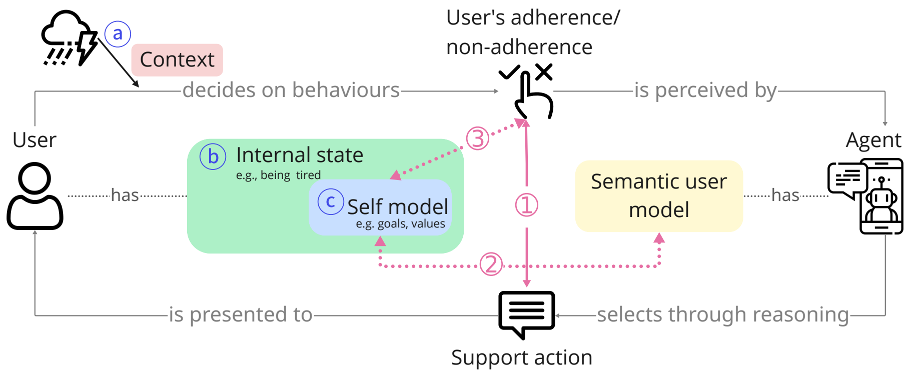


Figure 2.1: Alignment dynamics in user-agent interaction: types of alignment (1,2,3) and changes that could cause misalignment (a,b,c); see text for further explanation. The primary goal of alignment dialogues is to address the first type of alignment, as indicated by the solid pink arrow.

Reasons for misalignment In the example above, misalignment arises because the user changes their mind with regard to their desired exercise schedule. We identify three main reasons why misalignment can arise in general. These are derived from the way the agent's support actions would be chosen, i.e., using reasoning based on information in the user model. A misalignment between the agent's support action and the user's support needs can thus arise because, first, the agent's reasoning process itself can be wrong. Second, the agent's user model can be wrong initially. Third, something can change that requires an adapted interpretation of the situation compared to the information captured in the existing user model. Regarding the last one, we further identify three aspects that could change and cause misalignment (illustrated in Figure 2.1):

(a) **Context**: This includes the external factors or environments of the user, such as the weather, special occasions, or events. When the context changes, the original support may no longer match the user's needs.

(b) **User's internal state**: A user's internal states encompass the user's emotions, stress level, physical or mental conditions, etc.

(c) **User's desired behavior**: As time passes, the user may want to adjust their goals or other motivational attitudes. In this case, the user model may have been correct at the beginning; however, it is the user themselves who changes over time, requiring the agent to adapt the user model to ensure alignment.

Alignment dialogues We define alignment dialogues as *dialogues with which the agent and user try to achieve or maintain alignment.* This can include first establishing if there is a misalignment, as it might not always be obvious to the agent if an observation of the user's behavior points to a misalignment. If there is a misalignment, the conversation could shift to talking about how to solve the situation, where the user and the agent take on a question-answering approach. The cycle will continue until the misalignment no longer exists, and the agent will have obtained a better understanding of the current situation.

2.4 User Study

In the previous section, we have outlined what we mean by (mis)alignment and how we see the role of alignment dialogues. To better understand how we could shape alignment dialogues, we performed a qualitative user study to explore people's opinions and ideas about alignment dialogues.

2.4.1 Focus group with scenarios

The user study was performed in the form of focus groups using a scenario-based approach. We chose to conduct focus groups because the interaction between the participants could spark more discussions regarding what they like or dislike about certain aspects in alignment dialogues [162, 178]. The moderator encouraged participants to express different opinions and ensured participants got a chance to share their views. In the focus groups, the participants were presented with six scenarios with accompanying variants of human-agent dialogues described in textual form, similar to the example in Section 2.3. We followed the definitions and procedures by Schnaars [287] and Spaniol and Rowland [296] to create six scenarios. We identified factors that could lead to misalignment (Section 2.3) and what the corresponding dialogues could be. From a large number of possibilities, we chose six scenarios with differences and diversity to cover various challenges. For each scenario, we had one or two variants of alignment dialogues to address it. Details are further discussed in Subsection 2.4.3.

2.4.2 Participants

Eligible participants were those who were fluent in English and current or potential users of behavior support technology. A total of 13 adults participated in two focus groups, seven in the first group and six in the second group (eight males and five females, age = 26.08 years, SD = 2.72 years) from various countries of origin. The participants were recruited through our networks or through advertisements on social media. We obtained approval to conduct the study from the Human Research Ethics Committee of Delft University of Technology (ID nr 1673). We e-mailed the informed consent forms, including the request for consent to record videos, to the participants before the study and asked them to sign them.

2.4.3 Materials

The study was divided into two parts. The first part consisted of general questions with regard to behavior support agents. The purpose of the first part was to familiarize the participants with this type of agent and its role. For the second part, we focused on participants' opinions on alignment dialogues. Six misalignment scenarios and their corresponding alignment dialogues were shared with the participants in textual form over several pages.

Scenarios In all six misalignment scenarios, the behavior support agent had conversations with a fictional persona named John. John's age, profession, social relationships, hobbies, and the behavior change he needs were detailed in written form, alongside a picture of a white man. These misalignment scenarios occur because the user (John) deviates from their goals due to the unpleasant weather (Scenario 1), the user's mood (Scenario 2), an occasional birthday party (Scenario 3), and changes in desired activities (Scenario 4). Additionally, we included Scenario 5 in which the provided support is in fact in line with user needs but it is so only *by accident* (the agent suggests the user go to work by bike for health reasons but the user does so for an environmental reason). In Scenario 6 the agent is pushy by asking the user to exercise repeatedly even if rejected multiple times, which in itself might be a misalignment as it may deviate from the support the user needs.

Dialogues For each scenario, one or two versions of alignment dialogues were created, with variations in several aspects. One of the key differences between the versions is the **depth of reasoning**: In one dialogue variant, the agent asks surface-level questions, while in another it aims to elicit user values. Values are considered to be a driving factor in human behavior [273, 289]. Friedman et al. [122] defined values as "what a person or group of people consider important in life." For example, in an alignment dialogue, the user may say "because it's raining" when asked why they do not want to go running. If the agent continues asking, the user's values (e.g., *comfort*) may be revealed.

The second variant is the agent's **reactions to the user's non-compliance** with regard to their goals: in one dialogue, the agent acknowledges the importance of the values behind the action, while in the other dialogue, the agent suggests an alternative and asks if this is a one-time exception to know if the user model should be updated. The third variation relates to the **dialogue initiation**, specifically the timing and the initiator. The complete material can be found in the Figure A.1.

2.4.4 Procedure

Due to the measures regarding Covid-19, both sessions were conducted online via Microsoft Teams. All sessions were video recorded for the purpose of making transcriptions. The recordings were deleted once they were transcribed. The participants were given vouchers worth 15 euros as a thank-you for their contributions. We used only reading material in the focus group, with no other physical prompts. The session lasted around 1.5 hours. Therefore, we believe the online setting was appropriate.

At the beginning of the sessions, the overall objective of the study was explained to the whole group: to explore how end-users prefer to discuss misalignment with support agents. The material was shared with the participants while going through each page as per the moderator's instructions. In the first part, each participant was asked about their personal attitudes toward support agents. In the second part, the participants read misalignment scenarios and alignment dialogues, and were then asked to compare and discuss them by imagining themselves as the persona in the scenarios. This continued until all six scenarios were discussed. To guide the participants, discussion questions were prepared.

- Which version of the dialogue do you prefer, or which part of which dialogue do you prefer? Why?
- Is there a certain part of the dialogues that you particularly like/not like? Why? How would you want to do it instead?

Furthermore, we are interested in how users felt about their relationship with the agent after engaging in an alignment dialogue. Although the participants did not interact with a system, we asked them to answer the questions as if they were the user in the presented dialogues. Users' perception of the agent can ultimately affect the agent's effectiveness and resultant user behaviors. Our questions were inspired by the autonomous agent teammate-likeness (AAT) model, which aims to understand humans' perceptions of their intelligent partners. The questions we prepared were inspired by some of the constructs from the AAT model [336], which was a useful inspiration because it aims at understanding humans' perceptions of their intelligent partners. Applying to our use case, the user and the agent need to work together, as a team, to achieve the user's goal.

The AAT model comprises six constructs which are measured with a series of statements, where each statement represents one of the six AAT constructs using 5-point Likert scales [337]. As not all the constructs are relevant or applicable for the goal of our research and our study was qualitative in nature rather than quantitative, we incorporated statements of the AAT constructs into our discussion questions as follows:

- Which dialogue is more "intelligent", as in has more capability in providing support? (related AAT construct: perceived agentic capability)
- Do you feel one dialogue is more supportive than the other? (related AAT construct: perceived benevolent intent)
- After which dialogue do you think the agent would be more "on the same page" as you? (related AAT construct: synchronized mental model)

The validity and reliability of the AAT are not applicable due to the modifications to open questions. However, the primary focus of this study is to gather participants' attitudes and the underlying reasons for their responses, rather than to obtain ratings of the statements. Thus we used AAT only as an inspiration for preparing the discussion and did not rely on its predefined constructs in our data analysis.

2.4.5 Data analysis method

We transcribed the focus group sessions and analyzed the transcriptions using qualitative data analysis methods. Qualitative data analysis is sometimes criticized for being subjective and lacking reproducibility and generalizability [99, 206]. However, when it comes to understanding people's beliefs, attitudes, and values, a qualitative approach may be more appropriate than quantitative methods [99]. As our study aims to uncover the reasons behind individuals' opinions on engagement in alignment dialogues, we believe that a qualitative approach is suitable.

In our study, we seek to understand how users want to engage in alignment dialogues and what are the reasons behind their opinions *grounded* in the data. To achieve this, we chose inductive thematic analysis as our analysis method where the themes identified are strongly linked to the data themselves without trying to fit it into a pre-existing coding frame [250]. This form of thematic analysis is similar to the "lite" version of grounded theory [47]. We incorporated the coding stages from Strauss and Corbin [300] and Krueger [177] as they provide a clear series of steps and descriptions of how each step takes place.

1. Familiarization: at this step, we familiarized ourselves with the data by reading the full transcriptions several times in an effort to immerse in the details and get a sense of the interview as a whole before breaking it into parts [261].

2. Open coding: the data are chunked into small units and coded with a number of words that represent key points in the data.

3. Indexing and charting: the quotes are lifted from their original context and rearranged to prepare for the next step.

4. Axial coding: similar codes are grouped together to create categories from the open codes.

5. Selective coding: central categories that connect all the codes are identified.

6. Interpretation: emergent themes are linked and visualized. The focus is the relationship between the quotes, and the links between the data as a whole [261].

2

The coding results and the model were evaluated through a collaborative process involving the primary researcher and an independent researcher. Both researchers coded the passages independently using the established schema. The independent researcher also reviewed the terminology, consistency, completeness, and grouping of the codes. In instances of disagreement between the two coders, we engaged in further discussion to reach a consensus.

We followed the guidance from Lincoln et al. [191] to ensure the credibility, transferability, dependability, and confirmability of our study. We used data source triangulation (literature, potential users) to validate the credibility of our findings [55, 251]. Transferability and dependability assessments are supported with detailed descriptions of the research methods. Confirmability was ensured through a pilot study and a second coder [207].

2.5 Results

During the study, we asked participants to imagine themselves as if they were the users having an alignment dialogue with a support agent. In the following sections, we use the term *participants* when referring to the opinions of those participating in the study, and *user* when referring to their envisioned role in the human-agent dialogue.

2.5.1 Dialogue variants and how they are perceived differently

Prior to conducting the qualitative analysis, we examined participants' responses regarding the different variants of alignment dialogues (discussed in Subsection 2.4.3) by contrasting what has been said about different variants of dialogues. With regard to the first variation - **the depth of reasoning** - the majority of participants did not prefer the dialogue in which the agent probed further. They found it annoying or missing the point, even if they understood its purpose. In terms of the **reactions to the user's non-compliance**, participants preferred when the agent offered suggestions rather than simply accepting non-compliance. However, participants disliked user model-related conversations, finding them passive-aggressive or sarcastic when values were acknowledged, and weird when asked about the incident being a one-time thing. With respect to the **initiation of the dialogue**, participants preferred the agent to initiate the dialogue but with an option for the user to give input. The timing of initiation varied depending on the situation and severity of the outcome.

It is important to note the aforementioned summary is not intended to provide a conclusive "solution" as to which variant is better. As a result of the qualitative nature of the study, we did not conduct a quantitative analysis. Nevertheless, the observations were intended as an initial exploratory first step to identify which types of considerations and dimensions we need to take into account when designing alignment dialogues.

2.5.2 Tree of codes

We used QSR NVivo [1] to perform the qualitative analysis. First, we derived a preliminary coding schema from a thorough reading of the material (*Step 1* in Subsection 2.4.5). In the second round of analysis, we annotated each piece of text with appropriate codes (*Step 2*), and grouped relevant codes together, resulting in a tree of codes.(*Step 3 & Step 4*).

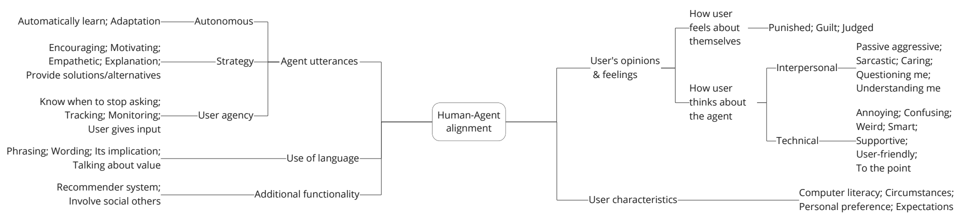


Figure 2.2: Final tree of codes. Due to space constraints, the codes at the lowest level are displayed side by side using semicolons.

The tree of codes is shown in Figure 2.2. At the highest level, there are five categories. These categories are groupings of codes that together represent the main elements emerging in alignment dialogues:

- **Agent utterances** include the codes with regard to the agent's support actions or utterances.
- **Use of language** focuses on how a sentence or a piece of information is expressed.
- **User's opinions & feelings** cover a rather broad theme that consists of the user's feelings or opinions arising in the alignment dialogues, such as the agent being annoying or the user feeling guilty.
- **User characteristics** represent attributes of the user that could play a role in human-agent alignments, such as their personality or personal preference.
- **Additional functionality** refers to functions desired by participants, such as recommendations for activities or integration of menstrual cycles for exercise advice.

2.5.3 Connections between the themes

In the last stage of analysis, we explored the relationships between categories (*Step 5* and *Step 6*). We queried the data with all the combinations of codes. The main intersections were found between *user's feelings* and *agent utterances*, *user's feelings* and *the use of language*, and *user's feelings* and *user characteristics*. We further looked at the quotes that contained these combinations of themes. Some example quotes are presented in Table A.1. By examining the quotes containing these intersections, we gained insights into how the categories are related and the potential impact of alignment dialogues on the user, as shown in Figure 2.3. The conceptual model explains how the different aspects relate to each other, focusing on *why* people have certain views and opinions towards the dialogue content, and *how* it has an impact on the user.

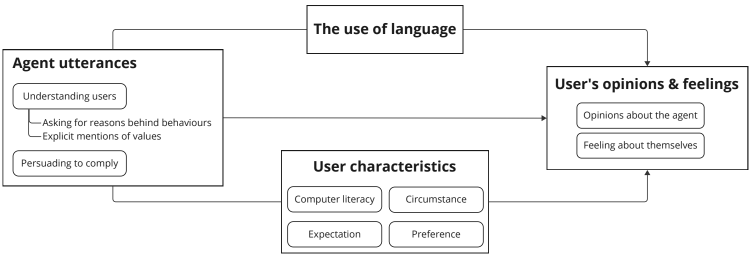


Figure 2.3: Overview of how alignment dialogues may affect user feelings.

Relationship between agent utterances and user's opinions and feelings

The content of alignment dialogues includes interactions aimed at understanding the user and persuading them to comply with their goals, as explained in Section 2.3. Participant responses indicate that these agent utterances could have an impact on the **user's opinions and feelings**, which can be further classified into *user's feelings about themselves* and *how they think about the agent.* It is important to distinguish between these two types of feelings as the objects to which the feelings are directed are essentially different, and separating them helps to gain deeper insights into their underlying causes. For instance, the dialogue may lead the user to feel guilty, or the user may perceive the agent as unsupportive. This differentiation is critical in comprehending the factors driving these feelings and in elucidating their interplay with the alignment dialogue.

Regarding the *user's feelings about themselves* component, it is worth noting that particular events can trigger various emotions. For instance, when the agent uses comparative language to describe the user's decision (see [Q1] in Table A.1), it is likely to generate a feeling of being judged. Similarly, negative emotions arise when the agent highlights the user's non-compliance behaviors [Q2]. When the agent asks the user about their values based on its observations of their behavior, participants indicated they would feel confused, disoriented, or annoyed [Q3].

At first glance, one may question why the alignment dialogues were designed to elicit negative emotions in participants. However, they were not created with such intention in mind, except for Scenario 6. Reflecting on the results, we identify two underlying facets of the nature of behavior support agents that could make them prone to evoking some negative feelings in the user: (1) the agent's role, at least at times, is to address non-compliance behaviors of the user; (2) the agent needs to support the user towards their goals even if it conflicts with their short-term desires.

These findings about alignment dialogues potentially triggering negative emotions in users align with existing research in psychology and behavior change/support. The presence of alternatives that simultaneously cue both long-term goals and short-term desires can lead to conflicts [112]. In such conflict scenarios, either option (i.e. compliance or non-compliance) will inevitably elicit both positive and negative emotions [245]. Additionally, Laschke et al. [183, 184] suggest that friction in interactive technologies is necessary to make people stop, think, and ultimately change.

Factors that moderate human-agent relationship
We observed that there are additional factors that could play a role in the relationship between the dialogue content and the user's opinions or feelings: the use of language and user characteristics.

Regarding **the use of language**, the participants' responses indicate that the way a sentence or a piece of information is expressed could have a significant impact on both the user's feelings about themselves and about the agent [Q1][Q4]. This is in line with Fogg [117], which suggests that the use of language in persuasive technologies sets the stage for outcomes. Jones et al. [152] also demonstrated the effect of source credibility and message framing on promoting physical exercise. It is noteworthy that our findings revealed that users perceived mentions of personal values by the agent as negative [Q5], regardless of whether the intention was to acknowledge the values or verify their accuracy. Users perceive such references as passive-aggressive or sarcastic and express their feelings of being judged.

User characteristics can explain how various aspects of users influence their feelings towards the agent. Although we did not ask the participants for their characteristics, we deduced four user characteristics from their quotes. The first one is the user's *computer literacy.* Participants who lacked computer literacy and did not understand the relevance of certain questions experienced confusion [Q6], while those with more knowledge did not report confusion but did express concerns about the user-friendliness of the agent [Q7].

The second characteristic is the *user's expectation* [Q8][Q9]. Zamora [341] has identified the need for "high performing, smart, seamless and personal" agents. However, the reality does not live up to these expectations. This coincides with Luger and Sellen [195]'s findings on the dissonance between user expectations and their assessment of the intelligence of the conversational agents. Norman's Gulf of Execution [234] illustrated this mismatch between the user's intentions and the allowable actions. The smaller the gulf, the more satisfying the user experience.

The third characteristic is *personal preferences.* Throughout different focus groups discussing different scenarios, it was repeatedly emphasized that personal preference plays a role in human-agent interaction. As one participant expressed, "What if we can choose how we are spoken to," highlighting the importance of personalization. This resonates with a large body of research on personalization. For behavior support, personalization plays an important role as effective strategies are likely to depend on user characteristics [40, 59, 202].

The last user characteristic is the *user circumstance* [Q10]. For instance, the agent gives a reminder when the user is not available, and then the user forgets, which in turn makes them think the agent is not useful. There is vast research on modeling and reasoning about context and situation, e.g., [34, 338]. Behavior support agents need to understand a user's situation to provide comprehensive support [170]. By improving the agent's context awareness, the richness and usefulness of the agent increase as well [7].

2.6 Discussion & Conclusion

Limitations The limited number and similar age of the participants means there are some limitations regarding the generalization of the results. Moreover, written dialogues were used which means that participants did not interact with a dialogue agent personally. Therefore, the results cannot be interpreted as yielding general a theory of how alignment dialogues affect users. Rather, this study is intended as a first step in the design process of alignment dialogues, and our results provide directions for further investigation in our next steps.

Discussion We have observed that the participants' attitudes can be negative regarding parts of the dialogue where the agent tries to ask about abstract, broader reasons behind the user's actions such as their values and what is important to users in general. However, it is not yet clear what the reasons for this are. It could be that participants do not like the parts of the dialogue in which the agents ask about values due to their abstract nature, that the unfamiliarity with this type of conversation causes misunderstandings, that the conversation becomes too deep or personal too quickly, or that the timing is wrong. Moreover, it could be that asking the user for an *explanation* of the reasons behind certain choices of action is perceived by participants as the agent asking for a *justification*. This might be reinforced by the dialogue-based setup, which could invoke the perception of the agent as a *social other* with opinions about the behavior of the user.

This gap in user expectations might be bridged by providing more explicit information about the agent's intent or by tailoring the communication style. As suggested by Ranjbartabar et al. [264] and Salman et al. [284], identifying *how* a user prefers to be spoken to is critical. Incorporating specific relational cues could mitigate negative perceptions and help users understand the relevance of the agent's questions.

Furthermore, following the approach of Abdulrahman et al. [6], the agent could offer potential reasons why a user might resist a certain action and affirm that viewpoint without judgment. By validating the user's perspective before tailoring recommendations and explanations according to their underlying beliefs, the agent may reduce the feeling of being judged and improve the overall acceptance of value-based inquiries.

Contributions and future work The research on alignment dialogues is still in its early stage. This study introduces and sharpens the notion of alignment dialogues and sheds light on what is needed for future development and research on an interactive approach to human-machine alignment in support agents, in particular regarding the potential effects of the dialogues on users' feelings. To further understand these effects, it is essential to conduct qualitative and quantitative user studies where the participants are experiencing the dialogues as they unfold, as opposed to reading pre-written dialogues.

3

Why Don’t You Do What You Said You Would? Conversational Strategies for Agents to Understand Users’ Reasons in Supporting Behavior

Effective support from personal assistive technologies relies on accurate user models that capture user values, preferences, and context. Knowledge-based techniques model these relationships, enabling support agents to align their actions with user values. However, understanding values in a single context is insufficient due to the dynamic nature of behavior. This study explores the use of dialogue strategies to update user models. Participants were randomly assigned to different strategies and they discussed one randomly chosen non-adherence situation with the agent. Then, their emotions, acquired information accuracy, completeness, and dialogue experience were rated. Our findings suggest that multiple-choice dialogues may limit response depth, reducing the perceived completeness of behavior reasons. In contrast, open-ended questions allow more detailed input but require more time and effort, potentially worsening the dialogue experience. Through inductive coding, we identified key topics, such as individual challenges, priorities, tangible outcomes, and values, essential for constructing personalized user models. We also analyzed conversation paths to improve dialogue-based user model updates in support agents. Further research is needed to refine the relationship between dialogue strategies and self-conscious emotions, considering diverse backgrounds and health goals, while enhancing dialogue design.

Pei-Yu Chen, Birna van Riemsdijk, Dirk KJ Heylen, Catholijn M. Jonker, and Myrthe L. Tielman. “Why don’t you do what you said you would? Conversational strategies for agents to understand users’ reasons in supporting behavior.” *Behavior & Information Technology* (2025), pp. 1-20.

3.1 Introduction

Behavior support technology is designed to guide users' actions, for example, by means of support agents that help people eat healthier. Previous research shows that personalized approaches outperform generic "one-size-fits-all" strategies in supporting behaviors (e.g., [176, 197, 243]). Personalization encompasses using various forms of information about an individual, such as socio-demographic characteristics, and personality traits [157, 176], to adjust the support. In this paper, we focus on acquiring information about the reasons for a user's behavior in specific situations in a user model. Acquiring these reasons and the specific details of the situation is critical as it allows for a more accurate representation of the user and thereby provides support that aligns with the user's needs.

In contrast to data-oriented user models that use different learning techniques to predict or classify users [9], the user model we refer to employs knowledge-based techniques [44]. These models explicitly capture the user's reasons and their relationships with desired behaviors. For example, a user model can describe the user's daily activities and which values are promoted or demoted by which activities [309]. If the user's behavior reasons and specific situational details are modeled, the agent can select support actions that align with the user's needs when a similar situation occurs again.

However, there is a research gap in how to acquire the behavior reasons and specific situational details for such a knowledge-based user model. Given the dynamic nature of user context and behaviors, we argue that interactive dialogues are promising because they facilitate the exploration of evolving user preferences and contextual nuances [71]. Additionally, conversational agents offer benefits such as instant availability, a gentle learning curve, and platform independence [166]. In the same vein, Chen et al. [62] proposed the concept of *human-agent alignment dialogues*, which are defined as dialogues with which the agent and user try to achieve or maintain alignment - support that aligns with the user's values and preferences. Through these dialogues, the agent attempts to acquire what is significant to the user in their ever-changing daily contexts. As these contexts continually evolve, misalignments between the support provided by the agent and the user's needs may occur. Alignment dialogues are proposed to address these misalignments, ensuring that the support remains relevant and effective.

Alignment dialogues include the user and the agent engaging in a question-answering conversation. The cycle continues until the misalignment no longer exists, allowing the agent to gain a better understanding of the current situation. For example, the agent might ask, "What factors influenced your decision to skip your workout today?" or "How does the current situation affect your motivation to eat healthily?" Different choices can be made about how these questions are phrased, ranging from open-ended to multiple-choice formats. The way these questions are structured can significantly impact the quality of the information gathered and the user's emotional response.

In this study, we explore different dialogue strategies for alignment dialogues within the context of healthy lifestyle change, but the concept of alignment dialogues can be beneficial in other behavior support domains too. These strategies vary in terms of the openness or closedness of the questions (detailed in Subsection 3.3.2). We focus on scenarios where the agent's advice, although in line with the user's goals as understood by the agent, is perceived as incorrect by the user. In such instances, the agent must capture the reasons behind the misalignment to identify if or how to revise the user model. This study explores how to structure a series of questions to acquire these reasons effectively.

One of the key behavior-related aspects we aim to capture is users' values. This is based on the premise that effective personalization of technological support requires an understanding of what is important to users [320]. Values are particularly useful in this regard, as they represent the criteria people use to make decisions and evaluate others and events [122, 317]. Additionally, we are interested in how individuals prioritize different values across various situations. Therefore, we also consider contextual information about the situation and the social context to be useful.

In addition to eliciting useful information effectively, alignment dialogues should minimize the arousal of negative emotions. As alignment dialogues often occur in situations where users fail to perform goal behaviors, negative emotions may arise [112, 245]. According to Tangney et al. [306], self-conscious emotions are evoked by self-reflection and self-evaluation; these could be particularly relevant when the agent probes into reasons behind users' non-adherence with goal behaviors [62].

In this study, we investigate the effectiveness of dialogues in eliciting the required information and which self-conscious emotions arise after these dialogues. Participants were randomly assigned to different dialogue strategies, discussed one randomly chosen non-adherence situation with the agent, and then rated their emotions, information accuracy, completeness, and dialogue experience. This work contributes to a deeper understanding of specific situational factors, in contrast to the general categories of barriers typically discussed in behavior change literature, that prevent users from doing their goal behaviors. Specifically, we formulate the following research questions:

- RQ1: What are the effects of different dialogue strategies on the occurrence of self-conscious emotions?

- RQ2: What are the effects of different dialogue strategies on the dialogue experience, and on the subjective accuracy and completeness of the information acquired?

- RQ3: How effective are different dialogue strategies in eliciting the targeted information (social and contextual aspects of the situation, and underlying values)?

Additionally, we formulate two exploratory research questions:

- RQ4: Besides the targeted information, what additional topics do users mention?

- RQ5: What patterns emerge in users' transitions between various topics when explaining their reasons for non-adherence behaviors?

The remainder of the paper begins with a discussion of Related Work (Section 3.2). Section 3.3 describes the design rationale of the dialogue strategies. We carried out a user experiment to investigate different dialogue strategies (Section 3.4). Subsequently, the results and discussion are presented in Section 3.5 and Section 3.6. Insights derived from this study and their implications for the design of alignment dialogues are discussed (Section 3.7). We discuss future work (Section 3.8) and conclude the paper in Section 3.9.

3.2 Related work

In this section, we review existing literature relevant to our study, namely alignment dialogues (Subsection 3.2.1), and emotions that can arise in such interactions (Subsection 3.2.2). We clarify the unique perspectives of alignment dialogues, the possible topics these dialogues encompass, and the resulting potential impacts on the users.

3.2.1 Human-Agent Alignment Dialogues and Knowledge-based User Modeling

Conversational agents have been developed to support a wide range of health-related activities, including behavior change. These conversational agents largely focus on the actual support functionalities such as instruction on how to perform a behavior, social support, and problem-solving [201]. These conversational agents were evaluated based on the effectiveness of the actual change of behaviors and the usability of the agents [218]. However, in this study, the primary focus of the conversational agent is to gather information about the user by obtaining the reasons behind their non-adherence behaviors through *human-agent alignment dialogues*. This process aims to construct a more accurate and comprehensive user model. Such dialogues can be viewed as a necessary step before the agent provides actual support functionalities.

Defining Alignment Dialogues and Their Unique Role

Human-agent alignment dialogues, as defined by Chen et al. [62], are dialogues where the agent and user try to achieve or maintain alignment, which is defined as a situation where the provided support matches what the user needs or wants. This could be accomplished by the agent acquiring characteristics of users through the exchange in the dialogue. These dialogues are a form of task-oriented dialogue, such as in Traum and Hinkelman [312], but are different from traditional applications in conversational agent research in several aspects. First, common task-oriented dialogues focus on completing specific operational tasks, such as booking facilities [133]. However, capturing high-level, situation-dependent concepts like values via a conversational agent, to the best of our knowledge, has not been extensively explored.

Meanwhile, in recent years there is a growing amount of research on conversational agents in the e-health domain. These agents typically focus on offering support through coaching, encouragement, and advice, facilitating various stages of behavior change [95, 147, 254, 329]. In contrast, alignment dialogues have a different focus: rather than supporting the users in their health, alignment dialogues focus on ensuring the user model is correct. The underlying motivation is that only when the user model is a correct representation of the user, can the agent provide support that aligns with the user's goals, norms, values, capabilities, and context.

To formally represent the connections between values, actions, and their relationships, Pasotti et al. [248] have introduced a framework that models activities in hierarchies. Building upon this framework, Tielman et al. [309] have demonstrated how values and contextual factors can be integrated. Values are linked to actions within a hierarchical structure. The model also specifies whether values are promoted or demoted by actions, and how a context influences the relationships.

Value Elicitation

A primary focus of alignment dialogues is to elicit users' values. Values are concepts that people use to evaluate people and events and make decisions [121, 317]. Capturing these values enables technology to offer personalized support [320]. Explicit elicitation of values, as demonstrated by Berka et al. [32], is necessary because inferring such high-level concepts from behavioral data is challenging [23], especially considering that each of us has different values and that humans can behave irrationally [53]. Moreover, values are abstract concepts that do not always carry the same meaning to all people. Therefore, incorporating values into a system should involve direct engagement with potential users during the design phase [315].

Traditionally, values can be acquired explicitly through various questionnaires with predefined value lists, such as Schwartz [290]. However, directly asking individuals about their values can lead to incomplete or hypothetical responses that may not accurately represent real-life behaviors [42]. Moreover, these methods may acquire value preferences but are often not grounded in a context [193, 257]. Liscio et al. [192] proposed a methodology to identify context-specific values by engaging human annotators in NLP tasks. While their work focuses on identifying values within a specific context, the term *context* is used more broadly compared to how we consider the specific situation a user is in when making a behavior-related decision.

In contrast, laddering interviews have been proven effective in eliciting deeper insights [277, 321]. These interviews involve a series of *why* questions to construct an attribute-consequence-value (ACV) chain [215]. Laddering interviews are rooted in personality psychology and are applied to explore user needs and values [33, 277, 321]. While traditionally used in requirements engineering for understanding consumer objectives, we conjecture that this method could be suitable for our context to guide users to describe the characteristics of the situation (**attribute**) and its impact (**consequence**), ultimately leading to a better understanding of their **values**. Inspired by the laddering interview technique, our first dialogue strategy employs a series of *why* questions.

Context and Social Awareness

While it is crucial to capture users' values, we should also consider how individuals apply different values in varying situations. The prioritization of values is heavily influenced by socio-cultural environments [93] and contexts [140, 193]. Therefore, besides capturing a user's values, an agent should also identify the *contextual features* that influence behavioral decisions, such as non-adherence with the user's stated goals. This aligns with the concept of context-awareness in system personalization and tailoring.

Research in computer science uses terms such as situation-awareness (e.g., [105]) and context-awareness (e.g., [15]) to describe approaches for enabling artificial agents to better understand their surrounding environment [169]. Op den Akker et al. [241] provides a comprehensive survey of various activity coaching systems that incorporate different forms of context-aware tailoring. Op den Akker et al. [240] identify a set of optimal contextual features that enhance feedback compliance in physical activity systems.

A particular aspect of context-awareness is *social situation awareness*. This aspect specifically focuses on the user's social context. Kola et al. [171] defined a social situation as one involving multiple individuals. Social situation awareness emphasizes factors like relationship quality and contact frequency between the user and others in the situation [168], and demonstrates how integrating social features into a decision tree can offer personalized and socially-aware behavioral support.

The contextual and social aspects of a situation inform our second and third dialogue strategies, in which we ask questions about these aspects, as well as the user's values.

3.2.2 Self-conscious emotions

When it comes to behavior change, a critical issue is user adherence with their behavior goals. Research has demonstrated that when individuals are presented with choices that trigger both long-term goals and immediate desires, it often leads to internal conflicts [112]. In such scenarios, choosing either option—adherence or non-adherence—can evoke both positive and negative emotions [245].

Chen et al. [62] highlight two aspects of behavior support agents that may inadvertently provoke negative emotions in users: (1) The agent sometimes needs to confront the user's non-adherence behaviors, and (2) The agent's role in guiding the user towards their long-term goals, which may conflict with their immediate desires. The emotions commonly experienced by users in these situations include feelings of being judged and feelings of guilt, which are classified as self-conscious emotions [311].

Self-conscious emotions differ from basic emotions because they are evoked by self-reflection and self-evaluation [306]. Tracy and Robins [311] proposed a process model explaining the emergence of self-conscious emotions, suggesting that individuals assess whether an event aligns with their goals and self-identity, which in turn influences the emotional outcome.

If the support provided by the agent is deemed appropriate from the agent's perspective (i.e., aligns with the user's long-term goals), but the user finds it unsuitable, it indicates a misalignment for the user. This incongruence could lead to negative self-conscious emotions. As a result, we consider self-conscious emotions to be a crucial measure in our study.

3.3 Dialogue Design

In this study, we have chosen health-related behavior changes, such as increasing physical activity or adopting a more nutritious diet, as our use case for studying different alignment dialogue strategies. In this section, we give an example scenario in this use case (Subsection 3.3.1) and describe the dialogues strategies used in the experiment (Subsection 3.3.2).

3.3.1 Use Case Scenario

Example Scenario 1 During Anna's interaction with her behavior support agent, she shares her goal of reducing sugar intake. The agent adjusts her online grocery options accordingly. However, Anna deviates from her routine and adds sugary items to her cart. This scenario highlights the need for human-agent alignment dialogues as the agent's support action may not always align with the user due to incomplete information. For instance, Anna may have bought sweets for her birthday celebration, which the agent's model doesn't account for.

Example Scenario 2 Another user John, whose goal is to regularly attend the gym, skips several gym sessions. This scenario could initiate the agent to start an alignment dialogues to acquire John's reasons. For example, John might have skipped his workouts because he's facing a tight deadline at work, which takes precedence as it could impact his chances for a promotion.

The alignment dialogue aims to uncover such context, enabling the agent to update Anna's or John's user model and provide more relevant support.

3.3.2 Dialogue Design Choices

Following the identification of non-adherence scenarios, the agent engages in dialogue with the user to explore the underlying factors contributing to the deviation from their goals. Three dialogue strategies were used in the experiment, namely Exploratory Dialogues, Focused Dialogues, and Structured Choice Dialogues.

Exploratory Dialogues

Intuitively, to capture why the initial support was inadequate, the agent needs to probe into *why* the support was wrong and elicit knowledge about what aspects of the current situation are significant to the user. In this dialogue variant, the agent aims to "dig deeper" by posing a series of why-questions, aiming to eventually uncover the user's values behind their decision in certain situations.

Exploratory dialogues involve the following questions. This setup was inspired by Rietz and Maedche [271]. Responses from the participants to these queries are in free-text form. A representative example of this interaction format, drawn from our empirical data, is presented in Figure B.1.

1. **Q1:** What particular factors related to the situation influence your choice?
2. **Q2:** Why is it important to you?
3. **Q3:** What feeling does it give you?
4. **Q4:** Why is that?

Focused Dialogues

Focused Dialogues employ a dialogue strategy that involves using open questions but with a targeted information-seeking approach. This selection of targeted information was based on our assumptions, drawn from existing literature about the aspects influencing user behavior, namely: social aspects [171], contextual aspects [241], and personal values [290, 320].

The agent initiates the dialogue with the user by outlining that it will explore three aspects of the scenario, accompanied by the definition from [171], [241], and [122] for social aspects, contextual aspects, and personal values respectively. Responses from the participants to these queries are in free-text form. This intentional structure progresses from tangible details to broader considerations, encouraging introspection. An authentic instance of this dialogue is shown in Figure B.2.

1. Social aspects: These encompass the elements within a situation that are related to social interactions, relationships, and social dynamics.

Q1: What social aspects in this situation contribute to your choice?

2. Contextual aspects: These can be any information within this situation that can influence, shape, or affect your choice.

   Q2: What contextual aspects in this situation contribute to your choice?

3. Personal values: Personal values are what a person considers important in life.

   Q3: What personal values contribute to your choice in this situation?

Structured Choice Dialogues

In addition to two open-ended strategies, we employed a structured approach by presenting participants with lists of predefined social aspects, contextual aspects, and values as multiple-choice. The questions in Structured Choice are identical to those posed in the Focused Dialogue. The difference lies in Structured Choice Dialogues offer participants predetermined options to select from. This method provides participants with predefined options, allowing them to choose the elements most relevant to their decision. Below we describe the lists used. The full list with all the options is shown in Section B.1. An empirical instance of this dialogue, reflecting an authentic interaction with a participant, is presented in Figure B.3.

1. **Social aspects:** This list was adapted from Kola et al. [168], who formalize elements of social situations, including the role a person plays, how often they are in touch with others, the nature of relationships, depth of acquaintance, quality of the relationship, and how formal or informal they are with those around them.

2. **Contextual aspects:** For contextual aspect, we used the lists from Op den Akker et al. [240] and Op den Akker et al. [241] which capture elements such as the weather conditions, the proximity to places where goals can be pursued, timing-related details (like day of the week, specific occasions), and past behaviors concerning the goal.

3. **Personal values:** We used the List of Values (LOV) for presenting values [154]. These values from the LOV are known to be related to consumer activities and healthy lifestyle choices [96, 143]. The LOV includes the importance of interpersonal relations, personal factors (e.g., self-respect, self-fulfillment), and personal enjoyment (e.g., fun, excitement) in value fulfillment.

3.3.3 Hypotheses

Based on the reviewed literature and the dialogue strategy design, we articulated the following hypotheses, corresponding to RQ1 to RQ3.

- H1: Participants in Exploratory Dialogues will experience greater self-conscious emotions compared to Focused and Structured Choice Dialogues, as open-ended questions may be perceived as more judgmental [62].

- H2a: Exploratory Dialogues will result in a lower dialogue experience than Focused Dialogues or Structured Choice Dialogues. Open-ended questions may require more time and effort to answer [30, 66], potentially inducing cognitive burden and leading to skipped or low-quality responses [30, 66, 266].
- H2b: Structured Choice Dialogues will lead to lower subjective accuracy and completeness compared to Exploratory and Focused Dialogues. This is because the closed nature of Structured Choice could limit users' ability to express themselves in a manner that feels accurate to them [66], and that users may find the use of values unrelatable or unnatural [42].
- H3: Structured Choice Dialogues will be the most effective in eliciting users' values, social, and contextual aspects, followed by Focused Dialogues, with Exploratory Dialogues being the least effective.

3.4 Method

A user experiment was conducted to investigate how effective different dialogue strategies are in eliciting information related to contextual and social aspects of a situation, and underlying values (RQ1). In addition, an inductive coding method was used to see if additional topics came up in the Exploratory Dialogues and Focused Dialogues (RQ2). Validated questionnaires were used to measure self-conscious emotions and dialogue experience (RQ3). Finally, participants were asked to rate the accuracy and completeness of the summary of their provided input (RQ4).

3.4.1 Participants

Participants were recruited from the crowdsourcing platform Prolific[1], where they received monetary compensation according to platform policies. Approval for the study was obtained from Human Research Ethics Committee of Delft University of Technology (ID nr 2795). A total of 234 participants were included in the analysis. 82.91% of individuals were below 35 years old, 14.10% were between 35 and 54 years old, and 2.99% were 55 years old or older. Educational backgrounds varied, with participants holding different levels of education (high school 30.34%, Bachelor 55.98%, Master 12.39%, PhD 0.85%). Additionally, 46.15% reported degrees or professional experience in STEM fields, while computer programming skills varied (no experience 32.48%, novices 37.18%, competent 27.77%, experts 2.56%).

3.4.2 Procedure

We used two platforms for the experiment: Qualtrics[2] for the questionnaires and Landbot[3] for testing the dialogues. Landbot is a rule-based platform that allows us to prototype and test different dialogue strategies without having to build a conversational agent from scratch. It follows the dialogue flows we created, as detailed in Subsection 3.3.2. The experiment had two main parts:

[1]https://www.prolific.co/
[2]https://www.qualtrics.com/
[3]https://landbot.io/

Participants first completed a Qualtrics survey via Prolific, providing background information, computer skills, and emotion traits (see Subsection 3.4.3). Subsequently, they engaged with a conversational agent. During the chat, they shared one health-related goal and three *non-adherence situations* where they thought it might be difficult to stick to their goals. They also rated their motivation level for the goal and the perceived realism of the non-adherence situations.

Participants were randomly assigned to different dialogue strategy conditions. While they initially provided three non-adherence situations, only one was discussed with the agent during the second part, chosen at random to simulate real-life scenarios. The questions they were posed during the human-agent dialogue in Part 2 were based on the assigned dialogue strategy. Following the dialogue, participants returned to Qualtrics to rate their self-conscious emotions, accuracy and completeness of the acquired information, and their dialogue experience. We detail these measures in the next section.

3.4.3 Measures

In this section, we detail the instruments employed to evaluate the study. These measures include control variables, self-conscious emotions measures, the Dialogue Experience Questionnaire (DEQ), and assessments of accuracy and completeness.

Control variables

Demographics Participants provided essential demographic information, including age, gender, and highest level of education. They also rated their technological proficiency on a 4-point scale from "No experience" to "Expert," and were asked about their academic or professional experience in STEM-related disciplines.

Self-conscious emotion proneness To establish a comparative baseline for emotional responses, we assessed participants' inclination towards self-conscious emotions using the Test of Self-Conscious Affect-2 (TOSCA-2) [305]. Participants rated their likelihood of reacting to hypothetical scenarios on a 5-point scale, focusing on Shame, Guilt, Authentic Pride, and Hubristic Pride-Proneness. This measurement helped compare proneness with actual emotional responses post-interaction.

Motivated & Realistic Assessment Additionally, after each participant shared their individual healthy lifestyle goal and identified three challenging situations that could hinder adherence, the conversational agent prompted them to evaluate, on a scale ranging from 1 (not at all) to 5 (extremely), both their level of motivation to accomplish their goals and the perceived realism of encountering the obstructive situations.

Self-Conscious Emotions Measures

Given our focus on behavior change for a healthy lifestyle, we employed the Body-related Self-Conscious Emotions Fitness instrument (BSE-FIT) [58]. Participants rated 16 statements on a 5-point scale from "strongly disagree" to "strongly agree." Adaptations were made to inquire about emotions triggered by the conversational agent, e.g., changing words to assess emotion "at the moment" instead of general.

Dialogue Experience Questionnaire (DEQ)
We assessed users' alignment dialogue experiences using the Dialogue Experience Questionnaire (DEQ) [307]. We mainly focused on the "Interaction: discussion satisfaction" section. In addition, we included two items from the "Interaction: reality" section: *The conversation with the agent felt natural* and *I had to adjust my way of interacting to communicate with the agent.* Participants answered on a 7-point Likert scale, ranging from "Strongly disagree" to "Strongly agree."

Accuracy and Completeness
To assess how well the dialogue acquired the reasons for non-adherence, the agent compiled a recap at the end of each interaction, template-filling the information provided by the participants. Participants then rated on a scale from 1 (not at all) to 5 (extremely) the accuracy and completeness of this recap in acquiring their reasons for non-adherence.

3.4.4 Text Responses Analysis Method
Exploratory Dialogues and Focused Dialogues yielded free-text responses from participants, which we analyzed using a combination of deductive and inductive analysis as demonstrated by Fereday and Muir-Cochrane [111].

Deductive coding
The deductive coding approach involves predefined coding schemes developed by previous research or theory to categorize data [77]. However, existing behavior change research primarily focuses on facilitators and barriers related to individuals' characteristics, views, or beliefs [88, 302], often overlooking the reasons behind behaviors in specific contexts, such as the non-adherence scenarios central to our study. Consequently, we developed our own coding scheme. First and foremost, we examined whether participants introduced *novel information* in their answers −defined as any new details or insights that the user has not previously mentioned at any point during the dialogue. This definition does not limit the novelty to the immediately preceding response; rather, it encompasses all parts of the dialogue, regardless of whether it pertains to social, contextual, or personal value aspects. These three aspects were only annotated if an utterance was first coded as novel information. Our deductive coding scheme included four categories: novel information, social aspects, contextual aspects, and personal values. We annotated the utterances with these predefined codes as either being mentioned or not.

Inductive coding
If a response coded as having novel information did not fit under the codes of social aspects, contextual aspects, or personal values, we used inductive coding. In this process, we analyzed the data to identify and establish new codes based on key points within the responses. We followed the inductive coding stages from Thomas [308].

Coding process
In summary, our text analysis method integrates both deductive and inductive coding approaches. We follow the following coding procedure. Additionally, an independent coder evaluated one third of the data. The independent coder was given the definitions of

novel information, social information, contextual information, and personal values. The independent coder was also given examples of utterances that should be coded as such. In instances where coders disagreed on the coding of an utterance, they engaged in further discussion to arrive at a consensus.

1. Each utterance is assessed to determine if it contains novel information in comparison to what participants had said.
2. If no novel information is identified, the coding process concludes for that utterance.
3. If novel information is present, proceed to annotate whether it relates to predefined codes, such as social, contextual, or personal values (deductive coding).
4. Additionally, we annotate any emerging topics the utterance contain, applying inductive coding.

To summarize, across all dialogue strategies, we measure self-conscious emotions, DEQ, accuracy, and completeness. Additionally, we coded the responses in both the Exploratory Dialogues and Focused Dialogues using deductive coding. For the Exploratory Dialogues, we further employed inductive coding, as we were particularly interested in identifying and merging topics that did not fall under social, contextual, or personal values. Figure 3.1 shows an overview of the study design and data analysis.

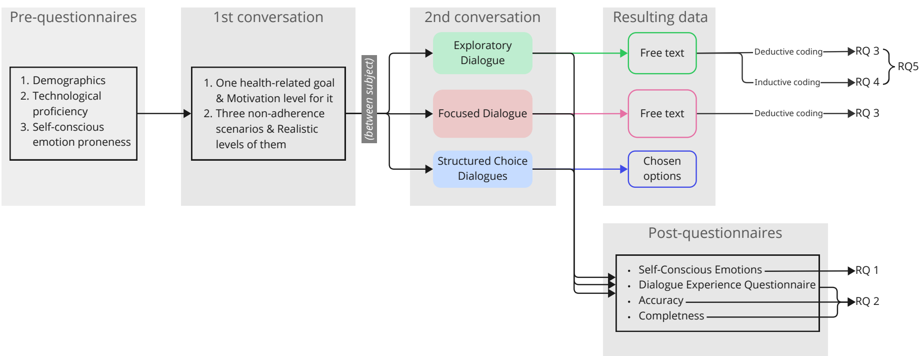


Figure 3.1: Overview of the study design and data analysis.

3.5 Results

A thorough assessment of the control variables, namely, gender, age, education, and emotion proneness, indicated no significant variations across the different dialogue strategy groups.

3.5.1 Self-Conscious Emotions Analysis

To answer RQ1, "What are the effects of different dialogue strategies on the occurrence of self-conscious emotions?", we first calculated the averages and standard deviations of each emotion in different dialogues, which are shown in Table 3.1 and illustrated by Figure 3.2. These demonstrate the participants' self-conscious emotions in response to various dialogue strategies after engaging in alignment dialogues. On average, across all dialogue conditions, participants reported the highest emotional experience of Guilt (mean = 2.90), followed closely by Authentic Pride (mean = 2.79), then Shame (mean = 2.5), with the least being Hubristic Pride (mean = 2.48).

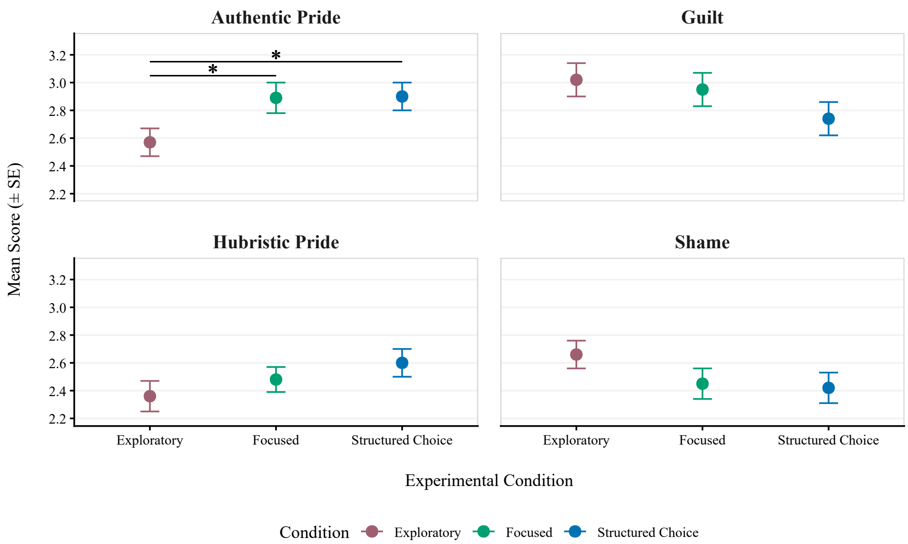


Figure 3.2: Plot of emotions by dialogue strategy with error bars.

Table 3.1: Mean (SD) of self-conscious emotions reported by participants across different dialogue strategies, measured on a scale of 1 to 5 (1 indicating "strongly disagree" and 5 indicating "strongly agree").

	Shame	Guilt	Authentic Pride	Hubristic Pride
Exploratory	2.66 (0.10)	3.02 (0.12)	2.57 (0.10)	2.36 (0.11)
Focused	2.45 (0.11)	2.95 (0.12)	2.89 (0.11)	2.48 (0.09)
Structured Choice	2.42 (0.11)	2.74 (0.12)	2.90 (0.10)	2.60 (0.10)
Overall	2.51 (0.93)	2.90 (1.05)	2.79 (0.92)	2.48 (0.90)

Furthermore, we applied a linear regression model to examine variables' influence, including the participants' self-conscious emotion proneness, their motivation levels to engage in the goal behaviors, and how realistic the challenging situations are (detailed in Subsection 3.4.3). Our predictor selection employed a stepwise procedure that merged both forward and backward selections, ensuring the identification of the optimal model.

Shame The regression analysis indicated that participants' shame proneness was a significant predictor of the level of shame they reported experiencing during the alignment dialogues ($\beta = 0.46, p < .001$). Other factors did not emerge as significant predictors in the final model.

Guilt Similarly, the regression analysis showed that participants' shame proneness was a significant predictor of the guilt experienced during the alignment dialogues ($\beta = 0.43, p < .001$).

Authentic pride Our analysis showed the dialogue strategies and motivation level to be significant predictors of feelings of authentic pride. Specifically, participant motivation to achieve their goals was positively associated with feelings of authentic pride ($\beta = 0.27, p < .001$). Participants in the Exploratory Dialogues group felt less authentic pride than those in the Focused Dialogues group ($\beta = -0.28, p < .05$). However, there was no clear difference in feelings of authentic pride between the Focused Dialogues and Structured Choice Dialogues groups, and we did not find any interaction between the dialogue strategies and motivation levels.

Hubristic pride For hubristic pride, the regression analysis indicated participants' motivation to achieve their goals as a significant predictor of hubristic pride ($\beta = 0.23, p < .001$).

3.5.2 Dialogue Experience Questionnaire (DEQ), Accuracy and Completeness

RQ2 addresses the effects of different dialogue strategies on the dialogue experience, as well as on the subjective accuracy and completeness of the information acquired. The results are descriptively summarized in Table 3.2, providing insights into how each strategy influences these aspects of user interaction.

Table 3.2: Descriptive statistics of Dialogue Experience Questionnaire (DEQ), measured on a scale from 1 (Strongly Disagree) to 7 (Strongly Agree) with statements regarding satisfaction and realism, Accuracy, rated on a scale 1 (Not at all accurate) - 5 (Extremely accurate), and Completeness, rated on a scale 1 (Not at all complete) - 5 (Extremely complete)

	DEQ	Accuracy	Completeness
Exploratory Dialogues	4.13 (0.93)	4.03 (1.13)	4.12 (0.98)
Focused Dialogues	4.37 (0.87)	4.06 (0.86)	4.17 (0.78)
Structured Choice Dialogues	4.41 (0.86)	3.96 (0.82)	3.78 (0.87)
Overall	4.30 (0.89)	4.02 (0.94)	4.02 (0.89)

We applied the same linear regression method used for emotions. In our analysis, we included dialogue strategy, motivation levels toward achieving the goals, and perceptions about the likelihood of encountering non-adherence scenarios. We also considered background factors like education level, computer programming skills, and possession of a STEM degree, believing that technological expertise might influence their assessment of the dialogue.

DEQ Our regression analysis revealed that participants' perception of how realistic the scenarios were, was negatively associated with their DEQ scores ($\beta = -0.17, p < .05$). Additionally, participants' motivation levels were positively associated with their DEQ scores ($\beta = 0.33, p < .001$). The dialogue strategy used had an impact too. Specifically, participants in the Exploratory Dialogues condition reported significantly lower DEQ scores than those in the Focused Dialogues condition ($\beta = -0.41, p < .05$). No significant difference in DEQ scores was observed between the Focused Dialogues and Structured Choice Dialogues groups. Interestingly, education levels themselves did not have a main effect. However, after aggregating these into two main categories: "High school" and "Above Bachelor" (which combines Bachelor, Master, and PhD levels), we found significant interactions between education level and dialogue strategies. The interaction effect is demonstrated in Figure 3.3.

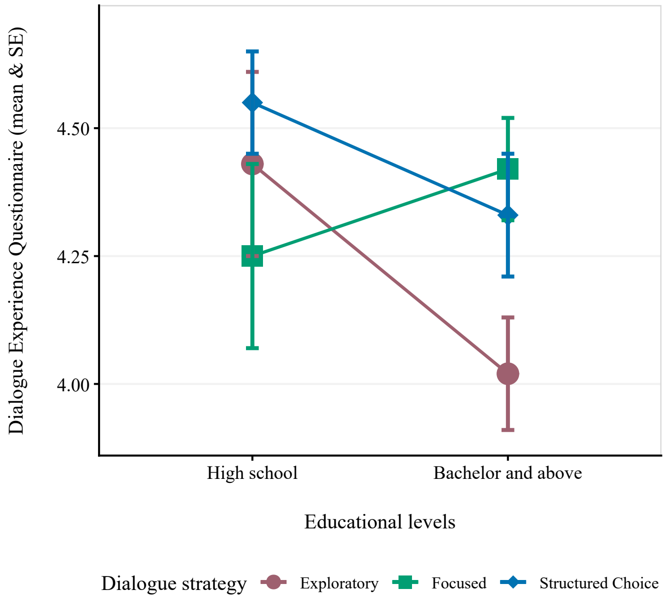


Figure 3.3: Interaction plot of DEQ by aggregated education and condition (Mean with error bars).

Table 3.3 shows the DEQ scores in different dialogue strategies with different educational levels. The interaction between the Structured Choice Dialogues and High school was significant ($\beta = 0.67, p < .05$), Similarly, the interaction between Exploratory Dialogues and High school also showed significance ($\beta = 0.60, p < .05$).

Accuracy We observed a significant correlation between participants' accuracy ratings and their perceived realism of non-adherence scenarios ($r = 0.19, p < .05$).

Table 3.3: Dialogue experience questionnaire (DEQ) scores across different dialogue strategies by educational level.

	High school		Bachelor and above	
	n	Mean (SD)	n	Mean (SD)
Exploratory Dialogues	17	4.43 (0.72)	58	4.02 (0.97)
Focused Dialogues	24	4.25 (0.86)	54	4.42 (0.88)
Structured Choice Dialogues	30	4.55 (0.67)	50	4.33 (0.95)

3

Completeness Regarding the completeness of the reasons, we found a significant correlation with participants' motivation levels ($r = 0.18, p < .01$) and the dialogue strategies. Specifically, participants in the Structured Choice Dialogues condition reported significantly less complete reasons compared to those in the Focused Dialogues condition ($r = 0.18, p < .01$). No significant difference in completeness level was observed between the Focused Dialogues and Exploratory Dialogues groups, and there was no interaction effect between dialogue strategies and motivation levels. Table 3.4 presents the summary of all the regression models.

3.5.3 Text responses analysis

In this section, we present the results of text analysis. Deductive coding reports the occurrences of predefined codes, answering RQ3. Inductive coding reveals emergent codes, answering RQ4.

Deductive coding

To address RQ3, "How effective are different dialogue strategies in eliciting targeted information?", we calculate the percentage of responses that yielded novel information. Then, within this set of responses containing novel information, we further calculate the percentages that pertain to predefined categories of interest: social factors, contextual factors, and personal values. Table 3.5 provides a detailed breakdown of these percentages, showcasing the effectiveness of each dialogue strategy in eliciting the predefined information.

For Exploratory Dialogues, the proportion of responses containing novel information varied across the dialogue questions (Q1 to Q4), with Q1 having the highest at 83%, followed by Q3 at 79%, Q2 at 62%, and Q4 at 27%. Social and contextual factors were most frequently mentioned in Q1, though less so in subsequent questions. Interestingly, while Q1 showed little mention of personal values, this topic was most predominant at 29% of novel information in Q4, with 23% of all novel information responses in Q2.

For Focused Dialogues, the novel information was predominantly in the answer to Q1 at 82%, out of which 48% provided information on this aspect. This was followed by Q2 at 56% (out of which 27% provided contextual information) and Q3 at 51% (out of which 30% provided value information).

Inductive coding

To address RQ4 —"Besides the targeted information, what additional topics do users mention?" —we employed inductive coding in the Exploratory Dialogues. This approach was

Table 3.4: Final regression models for assessing the effects of dialogue strategies on self-conscious emotions, accuracy, completeness, and DEQ. The table only displays predictors with significant effects. For categorical variables (e.g., dialogue strategy), all levels are shown for clarity and to indicate the reference level used in comparisons.

Measures	**Variables (Predictors)**	β	*SE*	p **value**
Shame	Shame proneness	0.43	0.10	$< .001^{***}$
Guilt	Shame proneness	0.46	0.11	$< .001^{***}$
Authentic pride	Dialogue strategy: Structured Choice	-0.02	0.14	0.91
	Dialogue strategy: Exploratory	-0.28	0.14	$< .05^{*}$
	Motivation level	0.27	0.06	$< .001^{***}$
Hubristic pride	Motivation	0.23	0.06	$< .001^{***}$
Accuracy	Realistic level	0.18	0.07	$< .05^{*}$
Completeness	Dialogue strategy: Structured Choice	-0.41	0.14	$< .01^{**}$
	Dialogue strategy: Exploratory	-0.02	0.14	< 0.89
	Motivation level	0.18	0.06	$< .01^{***}$
DEQ	Dialogue strategy: Structured Choice	-0.25	0.16	< 0.13
	Dialogue strategy: Exploratory	-0.4	0.16	$< .05^{*}$
	Education: High school	-0.20	0.20	0.32
	Realistic level	-0.17	0.07	$< .05^{*}$
	Motivation level	0.33	0.06	$< .01^{***}$
	Dialogue strategy: Structured Choice* Education: High school	0.67	0.28	$< .05^{*}$
	Dialogue strategy: Exploratory* Education: High school	0.60	0.30	$< .05^{*}$

used to identify topics in responses that did not fall under the targeted categories of social aspects, contextual aspects, or personal values. Below, we present a list of these emergent topics with explanations and illustrative examples from the responses.

1. **Individual Challenges** encompass intrinsic factors that are inherent to the individual and are difficult to ask of them to change.

2. **Environmental and Societal Challenges** refer to broader, external factors that shape user behavior. These challenges are generally stable and persistent, rather than tied to the user's immediate circumstances.

3. **Prioritization and Preferences** underscore the influence of personal tastes, likes, dislikes, preferences and attitudes.

4. **Norms and Obligations** cover actions and choices driven by societal/personal norms or obligations.

5. **Beliefs** encompass a wide range of beliefs, such as cultural beliefs and factual beliefs.

Table 3.5: Deductive coding in Exploratory Dialogues and Focused Dialogues. The first row indicates the percentage of responses containing novel information, calculated by dividing the number of responses with novel information by the total number of responses. The following rows indicate the percentage of these novel information responses that further discuss social, contextual, and personal value aspects, respectively, calculated by dividing the number of responses addressing each aspect by the total number of responses containing novel information. See Subsection 3.3.2 for what Q1-Q4 refers to in each dialogue condition.

	Exploratory Dialogues				Focused Dialogues		
	Q1	Q2	Q3	Q4	Q1	Q2	Q3
Novel information	0.83	0.62	0.79	0.27	0.82	0.56	0.51
Social aspects	0.06	0	0	0.10	0.48	-	-
Contextual aspects	0.13	0.02	0	0	-	0.27	-
Personal values	0.03	0.23	0.03	0.29	-	-	0.30

6. **Tangible Outcomes** encompass the practical, observable consequences of behavioral decisions.
7. **Emotional Outcomes** relate to affective responses from (lack of) behaviors.

3.5.4 Patterns in User Explanations

To answer RQ5 "What patterns emerge in users' transitions between different topics when they explain their reasons for non-adherence behaviors?", we analyzed the conversation paths in the Exploratory Dialogues which start with a user's non-adherence scenario and then four questions from the agent. The primary objective was to identify patterns in how users transitioned between various topics of explanations. Our goal was to understand these transition patterns better, thereby aiding in the selection of effective questions for eliciting informative responses. Below is the series of questions the agent asked after the participant provided the non-adherence scenario:

0. **User's Non-adherence Scenario**
1. **Q1:** What particular factors related to the situation influence your choice?
2. **Q2:** Why is it important to you?
3. **Q3:** What feeling does it give you?
4. **Q4:** Why is that?

For the non-adherence scenario and each response to a question, we formed four pairs of consecutive responses which we refer to as "Duos." We adopted two distinct perspectives for analyzing the Duos: Preceding Path Analysis and Following Path Analysis. The former examines the percentage of topics *preceding* a certain topic, while the latter analyzes the percentage of topics *following* a certain topic. In other words, Preceding Path Analysis can reveal, for instance, that topic x is most often preceded by topic y. If the agent wants information about topic x, they could lead with topic y. Conversely, Following Path Analysis can show that topic y often follows topic x. Asking about topics y after topic x could make the conversation feel more natural to the users.

For the Preceding Path Analysis, we calculated the percentage of instances where topic x was led by each specific topic out of all topics that led topic x. For the Following Path Analysis, we calculated the percentage of instances where each specific topic followed topic x out of all topics that followed topic x. We applied these two approaches differently depending on the nature of the dialogues. Below are the results of each Duo.

Table 3.6: Following Path Analysis of Duo (Scenario, Q1). Only the top few topics with a cutoff point around 0.1 are shown (Abbreviations used: "Indiv." for Individual Challenges, "Contxt." for Contextual Information, "Env. Soc." for Environmental and Societal Challenges, "Pref." for Prioritization and Preferences, "Norm. Obl." for Norms and Obligations, "Tang." for Tangible Outcomes.)

Scenario	Q1	Percentage
Env. Soc.	Env. Soc.	0.27
	Pref.	0.13
	Indiv. Contxt.	0.09
Indiv.	Indiv.	0.33
	Env. Soc.	0.33
	Pref.	0.22
Motiv.	Pref.	0.16
Pref.	Pref.	0.4

1. **Duo (Scenario, Q1)**: For the Duo that goes from the start to Q1, we exclusively employed the Following Path Analysis. This choice was made because our interest lay in understanding the explanations of why these scenarios led to their non-adherence behaviors. Table 3.6 shows the top explanations that were at least 10% in the analysis. Below is a list of summary.

   - Detailed Challenges: Participants often provide detailed explanations of challenges following scenarios involving Environmental and Societal Challenges or Individual Challenges.
   - Prioritization and Preferences consistently emerge as prominent topics across various scenarios, revealing participants' considerations in prioritizing or preferring elements over behavior goals.
   - Contextual Information is in some cases provided, showing situational factors influencing behavior decisions.

2. **Duo (Q1, Q2)**: we utilized both the Following Path Analysis and the Preceding Path Analysis. Using both analyses allowed us to understand the topics both preceding and following the participants' responses.

   (a) Following Path Analysis (Table 3.7)

Table 3.7: Following Path Analysis of Duo (Q1, Q2). See Table 3.6 for abbreviations.

Q1	Q2	Percentage
Env. Soc.	Tang.	0.23
	Pref.	0.13
Indiv.	Pref.	0.15
	Tang.	0.15
Pref.	Tang.	0.27
	Pref.	0.19
Tang.	Value	0.29
Norm. Obl.	Value	0.40
Social	Value	0.43
Contextual	Value	0.33

Table 3.8: Preceding Path Analysis of Duo (Q1, Q2). See Table 3.6 for abbreviations.

Q2	Q1	Percentage
Env. Soc.	Env. Soc.	0.33
Pref.	Pref.	0.33
	Pref.	0.27
Tang.	Env. Soc.	0.23
	Pref	0.23
Beliefs	Env. Soc.	0.25
Norm. Obl.	Env. Soc.	0.43
Value	Social	0.23
	Contextual	0.16
	Env. Soc.	0.16

- After discussions on Environmental and Societal Challenges or Individual Challenges, users often move to topics related to Tangible Outcomes or Prioritization and Preferences.
- Following discussions on Prioritization and Preferences, users frequently transition to Tangible Outcomes, indicating an intertwined nature of these topics.
- When users discuss Tangible Outcomes, Norms and Obligations, Social Information, or Contextual Information, the subsequent topic tends to be Values, suggesting a likely progression in the conversation.

(b) Preceding Path Analysis (Table 3.8)

- Some topics in Q2 are most frequently preceded by the same topics in Q1, such as Prioritization and Preferences, as well as Environmental and Societal Challenges.
- Tangible outcomes are preceded most by Environmental and Societal Challenges & Prioritization and Preferences.
- Norms and Obligations, as well as Beliefs are predominantly preceded by Environmental and Societal Challenges.
- Values are mostly preceded by Environmental and Societal Challenges, Context, or Social factors.
- Above observations imply that initiating the dialogue with questions related to Environmental and Societal Challenges or Prioritization and Preferences could effectively lead to discussions about Tangible outcomes, Norms and Obligations, Beliefs, or Values.

3. **Duo (Q2, Q3)**: We skipped this due to Q3's focus on emotions, making preceding or following analysis unnecessary.

4. **Duo (Q3, Q4)**: We employed Following Analysis, as Q3 focused on emotions, we were interested in what topics followed the emotional responses.

   - After discussing emotions, users provided Value Information more than other topics. This trend suggests that probing emotional consequences could lead users to articulate their values.

3.6 Discussion

This study explored how dialogue strategies affect self-conscious emotions, user satisfaction —comprising accuracy, completeness, and dialogue experience —and the aquisition of the user reasons behind non-adherence in specific scenarios. In the following sections, we discuss our findings related to self-conscious emotions (Subsection 3.6.1) and user satisfaction (Subsection 3.6.2). Subsequently, we summarize the qualitative information regarding reasons for non-adherence behaviors (Subsection 3.6.3). Finally, we highlight other noteworthy observations that emerged during our study (Subsection 3.6.4 & Subsection 3.6.5).

3.6.1 Emotions in Different Dialogue Strategies

Our analysis reveals that emotions of shame, guilt, and hubristic pride remain largely consistent across the different dialogue strategy conditions (see Figure 3.2). This suggests that the dialogue strategies did not have a significant impact on the manifestation of these self-conscious emotions.

However, a notable observation is that participants in the Exploratory Dialogues condition reported lower levels of authentic pride in comparison to those in the Focused Dialogues and Structured Choice Dialogues conditions. Authentic pride is often associated with feelings of accomplishment and the motivation to pursue goal-oriented actions [56–58]. The straightforwardness inherent in the Exploratory Dialogues might give participants the impression that the agent is subtly challenging the value or importance of their goals. This finding may be attributed to the fully open-ended nature of the why-questions in the Exploratory Dialogues, which may be perceived as seeking justification. The ambiguous nature of why questions may lead to unproductive or even hostile responses [90]. Conversely, the Focused Dialogues and Structured Choice Dialogues strategies, which employ more specific and targeted questions, could be perceived as being less direct and intrusive.

To summarize the implications with respect to self-conscious emotions: individual shame proneness is a clear predictor of shame and guilt. Beyond that, motivations seem intertwined with authentic pride and hubristic pride. How the open questions are framed in the Exploratory Dialogues strategy seems to negatively impact authentic pride.

3.6.2 DEQ, Accuracy, and Completeness in Different Dialogue Strategies

Regarding user satisfaction, our observations reveal that there were not any significant differences in the accuracy of the reasons for non-adherence. However, when it comes to the completeness of the reasons, participants in the Structured Choice Dialogues tended to perceive the summary as significantly less complete compared to their counterparts in the Exploratory Dialogues or Focused Dialogues. This result aligns with our expectations since Structured Choice Dialogues participants were presented with predefined multiple-choice options that covered social, and contextual aspects of the situation, and values. The restrictive nature of these options may have hindered participants from providing more expansive responses, given the inherent limitations of the setup [66].

Furthermore, we identified a significant decrease in the dialogue experience score (DEQ) for the Exploratory Dialogues compared to the Focused Dialogues and Structured Choice Dialogues conditions. This trend mirrors the patterns observed with authentic pride. Further examination revealed a strong correlation between authentic pride and DEQ ($\beta = 0.4, p < .001$), reinforcing the interplay between the dialogue strategy, and resultant pride emotions and DEQ. This might be because Exploratory Dialogues demand additional time and cognitive effort [30, 66], contributing to a worse dialogue experience compared to other conditions.

To summarize the implications on user satisfaction: the realistic level of non-adherence scenarios and participants' motivation level to achieve their goals emerge as key predictors for the overall user experience. Additionally, participants in Exploratory Dialogues and Focused Dialogues perceive the summary as significantly more complete than Structured Choice Dialogues, while participants in Exploratory Dialogues had a significantly lower DEQ compared to those in both the Focused Dialogues and Structured Choice Dialogues conditions.

Interaction between Education and Strategy on DEQ

Figure 3.3 depicts the interaction between education levels and dialogue strategy in terms of DEQ. These results suggest that, for the High School education level, being in the Structured Choice Dialogues or Exploratory Dialogues leads to a significant increase in the DEQ mean compared to Focused Dialogues. A plausible interpretation of this observation is grounded in the inherent structure of each strategy. As the education level transits from High School to "Bachelor and above" education, we see the DEQ mean for Exploratory Dialogues decrease, for Structured Choice Dialogues slightly decrease, for Focused Dialogues increases. This shift suggests that Focused Dialogues might be more suitable or preferred for those with higher education, while Exploratory Dialogues and Structured Choice Dialogues might be more effective for those with a High School education level.

Summary Quantitative Analysis and Real-world Implication

The insights highlight the nuanced interplay between dialogue strategies, self-conscious emotions, and the richness of user responses. Interestingly, while the study did not reveal significant distinctions in self-conscious emotions such as shame, guilt, and hubristic pride across different strategies, certain trends emerge in relation to authentic pride and user experience.

The Exploratory Dialogues strategy, in particular, registered lower scores for authentic pride and DEQ, while Structured Choice Dialogues showed a noticeable decline in the completeness of the provided information. This suggests a trade-off exists between the level of restraints imposed by the dialogue agent's questions and the likelihood of acquiring necessary information from users. Employing explicit prompts, as seen in the Focused Dialogues or Structured Choice Dialogues approaches, tends to yield more precise information. However, such specificity might constrict the depth or breadth of user elaboration, with the Structured Choice Dialogues method being a prime example. Focused Dialogues, while beneficial for many, can pose interpretative challenges, particularly for individuals with solely a high school education. Contrarily, the unrestricted nature of the Exploratory Dialogues method, though fostering freedom of expression, appears to diminish feelings of pride in respondents.

In essence, each dialogue strategy presents unique advantages and limitations. Real-world implementations must cater to diverse users and contexts. For instance, a hybrid approach could be employed, combining Structured Choice Dialogues as guiding questions followed by open-ended questions to allow users to articulate freely or provide supplementary information. Additionally, Dewdney and Michell [90] suggest the importance of contextualization and using neutral phrasing when employing why-questions to mitigate the risk of eliciting unproductive or even hostile responses.

3.6.3 Text Responses Comparison between Exploratory Dialogues and Focused Dialogues

The deductive analysis aimed to identify and categorize the emergence of novel information related to social contexts, environmental contexts, or personal values within the responses. Utilizing an inductive coding approach, we analyzed responses that introduced emerging topics. This in-depth examination allowed us to understand the breadth of content participants shared during alignment dialogues.

From the results, both the Exploratory Dialogues and Focused Dialogues, across each query round, showed an average of approximately 62% inclination to provide novel information. It is worth noting that the last *why* query within the Exploratory Dialogues condition exhibited a mere 27% introduction of novel information. Such a decline might hint that participants were no longer able to add more reasons by the fourth question or an ambiguous interpretation stemming from the preceding (Q3) query. Another observation of Q3 is its high percentage of novel information elicited. This could be because Q3 is very specific in asking "What feelings does it give you." Our intention behind leaving Q2 and Q4 ("Why is it important to you?") ambiguous was to gauge how participants would respond when presented with very open-ended queries. The significant reduction in novel information suggests there might be a limit to the effectiveness of repeatedly asking open-ended questions. This could be attributed to participants reaching a point where they no longer provide new insights or being fatigued from the continuous questioning.

Within the Exploratory Dialogues condition, the findings showed that on average only 7% of the novel responses touched on social aspects or contextual facets, while a relatively higher 15% on personal values. Such numbers, being on the lower side, suggest the difficulty in sourcing comprehensive data for constructing an user model using very open-ended questioning. In contrast, the Focused Dialogues condition showcased better results: 48% novel responses introduced social nuances, 27% provided context, and 30% on personal values.

An interesting observation from the Exploratory Dialogues data is the higher number of value-laden responses compared to mentions of social or contextual factors. The recurrent theme in Exploratory Dialogues's questions, centering on "why is it important to you," might have steered participants to introspect and articulate their intrinsic values more. However, nuanced specifics about the situation, especially those tied to social or environmental contexts, seldom emerged organically from the respondents.

Summary Deductive Analysis and Real-world Implication

Our experiments show that specificity in questions increases the chances of acquiring the targeted information. However, a more restrictive approach, like the multiple-choice format used in the Structured Choice Dialogues, might limit the depth and variety of responses. Real-world implementations need to find a balance between providing structure and allowing flexibility. Combining specific, guiding questions with opportunities for open-ended responses could help gather comprehensive insights while accommodating diverse user needs and contexts.

3.6.4 New Topics for User Model

The inductive coding analysis revealed new topics (Subsection 3.5.3 that can enrich the user model. Below we group similar topics to discuss.

Individual Challenges and Environmental and Societal Challenges

A significant finding was the emergence of Individual Challenges and Environmental and Societal Challenges. These topics underscore the importance of considering various constraints that users may face. These constraints often fall beyond users' control, such as personal challenges, health conditions, or specific work-related restrictions. Including these in the user model offers a deeper capture of a user's background and environment. Notably, these findings align with the concept of *barriers* discussed in the behavior change literature, emphasizing factors like family commitments, societal pressures, limited facilities, or perceived time constraints (e.g., [78, 295, 302]).

However, we suggest that such factors should be treated separately from the information elicited in alignment dialogues. Social and contextual aspects are situation-specific, while personal and environmental constraints represent more general background information that could be collected during the user initialization phase, as discussed in Section 3.1.

Preferences, Norms, and Obligations

We also identified a topic revolving around attitudes, preferences, obligations, and the like. This topic delves into users' preferences for various activities and how they prioritize these activities in their lives, similar to the findings in Kearney and McElhone [158]. While some preferences can reflect underlying values, others are simply personal choices devoid of deeper meaning. The related topic, Norms and Obligations serves similar functions but instead of personal preference, they are usually driven by societal and personal norms, obligations, and external pressures [180]. Whether it is a matter of prioritization, preference, or adhering to norms, these insights can be integrated as Norms in the user model, guiding the selection of support actions by the agent.

Tangible Outcomes & Emotional Outcomes

The remaining topics primarily revolved around the outcomes of participants' actions or behaviors, categorized as desires (typically positive outcomes) and consequences (often negative outcomes). Some of these outcomes are deeply intertwined with personal values, as depicted by statements like "So that I can take care of my beloved ones." Conversely, others reflect a more pragmatic stance and do not necessarily embody underlying values, as in the statement "I will go to jail for negligence." In the Exploratory Dialogues condition, one of the question rounds (Q3) asked "What feeling does it give you?" Such a question yielded valuable insights in terms of emotional consequences. Nearly 80% of responses included novel information, with 2.6% containing topics related to values.

Additionally, subcategories (see Subsection 3.5.3) – specifically "Outcomes about Work and Career/Social Relationships with Others/Appearance" – are closely intertwined with values like warm relationships with others, security or self-fulfillment, and self-respect. Based on this, we posit that exploring users' anticipated outcomes (both desires and consequences) could be a strategic method to delve deeper into their core values.

The incorporation of these novel topics into the user model introduces a more holistic and personalized understanding of users, enabling agents to offer personalized support that respects individual constraints, priorities, and motivations. However, it is crucial to strike a balance between the comprehensiveness of the user model and the efficiency of interactions. Careful consideration is needed to avoid overwhelming users or agents with excessive detail.

3.6.5 Pitfalls of Alignment Dialogues

Our analysis identified some challenges with alignment dialogues. In Structured Choice Dialogues, we observed instances where participant responses did not logically align with their stated non-adherence scenarios. For example, a participant whose goal is to go to the gym cited "contact frequency" as a contributing factor while being sick—a scenario where social factors seem irrelevant. Such responses suggest a need for deeper user engagement or point to possible misunderstandings of the dialogue prompts.

Furthermore, in Exploratory Dialogues, approximately 15% of the responses were not about why not adhering to goals but about the reasons behind setting these goals in the first place. This underscores the inherent ambiguity in open-ended questioning and the difficulty in asking clear, purpose-driven questions.

Despite these issues, we chose not to exclude these "incorrect" responses from our analysis for two reasons. First, the proportion of these responses was not substantial. Second, all responses, regardless of their direct relevance to non-adherence reasons, contribute valuable insights into users' reasons and can enhance our understanding of their behavior choices.

3.7 Reflection on Alignment Dialogue Design

In this section, we further discuss the insights from this study and how they may inform the design of alignment dialogues. We begin with an analysis of non-adherence scenarios Subsection 3.7.1, as they serve as the starting points for initiating an alignment dialogue. Lastly, we analyze the conversation *paths* by examining the percentages of the sequential occurrence of topics Subsection 3.7.2.

3.7.1 Non-adherence Scenarios

We classified the non-adherence scenarios reported by participants during the initial phase of our experiment. Our rationale stems from the understanding that distinct scenarios might require different dialogue strategies to be effective.

Many of the non-adherence scenarios reported by participants align with topics identified in the inductive coding of responses (Subsection 3.5.3). This overlap was expected, as participants' explanations expanded upon the initial scenarios, providing additional insights. Nonetheless, we also identified several unique non-adherence scenarios that distinguished themselves from the common categories:

1. Lack of Motivation
2. Lack of Energy
3. Emotional State
4. Time constraints

These scenarios primarily highlight participants' mental or motivational states hindering their commitment to goal behaviors. Participants often provided general reasons without delving into specific circumstances or underlying causes of non-adherence. Vague explanations lack specificity, failing to offer actionable insights into user behavior. For instance, the statement "I don't have time" lacks clarity on activities consuming time and potential adjustments. Deeper exploration is needed to uncover underlying factors in such scenarios.

3.7.2 Dialogue Paths for Alignment Dialogue

Drawing from the Preceding Path Analysis and Following Path Analysis of the duos (detailed in Subsection 3.5.4), several patterns emerge that can inform effective dialogue design. Firstly, when participants present vague scenarios discussed in Subsection 3.7.1 like "I don't have time", the agent's inquiries should aim to extract specific details. Following discussions on Individual Challenges or Environmental and Societal Challenges, further inquiries often yield Contextual specifics about the situation. This indicates that in real-world implementations, when users discuss Individual Challenges or Environmental and Societal Challenges, it would be effective for the agent to follow up with inquiries about the Contextual specifics of the situation.

Furthermore, discussions about Tangible/Emotional Outcomes, Norms and Obligations, Social aspects, or Contextual aspects often lead to conversations about Values. This pattern indicates that these topics can pave the way for discussions about values.

Overall, the recommended dialogue flow in practical applications might progress from addressing Challenges to exploring Prioritization and Preferences, followed by Tangible Outcomes, Norms and Obligations, Beliefs, Emotional Outcomes, and finally, Values. However, the progression is not linear due to the intertwined nature of these topics, necessitating the agent to adapt its approach based on context and user cues.

3.8 Limitations and Future work

3.8.1 Emotions in Alignment Dialogues

This study suggests that self-conscious emotions do not significantly differ across different dialogue strategies. However, more research is needed to understand the precise relationship between dialogue strategies, setup, and the experience of self-conscious emotions.

Furthermore, we did not measure self-conscious emotions before and after the participants engaged in the dialogues. As a result, we could not determine which specific dialogue strategies evoke particular emotional responses. Although it isn't the agent's goal to evoke negative self-conscious emotions, understanding the precise relationship between dialogue strategies and these emotions could further strengthen the design of alignment dialogues. This would ensure that the dialogues are both effective and sensitive to users' emotional states.

A key limitation of this experimental setting is that the scenarios are not real, in the sense that participants engaged with them as part of a study rather than a personal health journey. Consequently, the stakes of non-adherence for the individual remained low, which likely produced less socio-emotional engagement than would be required to elicit deeply held values, beliefs, or reasons.

Additionally, future studies could benefit from including variations in ethnicity, geographic location, cultural and socioeconomic backgrounds into the analysis. These factors could influence how self-conscious emotions are experienced and expressed during alignment dialogues, providing a more comprehensive understanding of the user experience across different demographics.

Another analysis is to explore the types of health goals participants set and identify any patterns associated with specific goals and the challenges they face. Understanding these patterns could help tailor dialogue strategies to better support users in achieving their health objectives.

3.8.2 Alignment Dialogue Design

While this study addresses the *what* aspect of alignment dialogue (i.e., "what information should the agent seek in alignment dialogue?"), it provides limited guidance on the *how*. It is important to recognize that the *how* of behavior change dialogues can be addressed at different levels. Existing research, such as the work of Ranjbartabar et al. [264] and Salman et al. [284] focus on how the use of relational cues can build rapport and influence the effectiveness of the interaction. There is, however, the *how* concerned the structural modeling and logic of the dialogue that requires further investigation. Future work is needed to enrich and refine the dialogue path we suggested. Implementing the dialogue in practice will require a more detailed structure, along with clearly defined relationships between each topic. Our work primarily outlines the topics but does not model how these topics interconnect. Additionally, the relevance of each topic may vary depending on the situation. Identifying which topics are pertinent in specific contexts is a critical area for future exploration.

3.9 Conclusion

The purpose of the present work was to investigate 1) the effects of different dialogue strategies on users' self-conscious emotions, dialogue experience, perceived accuracy, and perceived completeness 2) which reasons users mention to explain their non-adherence behavior. The ultimate goal was to better capture *what* constitutes a comprehensive user model and how agents could effectively gather this information without triggering negative emotional responses.

Our findings show that the open-ended general question had significant negative effects on authentic pride, dialogue experience, and the perceived completeness of the explanations collected. For other aspects namely shame, guilt, and perceived accuracy, the dialogue strategies did not show a significant effect.

Furthermore, we used inductive coding to analyze what topics participants mentioned to explain their non-adherence behavior. This analysis highlighted key topics that should be incorporated into user models. This list of topics is aimed at enabling the agent to form a comprehensive model of users, encompassing their personal constraints, beliefs, values, and preferences. The primary objective of alignment dialogues, therefore, is to gather information within this ontological framework, thus facilitating a deeper and more nuanced understanding of user behavior in various contexts.

4

Expert Insights on Conversational AI Systems as an Information Intermediary for Patients and Healthcare Providers for Diabetes Lifestyle Change

This paper explores the potential of conversational intermediary AI (CIAI) between patients and healthcare providers, focusing specifically on promoting healthier lifestyles for Type 2 diabetes. CIAI aims to address the constraint of limited healthcare provider time by acting as an intermediary in-between infrequent consultations. CIAI enables healthcare providers to understand patients better and offer personalized support. Through an exploratory focus group with healthcare domain experts, we gather insights into CIAI's envisioned in diabetes care. Our findings highlight the potential benefits of CIAI in diabetes care.

4.1 Introduction

Recent advances in digital technology provide opportunities for digital healthcare solutions [203]. In particular, conversational AI is a promising tool for facilitating behavior change interventions due to its ability to engage in natural conversations and build user relationships [342]. This paper explores experts' insights on the potential of conversational AI as an intermediary between patients and healthcare professionals, specifically in

Pei-Yu Chen, Sophie van Gent, M. Birna van Riemsdijk, Myrthe L. Tielman, and Tjeerd Schoonderwoerd. "Expert Insights on Conversational AI Systems as an Information Intermediary for Patients and Healthcare Providers for Diabetes Lifestyle Change." In: Kiemute Oyibo, Wenzhen Xu, Elena Vlahu-Gjorgievska (eds.), *The Adjunct Proceedings of the 19th International Conference on Persuasive Technology, April 10, 2024, Wollongong, Australia.*

promoting healthier lifestyles for Type 2 diabetes (T2D) patients. Given the importance of maintaining a healthy lifestyle for long-term T2D management [84, 145], exploring the role of AI in this area is essential. The Conversational Intermediary AI (CIAI) proposes that it will, via dialogues, learn about the users within the context of their daily lives and select relevant information to communicate back to healthcare providers.

Personalized interventions have demonstrated greater effectiveness in improving T2D-related parameters compared to usual care [85, 98]. This effectiveness is shown in digital interventions as well. Previous research consistently shows that personalized approaches outperform generic "one-size-fits-all" interventions in promoting health behaviors (e.g., [120, 176, 197, 243]). Personalization strategies encompass using a variety of information, including socio-demographic characteristics, personality traits, behavior determinants, and habits [157, 176].

While integrating these factors is beneficial, they often represent static characteristics that may not fully reflect the dynamic context and nuances in which users operate. This is of particular importance for healthy lifestyle changes, which are very intertwined with users' daily lives. Chen et al. [62] proposed alignment dialogue, a conversational AI approach between AI and users, as a solution aiming to ground the user model in the current context. However, the practical implementation of acquiring a comprehensive user profile through conversational agents remains a topic for further investigation.

To address this gap in how AI could acquire a comprehensive user profile, in this work, we explore how healthcare providers converse with T2D patients to understand them better for lifestyle changes. By examining current practices, we seek insights that could inform conversational AI systems in acquiring a comprehensive user profile. However, in this domain, a significant challenge is the infrequent occurrence of healthcare consultations[1]. This scenario presents a good opportunity for Conversational AI as an intermediary (CIAI) between healthcare providers and patients in-between consultations: the role of CIAI consists of capturing patients' needs in their daily lives and conveying this comprehensive patient information back to healthcare professionals. In this way, healthcare providers can make tailored lifestyle change suggestions that are easier to adopt and maintain by the patient.

Using digital systems to bridge gaps between consultations and facilitate information flow between different healthcare levels has been previously explored. For example, Richards and Caldwell [267] developed a "virtual specialist" system that shared diagnostic insights with General Practitioners while patients awaited specialist appointments, demonstrating that interactive digital intermediaries can improve health outcomes by supporting ongoing patient management.

However, to our knowledge, there has been limited research on CIAI designed to capture users in their everyday lives and subsequently summarize this insights for healthcare providers. To explore this concept, we conducted an exploratory focus group with healthcare domain experts. This study aimed to gather their perspectives on the potential of CIAI to enhance care for diabetes patients and inform the practices of healthcare professionals. Broadly, there are two primary research questions.

RQ1. How are lifestyle changes currently managed in diabetes care, and what challenges

[1]https://richtlijnen.nhg.org/standaarden/diabetes-mellitus-type-2

do healthcare providers face in practice?

RQ2. What are the expectations and concerns of healthcare experts regarding the proposed conversational Intermediary AI (CIAI) system?

The findings of this expert study shed light on the dynamics of provider-patient conversations and highlight the opportunities for CIAI. However, alongside these opportunities, there are also notable concerns and ethical considerations. These insights point toward the need for further research directions.

4.2 Related Work

The concept of CIAI between patients and healthcare providers can be related to the research on the user (patient) modeling (Subsection 4.2.1) and ecological momentary assessment (Subsection 4.2.2).

4.2.1 User Modeling

User modeling aims to provide personalized interventions by capturing various aspects of users, such as employing different persuasive strategies to resonate with diverse personality traits [196, 325] and welcoming each participant using personalized messages [174]. Another example is to capture users' motivational attitudes [320], such as values and preferences. Capturing these aspects is crucial for healthcare providers to tailor interventions effectively. To capture these with AI, AI needs a user model to know what to ask for. Traditionally, values can be acquired explicitly through various questionnaires. However, these methods are often not grounded in a context [193, 257] and may not accurately represent real-life behaviors [42]. It is crucial to consider how individuals apply different values in varying situations and contexts [140, 193].

4.2.2 Ecological Momentary Assessment

The concept of getting information relevant to the user's behaviors in a situation can be likened to the Ecological Momentary Assessment (EMA) in clinical psychology. Instead of the conventional retrospective self-report assessment in clinical psychology, EMA involves gathering subjects' current behaviors and experiences in real-time within their natural environments [294]. EMA offers advantages such as providing more valid and detailed data about real-world behavior and experience [294]. By having an AI with the patients in their daily lives, it can gather extensive insights into the factors influencing patients' behavior choices across various daily situations. It complements EMA, which collects real-time self-reports but may lack nuanced motivations and contextual details.

4.3 Conversational Intermediary AI functionalities

In this section, we describe the envisioned CIAI system. Although the focus of this work is on AI as an intermediary information-collecting system for healthcare professionals, we added some *support* functionalities to envision how generally AI could interact with patients and healthcare providers. Figure 4.1 shows these envisioned interactions. The envisioned system has five functionalities. These functionalities were discussed in the focus group.

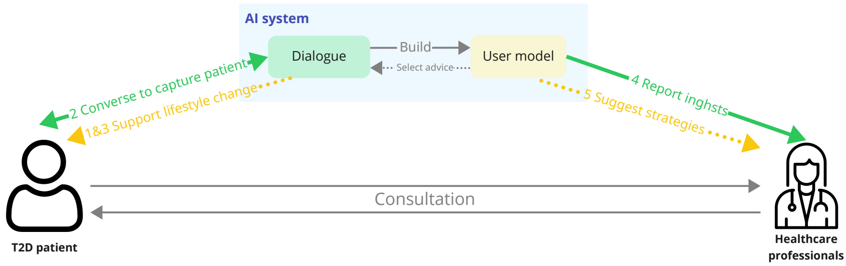


Figure 4.1: An overview of the functionalities. The green solid lines indicate the main functionalities of AI as an intermediary system between patients and healthcare professionals. The yellow dotted lines indicate the support functionalities.

1. **Remind the user to adhere to their healthy behavior goals.**
   This basic function involves the AI system asking the user about their healthy lifestyle goals and providing support accordingly.

2. **Capture reasons why the user is not adhering.**
   The AI system engages in dialogue with the patient to capture the reasons behind non-adherence and updates the patient profile with additional information such as needs and values in different contexts.

3. **Generalize from previous insights and apply to future occurrences.**
   Based on past interactions, the AI system proactively advises the patient and healthcare providers accordingly when similar contexts occur.

4. **Summarize the insights for healthcare providers.**
   The AI system provides a summary of its insights on the patients to healthcare providers.

5. **Suggest healthcare providers on possible treatment/strategies.**
   Using its comprehension of the patient, the AI system suggests the healthcare provider treatment strategies that are likely to be adopted by the patient.

4.4 Methods

To better understand experts' perspectives on this kind of AI system in diabetes care, we performed a focus group study to explore their opinions and ideas. We conducted the focus group in a hybrid format to accommodate both in-person and online participants.

4.4.1 Participants

Five healthcare experts contributed their expertise to the focus group, representing a diverse range of specialties within the field. The group included four senior scientists in personalized health, two of whom work part-time as healthcare professionals. One mid-level researcher experienced in customized machine learning for the analysis and integration of health data. These experts with healthcare practice and/or expertise in e-health provide valuable insights into the design considerations for conversational AI systems in healthcare.

4.4.2 Material

The focus group session was structured into three distinct parts. In the first part, we explored the current state of diabetes care in practice. This was followed by presenting a scenario about the envisioned AI system. Lastly, we discussed participants' opinions regarding this AI system.

Work sheets

To facilitate the discussion and note-taking, we provided worksheets in both digital formats (made in Miro[2], for the participants joining online) and paper (for participants joining in person). The first set of worksheets focused on current practices in diabetes care and the second set of worksheets centered on the envisioned AI system. These worksheets contained pre-formulated discussion questions (Subsection 4.4.3) and dedicated spaces for note-taking.

Scenario

For the second part of the focus group, we presented a storyboard depicting the interaction between a patient, a GP assistant, and a system for healthy lifestyle change (visualized as a chat interface app on a phone). This storyboard consisted of six scenes, corresponding to the functionalities in Section 4.3, with some additional scenes created or combined. The scenario was created to facilitate the discussion. It illustrates how the AI system could capture the patient's various aspects in their daily context and offer insights to healthcare providers.

1. **Scene 1** Via dialogue, the patient tells the AI system that his goal is to exercise three times per week.

2. **Scene 2** The patient adheres to his exercise plan. He registered with the agent when he exercised and he did so three times a week.

3. **Scene 3** The AI system notices that the patient has not registered any exercise activity for five days in a row.

4. **Scene 4** The AI system initiates a dialogue with the patient, trying to capture the reasons for the deviation from the routine. The patient explained that he had family visiting, which took priority over his exercise routine. The AI system updates this information in the patient's profile.

5. **Scene 5** A few weeks later, the agent notices another upcoming family event in the patient's calendar - the birthday celebration of his grandson. The AI system recognizes a similar context to the previous family visit and suggests that the patient discuss with the GP assistant how to balance enjoying the celebration with managing his nutrition.

[2]https://miro.com/

6. **Scene 6** The AI system proactively communicates with the patient's GP assistant about the significance of family in the patient's life. This information enables the GP assistant to have a more informed discussion with the patient, plan his nutrition around the celebration, and offer guidance on healthier food choices or necessary adjustments in medication.

4.4.3 Procedure

At the beginning of the session, the study's overall objective was explained to the whole group. For the in-person participants, working sheets were displayed on the walls with post-its for notes and comments. Online participants were provided with digital versions of these sheets and post-its. The study was divided into two parts: current diabetes care, and a scenario with an envisioned AI system.

Part I: Current Diabetes Care

The first part focused on the current state of diabetes care. The in-depth discussion questions covered various phases of diabetes care, the roles of different healthcare professionals, the objectives of each care phase, current assisting methods, and the challenges encountered. Additionally, we delved into how healthcare providers gather and utilize information throughout the diabetes care journey, their engagement strategies, and the challenges they face in this process.

Part II: Scenario with Envisioned AI system

After concluding the first part, we presented the scenario (detail in Subsection 4.4.2). This part introduced several AI functionalities, stimulating a discussion centered on their importance, benefits, expectations, and concerns from the perspectives of both healthcare professionals and patients. The discussion encouraged participants to critically evaluate each functionality and its potential impact.

4.4.4 Data Collection and Analysis

During the focus group session, participants were not only engaged in discussions but also wrote their inputs on post-its. To complement this, a dedicated note-taker was assigned to record the participants' verbal contributions.

After the session concluded, the written inputs on the post-its were digitized to streamline the analysis process. All raw data sources, including the participants' written inputs and the notes taken during the session, were aggregated in a single document, which is presented in Appendix C. This structure of the aggregated document was in line with the setup of the focus group: Current Diabetes Care and Scenario with Envisioned AI.

For the analysis, a color-coding system was applied to the text to highlight and differentiate themes. This facilitated the grouping of responses by similar themes, streamlining the organization of insights. The insights derived from this thematic organization are further explored and discussed in the following section.

4.5 Results

The following sections discuss the results from Part I (Subsection 4.5.1) and Part II (Subsection 4.5.2) of the focus group study.

4.5.1 Part I: Insights into Current Practices in Diabetes Care

The first part of the focus group addressed the current state of T2D care. The discussion revolved around the different phases of T2D care, the nature of patient consultations, the challenges in practice, and the opportunities for conversational AI.

Phases of Diabetes Care

Participants referred to established frameworks and guidelines prevalent in T2D care, citing sources from the Dutch General Practitioners Association (NHG)[3]. Broadly, the T2D care trajectory encompasses three phases: Diagnosis, Initial Treatment, and Chronic Treatment.

The participants explained that the patient journey of a patient with T2D usually begins with a Diagnosis, primarily made during consultations with a General Practitioner (GP). Treatment strategies vary based on the severity and stage of diabetes, ranging from lifestyle adjustments to immediate medication. The participants stressed that lifestyle adjustments are highly effective for preventing and treating diabetes. However, despite the effectiveness of lifestyle interventions, the participants acknowledged that medication is often prescribed due to the numerous barriers associated with lifestyle changes, which can be overwhelming for healthcare providers to address comprehensively during consultations. During the Initial Treatment phase, the primary aim is to raise awareness, potentially prevent comorbidities, and delay medication whenever feasible. In the Chronic Treatment phase, the emphasis lies on avoiding the escalation of medication and maintaining optimal blood glucose and HbA1c levels (average blood glucose (sugar) levels for the last two to three months).

Dynamics of GP-Patient Conversations

The dialogue between GPs and T2D patients typically covers a range of topics, including the patient's goals, their health literacy or knowledge, available social support, and their perspectives on their current lifestyle and its potential changes. However, patients' values are often left unexplored, despite the importance of patients' values in the effectiveness of lifestyle interventions. This is primarily due to time constraints and a lack of expertise in probing this area. Moreover, patients might not be adequately prepared or equipped to engage in this kind of deeper-level conversation. Typically, healthcare providers initiate conversation at a surface level and probe deeper in subsequent consultations if initial approaches are insufficient.

Identified Challenges

Several challenges were highlighted during the focus group. One significant challenge is the limited time allocated per patient encounter. GPs typically meet their T2D patients only three to four times a year, each session lasting approximately 15 minutes. This confirms the potential for leveraging conversational AI as an intermediary for patients and healthcare providers, especially between scheduled consultations. Additionally, a common problem is that it is often too late when patients are diagnosed. Furthermore, patients could be inconsistent in keeping up with follow-up appointments, causing delays in their care. Another hurdle is the inherent difficulty some patients face when meeting with a doctor, whether due to logistical, personal, mental, or socio-economic reasons.

[3]https://www.nhg.org/

Opportunities for Conversational AI

These aforementioned factors highlight the potential of conversational AI by facilitating earlier diagnosis, encouraging regular follow-up, offering comfort when meeting with doctors, and providing suggestions for continuous patient engagement. In addition, two key observations surfaced. Firstly, there is no standardized tool used by healthcare providers for patient conversations. This poses challenges in creating conversational agents that can effectively model these interactions but it also opens up a significant research opportunity to develop conversational AI solutions tailored to this need. Secondly, it was noted that many healthcare providers, such as GPs, often have limited time to focus on discussing lifestyle changes with their patients, even if it would be beneficial. This highlights another potential area where the *intermediary* role of conversational AI could be particularly helpful, offering continuous support for lifestyle management beyond regular consultations.

4.5.2 Part II: Experts' Viewpoints on the Envisioned AI Scenario

The second part of the focus group captured the expert perspectives on the proposed AI scenario in T2D management and addressed experts' concerns and ethical considerations about its integration in practice.

Perceptions AI Functionalities

As discussed in Section 4.3, the envisioned conversational AI system comprises several functionalities. These functionalities were discussed with the experts. Considerable attention was specifically focused on the function "Summarize the insights for healthcare providers." Experts indicated that this functionality could be critical. It was suggested that it could present a potential time-saving advantage for healthcare providers. One of the highlighted discussion points emphasized the need for the content of these summaries to be profession-specific. For example, GPs care most about things that directly affect medical choices, such as how well medicine is working or why a patient might not follow advice. Hence, when providing summaries to GPs, given the short time in each consultation, the AI system should focus solely on data influencing these medical decisions.

For the AI system to make these summaries, it needs to "capture reasons why the user is not adhering to their goal behavior." The experts indicated that this functionality would benefit patients by preventing them from repetitive discussions and potentially making them feel better understood. However, the absence of standardized conversational "pathways" between healthcare professionals and patients (as mentioned in Subsection 4.5.1) poses a challenge for AI. This absence inhibits rule-based conversational AI systems, which rely on predetermined pathways. There is a need for more sophisticated AI models that are capable of navigating this kind of diverse, complex, and nuanced dialogue interactions.

The functions "Remind the user to adhere with their healthy behavior goals/medication/mental wellbeing" and "Suggest healthcare providers on possible treatment/strategies" were also briefly discussed. The experts consider the former to be not very novel in nowadays applications but it often falls short due to its inflexibility in adapting to patients' daily lives. The latter raises skeptical attitudes among healthcare professionals because of the complexity of medical problems and ethical concerns such as patient privacy and confidentiality.

Addressing Concerns and Ethical Considerations
Some significant concerns arose in the discussions regarding patient privacy, especially in terms of the information that the AI system shares with GPs. This highlights the need for privacy protocols and guidelines on data sharing. Moreover, experts considered it essential to ensure transparency with patients about the data collection process and its relevance to their T2D management care. Patients need to be well-informed about why their data is being collected, how it will be used, and the benefits this brings to their treatment plan. This transparency was considered crucial in establishing trust and engaging patients with the AI system.

Another critical aspect that was highlighted is the nature of interactions between patients and healthcare professionals, which often begin with a specific topic and unfold from there. It was considered critical to consider this natural flow in building conversational AI that aims to capture patients. Rigid and predefined dialogue structures might hinder information acquisition.

4.6 Summary and Future Work

In this section, we summarize the results of the focus group and provide future research directions.

4.6.1 Summary

The expert focus group has highlighted the potential benefits of conversational AI in the context of T2D care. These agents could play a critical role in facilitating earlier diagnoses, encouraging regular follow-up, and offering comfort or advice regarding going into consultations with healthcare professionals. Additionally, conversational agents might mitigate the discussion of sensitive topics, which are often challenging to address during traditional consultations. However, the experts expected the effectiveness of AI in bridging the gap between healthcare providers and patients to be dependent on the specific types of healthcare providers involved and the phases of the diabetes patients, as each requires a different approach.

Implications for Patients From the patient's perspective, there is currently no standardized tool for capturing the reasons behind non-adherence. Existing techniques, such as goal setting (e.g., [38]), motivational interviews (e.g., [156, 247]), and questionnaires (e.g., [104, 198]) - well established in behavioral change practices - have been implemented in conversational agents. However, they may not fully capture the nuanced, context-specific factors such as values, preferences, norms, and beliefs required in real-world scenarios, as envisioned in our study.

Progress has been made in addressing this gap through agents that elicit and confirm user-specific reasons for behavior. Notably, Abdulrahman et al. [6] explored how agents can elicit a user's specific goals and beliefs to tailor explanations, treating these beliefs as the primary motivations for or barriers against the recommended behavior. The exploratory work by Chen et al. [62] presents an initial step in this direction, investigating dialogues aimed at capturing high-level, situation-dependent concepts like values via conversational agents. However, this study used hypothetical written dialogues rather than real patient interactions. There is a need for future research involving real patient interactions.

Implications for Healthcare Providers It was mentioned that healthcare professionals might prefer to receive concise reports over lengthy reports summarizing the AI's interactions with patients since the last consultation. These reports should focus on areas relevant to driving medical decisions. Additionally, the potential of the 360° diagnostic tool, developed by Harakeh et al. [136], was brought up. This tool provides an overview of critical T2D-related factors, including behavior and environment. It is intended as a decision support tool for T2D patients and GPs, helping them identify and address relevant factors and determine suitable interventions. One potential idea could be to incorporate this tool with the conversational agent. This integration could enable the AI to translate its insights into a format compatible with the diagnostic tool. Future research should investigate which presentation styles, i.e., presentation via a textual report or via the 360° diagnostic tool, are most effective for which types of healthcare professionals and under what circumstances. Understanding healthcare providers' preferences for receiving AI-generated summaries could inform the design of conversational agents, ensuring the questions are structured in a way that facilitates easy translation into these preferred formats.

4.6.2 Future Research Directions

Future research concerning the conversational AI system as an intermediary for patients and healthcare providers can be summarised along two main dimensions.

Research Focused on Patients Research for patients should focus on developing conversational AI agents that can effectively capture the unique contexts of patients and what it is about the current context that is important to them. This includes using dialogues to explore possible situational variables and patients' values, along with how these elements interrelate.

Research Focused on Healthcare Professionals Research for healthcare professionals should investigate the optimal ways to present AI-collected data to healthcare providers. This involves considering the various requirements based on healthcare professionals' specific roles and the treatment stages of T2D patients. This understanding is crucial as it would not only aid in the design of more effective AI system but also potentially enhance the overall efficiency and effectiveness of T2D care.

One step further could be presenting this data back to the patients themselves. This could enhance transparency if the patient could understand how the information collected by the agent is used, possibly leading to improved privacy and trust.

5

Presenting User Behavior Information Collected by a Conversational Agent: Impact of presentation format on comprehension quality and speed

Behavior support systems increasingly use conversational agents to collect information about users' daily contexts and behavioral challenges.While this information is typically processed by the agent to provide direct support, it can also be shared with a human party, such as a lifestyle coach or healthcare provider, to facilitate collaborative care. This raises the question of how best to present the information. This study investigates how three presentation formats—the original dialogue form, a presentation of information in the form of a structured list, and an interactive 360° tool which presents information visually in a circular form-affect comprehension quality and speed of the person aiming to understand this data. We conducted a between-subjects experiment where participants reviewed eight agent-user dialogues in one of three formats: original dialogue, structured list, or interactive 360° tool. We measured comprehension using adapted Situation Awareness Global Assessment Technique (SAGAT). Results showed that original dialogues led to the highest SAGAT Level 1 & 2 accuracy, while structured lists enabled faster responses, For SAGAT Level 3, no presentation formats consistently outperformed the others, highlighting a trade-off between detail and efficiency. Participants' qualitative feedback suggests that while structured lists and 360° tool aid speed, contextual richness as offered by the original dialogue remains important. These

Pei-Yu Chen, M. Birna van Riemsdijk, and Myrthe L. Tielman. "Presenting User Behavior Information Collected by a Conversational Agent." *Under review.*

findings offer early insights for designing presentation formats in behavior support systems, though further research with professional coaches is needed.

5.1 Introduction

Achieving and maintaining a healthy lifestyle is a complex challenge because it is deeply integrated into the routines and choices individuals make each day. Lifestyle coaches provide essential support by working closely with clients to understand their unique circumstances, motivations, and challenges. During an intake session, a coach typically explores the client's health goals, habits, and potential barriers to change [118, 249]. This allows the coach to craft an initial plan tailored to the client's lifestyle and readiness for change. However, after these sessions, clients are often left to manage their goals on their own, with limited guidance until the next check-in [123].

Advances in technology offer promising ways to bridge this gap, particularly through behavior change support systems (BCSS) [237]. These systems are typically known for guiding users toward healthier habits through automated reminders or suggestions. However, their potential extends beyond guidance alone: BCSS can also play a valuable role in monitoring users' behavior and context throughout daily life. In doing so, they can collect rich data about when and why people deviate from their intended health goals, a perspective that is often missing from routine coaching sessions.

To capture not only what people do, but why they do it, conversation is crucial. A promising direction is to use dialogue between the system and the user to explore reasons behind non-adherence. For example, when a user fails to follow through on a goal, the system could initiate a short conversation to understand the reasons, as proposed by Chen et al. [62]. These conversations can uncover nuanced contextual factors, such as work obligations, family responsibilities, or motivational shifts, that cannot easily be captured through passive tracking alone [97, 226]. Through such alignment dialogues, the system may better support lifestyle coaches by offering a more contextualized and person-centered view of the user's situation.

Since lifestyle coaches have limited opportunities to interact with clients directly, the information gathered through these dialogues could serve as an important bridge. Sharing this enriched data with coaches could enable them to better understand their clients' challenges, fostering more productive discussions and tailored interventions [64]. This integration of daily context into coaching strategies may significantly enhance the support provided to clients in achieving their health goals.

While gathering this information is essential, an equally important challenge is how to present it to the lifestyle coach in a way that supports comprehension and decision-making. The format in which data is presented plays a crucial role in how well the coach can interpret and apply these insights. Research in various contexts has shown that the way information is structured significantly influences a person's ability to process and utilize it effectively [35, 160, 223]. One key factor is the amount of information provided: too much data can overwhelm the end user, leading to cognitive overload [61, 186, 199].

A key challenge in behavior support systems is how to effectively present the information they collect. The system gathers data from one individual (the client) and presents it to another person (the lifestyle coach or, in this study, a lay participant). This setup reflects real-world scenarios where lifestyle coaches must interpret information about clients' experiences, challenges, and progress based on data they did not directly gather. How this information is structured and presented plays a crucial role in how well the recipient can comprehend and use it effectively.

Existing research has explored how data presentation formats affect comprehension and decision-making [35, 160, 223]. In fields like eHealth and mHealth, visualizations have been proposed as a way to support human information processing [16, 54, 110]. However, much of this research evaluates visualization designs in isolation, without comparing them to baseline formats such as the original, raw data [29, 43, 131, 145, 318]. This gap raises an important question: How do different presentation formats influence comprehension and usability?

To investigate this, we examine three presentation formats:

- Original dialogue –the unaltered conversation between the agent and the user.
- Structured list –a text-based summary categorizing key points from the dialogue.
- Interactive 360° tool –a visualization tool that combines text and graphical elements to present a holistic overview, inspired by Harakeh et al. [136].

These formats differ in their level of abstraction: from retaining full conversational context (original dialogue) to structured textual summaries (Structured list) to a more visual representation (interactive 360° tool). By comparing these approaches, we aim to answer the following research questions:

1. RQ1: How do different presentation formats influence a person's comprehension of another person's behavioral choices in a coaching context?
2. RQ2: How do these presentation formats affect the speed at which a person understands another person's situation?
3. RQ3: Which presentation format is perceived as most useful in helping a person understand another person's behavioral choices and support their healthy lifestyle changes?

The remainder of this paper is structured as follows. Section 5.2 reviews related work on user modeling, conversational agents, and the presentation of dialogue-derived information. Section 5.3 outlines the experimental setup and the study measures. Section 5.4 presents the results. We discuss the findings and their implications in Section 5.5, as well as the limitations and directions for future research.

5.2 Related Work

5.2.1 User modeling for behavior change support systems

Behavior change is a multi-faceted challenge [298]. To change and support behavior, information is needed not only about physical health factors of a person but also about lifestyle factors, e.g., [24, 260, 331].

To provide effective and personalized support, it is critical to capture the diverse factors influencing an individual's behavior, whether through human assessments or technology-driven approaches. When technologies are employed, this information is stored and managed within a user model, which serves as the foundation for delivering tailored interventions.

This need to consider multiple influences aligns with the concept of *positive health* [145], which highlights that well-being is shaped by a combination of lifestyle-related behaviors (e.g., eating patterns, physical activity, and sedentary behavior), mental health factors (e.g., stress, anxiety, and depression), and socioeconomic conditions (e.g., neighborhood environment) [136, 145]. Recognizing these interconnected factors is crucial for designing behavior change interventions that address individuals' real-world challenges holistically.

User modeling is widely used across various domains, including health applications, to systematically collect and organize user-specific information for personalization. In the context of behavior change support, user models help tailor interventions by adapting to personality traits [196, 325], adjusting recommendations based on users' progression rates [76], and incorporating individual preferences [109]. Another key aspect is capturing users' motivational attitudes, such as values and preferences, which influence engagement with interventions [320]. These models often include both static components (e.g., demographics, preferences) and dynamic components (e.g., daily routines or contextual constraints) [130].

User models can be constructed automatically by inferring information from users' interactions with the system [345, 346], or explicitly through direct user input [119]. While explicit data collection yields self-reported and interpretable information, it may overlook the situational nuances that shape real-life behavior [193, 257]. For instance, traditional questionnaires often rely on fixed categories, limiting their ability to capture users' evolving motivations and barriers. In contrast, dialogue-based approaches, such as alignment dialogues proposed in Chen et al. [62], allow users to articulate their reasoning in their own words, potentially leading to a more contextualized and nuanced user model.

Once collected, this information must be passed from the system to another human, typically a lifestyle coach—who did not directly observe the interaction. This highlights a second challenge: not just collecting information, but presenting it effectively. Prior work in psychology and information systems has shown that the structure and format of presented data strongly affect how well people can understand and act on it [35, 160, 223]. Overly detailed presentations can cause cognitive overload [61, 186, 199], while overly simplified summaries may omit critical context.

In this study, we focus on how to present dialogue-derived user model information to a third party, such as a lifestyle coach. While many studies have examined how to visualize behavior data or present survey results, relatively little is known about how to present conversational data effectively, particularly when the information must balance richness with usability.

5.2.2 Conversational agent for collecting user information

Understanding a user's real-life challenges and decision-making processes requires more than just collecting static data points. To overcome the limitations of these static user modeling approaches, Chen et al. [62] proposed alignment dialogues: a conversational AI approach designed to acquire information about the user's behavior and motivations in the context where it is occurring. This approach aligns with the principles of Ecological Momentary Assessment (EMA) in clinical psychology, which emphasizes gathering real-time self-reports of individuals' behaviors and experiences within their natural environments [294]. Like EMA, alignment dialogues aim to capture information as the situation unfolds, offering more valid and detailed insights into users' real-world behaviors. However, while EMA typically relies on structured self-report prompts, alignment dialogues enable a more interactive and flexible exchange that can uncover the nuanced motivations behind user behavior.

While the concept of collecting information via conversational agents is not new, previous implementations have typically focused on supporting specific tasks, such as booking flights or finding restaurants [340]. In contrast, alignment dialogues aim to elicit high-level concepts such as goals, values, and behavioral barriers, information that is difficult to infer solely from behavioral logs or retrospective reports.

What sets alignment dialogues apart is their focus on capturing high-level concepts such as user values, goals, and preferences. Collecting such information is important because context plays a crucial role in shaping behavior, as motivations, barriers, and external factors shift over time. Conversational agents are particularly promising for capturing this evolving context, as they enable interactive, real-time exploration of user preferences and situational nuances [71].

By engaging in dialogue, the agent seeks to acquire insights into what is significant to the user within their dynamic, ever-changing daily contexts. This makes conversational agents well-suited for behavior support settings, where capturing not only what a user does but also why they make certain choices is essential. Importantly, the information gathered through such interactions is useful for serving as a rich source of contextual data that can be relayed to lifestyle coaches. For this reason, the question of how to present such information effectively to a human coach becomes critical.

5.2.3 Information presentation and comprehension

Since conversational agents collect information in natural language, a key challenge lies in how to present this data in a way that supports comprehension and usability, particularly when another person, such as a lifestyle coach, needs to review the information. Prior research provides several insights into the role of presentation structure, but these findings stem from varied contexts and concern various different types of information.

One area of focus has been on the visual and structural organization of text. For example, Hunter [146] demonstrated that making text structure visually explicit—through cues and organized document layouts—can reduce cognitive burden and improve accessibility. Similarly, Lemarié et al. [189] emphasized that segmented, well-structured text presentations facilitate better comprehension compared to unstructured or linear formats.

Another line of work compares the impact of different text formats on processing efficiency. For example, Wogalter and Shaver [334] found that list-based formats reduced search time compared to paragraphs, suggesting that even simple formatting choices can influence comprehension and decision speed.

In contrast, dialogue-based data presents a unique challenge: it typically needs to be presented sequentially to represent the natural flow of conversation. To address this, Lee and Chen [187] proposed strategies for organizing spoken documents (such as course lectures, movies, news episodes) into concise and coherent representations that support efficient browsing and information retrieval. Extending this, Demberg et al. [86] proposed methods for structuring information according to user preferences, helping users more efficiently navigate complex and diverse information.

While these studies offer useful foundations, they largely address structured documents or spoken content in isolation. A critical gap remains in how to present natural language content collected through dialogue, such as the conversations between a behavior support agent and a user. Unlike questionnaire responses or system logs, such dialogue data is often richer, but also more ambiguous or fragmented. As a result, the design of presentation formats for this kind of data is both especially important and underexplored.

5.3 User study

We conducted a user experiment to investigate how different formats for presenting dialogue content collected by a conversational agent influence participants' understanding of another person's situation. We also examined how these formats affect comprehension time.

5.3.1 Participants

Participants were recruited from the Prolific crowdsourcing platform. The participants received monetary compensation according to the platform policy. The study was approved by the Human Research Ethics Committee of of Delft University of Technology (ID nr 4388). A total of 94 participants were included in the analysis. 45 of individuals were below 35 years old, 48 were between 35 and 54 years old, and one 55 years old or older. Educational backgrounds varied, with participants holding different levels of education (high school 31, Bachelor 46, Master 14, PhD 3).

5.3.2 Experiment Setup

The study employed a 2 (**Information Volume**: Long, Short) × 3 (**Format**: Original dialogue, Structured list, Interactive 360° Tool) mixed factorial design. The Format variable was manipulated between subjects, such that each participant was randomly assigned to experience one of three presentation formats: Original dialogue, Structured list, or Interactive 360° tool. In contrast, the Information Volume variable was manipulated within subjects, with all participants experiencing both Short and Long versions of the story across different rounds.

The experiment consisted of eight distinct scenarios. Each scenario depicted a situation in which a user failed to follow through on a healthy behavior goal, such as skipping exercise due to work stress or eating unhealthily during social events. In each case, a conversational agent engaged with the user to understand the underlying reasons for non-adherence. These agent-user dialogues formed the basis for the content that was then presented to participants in one of the three formats.

For any given scenario, participants saw either the Short or the Long version, not both. The order of scenarios was randomized for each participant to avoid systematic bias. The randomization and counterbalancing ensured that each scenario appeared an equal number of times in both volume conditions across the entire participant pool.

In summary, the following independent variables were used:

1. Presentation Formats (between-subjects)
   (a) Original Dialogue: Written-out dialogue interaction between the agent and the patient.
   (b) Structured List: Structured list summarizing reasons for non-adherence under predefined categories.
   (c) Interactive 360° Tool: Visual representation of the reasons for non-adherence, organized into five categories in the interactive 360° tool format.
2. Information Volume (within-subject)
   (a) High: Around 250 words in the original dialogue.
   (b) Low: Around 125 words in the original dialogue.

5.3.3 Presentation Format Design

As explained, in this study, we explored three distinct formats for presenting information collected during a conversation between a behavior support agent and a client to a lifestyle coach: (1) Original dialogue, (2) Structured list, and (3) Interactive 360° tool. Below, we describe the design and purpose of each format.

The content of the dialogues was written by the authors and inspired by findings from Chen et al. [65], which collected data on common healthy lifestyle goals and the circumstances under which people fail to comply with those goals through human-agent conversations. The topics covered in the eight scenarios reflect frequently mentioned challenges from that study, such as stress from work, social obligations, or low energy levels. Each dialogue followed a consistent structure: it began with a check-in or notification from the agent, followed by the client disclosing that they had not followed their goal. The agent then inquired about the reasons for non-adherence, to which the client responded, and the agent continued with follow-up questions to elicit more contextual detail. This structure was designed to resemble realistic support dialogues and to ensure consistency across scenarios.

Original dialogue

The original dialogue format presents a written-out conversation between a support agent and a client. It preserves the full back-and-forth exchange, reflecting how people discuss challenges related to behavior change.

As shown in Figure 5.1, participants receive the complete dialogue text, allowing them to read the interaction as it could unfold between an agent and a client. This format maintains the conversational flow and retains contextual cues that may help in understanding the reasoning behind the client's behavior.

Structured List

The structured list format extracts and categorizes key information from the dialogue, with the intention to make it easier for a person to quickly grasp another person's challenges. The categories are based on [65] which categorized reasons for non-adherence to healthy lifestyle behaviors. The ones used in this experiment include:

Agent:	Have you been able to keep up with your exercise routine?
Patient:	Not really. I've been working late a lot recently.
Agent:	How do late work hours impact your exercise routine?
Patient:	By the time I get home, I'm exhausted. I just don't have the energy to work out.
Agent:	What do you usually do instead of exercising when you get home late?
Patient:	I usually just collapse on the couch and watch TV to unwind.
Agent:	Why do you choose to relax with TV instead of exercising?
Patient:	It's the easiest way to switch off after a long day. Exercise feels like too much effort when I'm already so tired.
Agent:	Have you tried any strategies to fit in exercise despite your late hours?
Patient:	I've thought about shorter, less intense workouts, but I haven't been able to stick with them consistently.

Figure 5.1: Original dialogue example.

- Individual challenges: intrinsic factors that are inherent to the individual and are difficult to ask of them to change.
- Societal challenges and environment: external factors that shape user behavior.
- Contextual/social aspects: elements within a situation that can influence, shape, or affect your choice. Social aspects are specifically related to social interactions, relationships, and social dynamics.
- Values: A sense of accomplishment, Security, Excitement & fun, Being well-respected & Self-respect, and Sense of belonging & warm relationship.
- Other motivational attitudes: Preference & prioritization, Norms, and Belief.

These categories are presented as a list where the explanations for their behavior are grouped under the relevant predefined categories. Each category is labeled in bold, followed by a concise sentence summarizing the essence of the explanation, as shown in Figure 5.2. This design aims to improve information accessibility and reduce the cognitive load associated with reading a full dialogue.

- **Capability:** I've thought about shorter, less intense workouts, but I haven't been able to stick with them consistently.
- **Mental state:** By the time I get home, I'm exhausted; Exercise feels like too much effort when I'm already so tired.
- **Work-Related Constraints:** I've been working late a lot recently.
- **Preference & Prioritization:** I usually just collapse on the couch and watch TV to unwind; It's the easiest way to switch off after a long day.

Figure 5.2: Structured list example. The scenario is the same as in Figure 5.1.

Interactive 360° Tool

The Interactive 360° Tool provides an interactive visual circle-shaped representation of the information elicited from the user by the agent. This format is adapted from a diagnostic tool originally developed to support conversations between patients with type 2 diabetes and their healthcare providers [64].

In our version, the Interactive 360° tool represents general healthy lifestyle challenges which are divided into five slices, each representing one of the categories from the structured list format. Within each slice, icons represent specific reasons for non-adherence. When participants hover over an icon, a tooltip displays what the icon represents. Two icon colors are used: beige and purple. Purple icons indicate that the content contains new or particularly relevant information from the original dialogue. When clicked, a pop-up window appears in the center of the interactive 360° tool, displaying a short sentence summarizing the client's explanation, as shown in Figure 5.3. This sentence is identical to the one used in the structured list format.

5.3.4 Measures

To evaluate the effectiveness and usability of the different presentation formats, we collected a range of quantitative and qualitative measures. These included both predefined dependent variables related to comprehension and usability, as well as exploratory variables and open-ended feedback.

Control/Exploratory variables

We use these variables to describe our sample and for exploratory analyses, but we do not conduct any conclusive hypothesis tests on them.

1. Age group (categorical)
2. Gender (categorical)
3. Level of education (categorical)
4. Coaching knowledge (continuous)

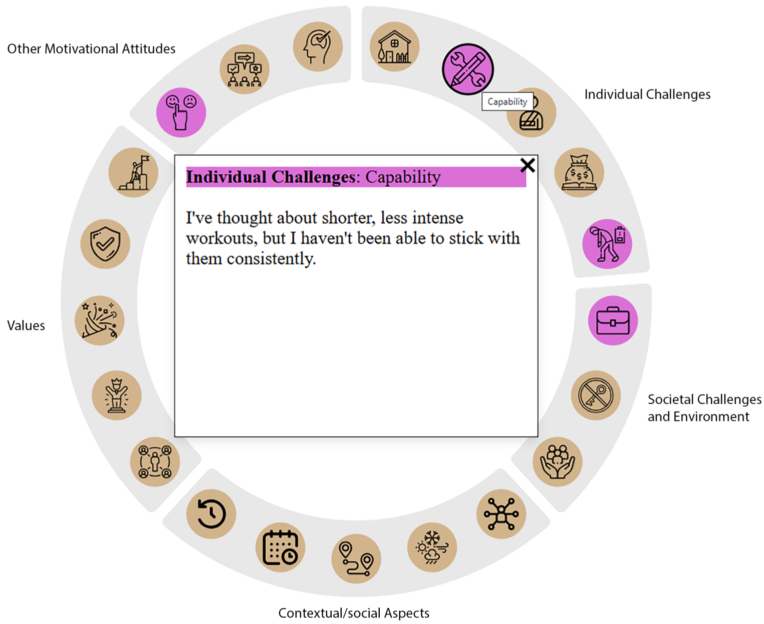


Figure 5.3: Interactive 360° Tool example. This visualization corresponds to the same scenario shown in Figure 5.1. The name of each icon is revealed when the cursor hovers over it.

5. Healthy lifestyle knowledge (continuous)

Dependent variables

1. **Accuracy of comprehension questions** (after each scenario): We measured participants' comprehension using the Situation Awareness Global Assessment Technique (SAGAT) [106]. Originally developed to assess operators' awareness in complex, real-time systems, SAGAT is not a fixed questionnaire but rather a framework for designing questions that probe different levels of situational awareness. Although our study context differs, SAGAT remains relevant as a framework for creating evaluating how well someone understands a situation, which is central to the goals of our study.

   In our adaptation, we combined SAGAT Level 1 (perception of elements) and Level 2 (comprehension of meaning) into a single question that assessed participants' understanding of the dialogue content and the reasons for non-adherence. We retained the original Level 3 questions (projection of the near future), which assessed participants' ability to predict what the client would likely do in a similar future situation. While the answer options were based on information provided in the scenario, the questions were phrased hypothetically to require inferential reasoning. The "future" scenario presented in the question maintained the same contextual conditions, so that the expected behavior could be logically inferred from the original dialogue.

Accuracy was assessed by comparing participants' answers to the pre-established correct choice based on the dialogue content.

Each scenario included two comprehension questions: (1) One that combines Levels 1 & 2; (2) One for Level 3. These questions evaluated how well participants understood key elements of the scenario as presented in the different formats. To ensure meaningful comprehension rather than superficial skimming, all multiple-choice options included keywords or phrases drawn from the dialogue presentation formats. This was done deliberately so that simply spotting familiar terms would not help identify the correct answer. Since all answer choices contained such cues. As a result, participants needed to read and interpret the content carefully to select the correct response.

Below are sample comprehension questions based on the same scenario shown in Figure 5.1:

(a) SAGAT Level 1 & 2 (Combined): What are the main reasons the client finds it difficult to keep up with their exercise routine?
   i. Late work hours, feeling exhausted, and choosing to relax with TV instead of exercising.
   ii. Lack of interest in exercise, no suitable place to work out, and a preference for other activities.
   iii. Health issues that prevent exercising, dislike for physical activity, and no motivation to work out.

(b) SAGAT level 3: If the client comes home late from work and feels exhausted, what are they most likely to do?
   i. Find a way to exercise every day despite the exhaustion.
   ii. Collapse on the couch and watch TV to unwind, avoiding exercise.
   iii. Start exercising more intensively to compensate for the missed workouts.

2. **Response time** (after each scenario): We recorded participants' response time for each SAGAT comprehension question. This allowed us to assess not only comprehension accuracy but also how quickly participants processed the presented information (in cases where their answers were correct) across different formats and scenario lengths.

3. **Usability** (after each scenario): To assess the perceived usability of the presentation formats, participants were asked the following question after each scenario. Responses were given on a 7-point Likert scale ranging from 1 (Extremely unlikely) to 7 (Extremely likely):

   - Imagine your job is to advise people about their healthy lifestyles. How likely would this dialogue/list/360° Tool enhance your effectiveness on the job?

4. **Representability** (after each scenario): For the Structured List and 360° Tool formats, participants were shown the original dialogue after completing the SAGAT and usability questions. They were then asked to evaluate how well the alternative

5

presentation represented the original content. Responses were given on a 7-point Likert scale ranging from 1 (Not well at all) to 7 (Extremely well):

- How well does the list/360° Tool represent the original dialogue content?

5. **Technology Acceptance Model (TAM) Questionnaire** (after all scenarios): After completing all eight scenarios, participants filled out the TAM questionnaire, which measures user acceptance of technology through two key constructs: Perceived Usefulness and Perceived Ease of Use.

   - Perceived Usefulness reflects the extent to which participants believe that using the presented format (dialogue, list, or 360° tool) would enhance their ability (imagined role as a coach) to advise clients effectively.
   - Perceived Ease of Use evaluates how effortless and straightforward participants find it to interact with the format. We adopted the TAM questionnaire and made slight modifications to tailor it to the context of our study.

   We adapted the original TAM questionnaire to better fit the context of this study. Specifically, participants were asked to "Imagine your task is to advise your client about their lifestyles...", and the phrase "this product" was replaced with "this dialogue/list/360° Tool" to align with the format they experienced.

5

Open Questions
To gather qualitative feedback, participants were asked several open-ended exploratory questions at the end of the experiment. These questions encouraged reflection on the usability and preferences for the different presentation formats. Specifically, participants were asked:

- Do you have any suggestions or improvements for the original dialogue/the list/the 360° tool between the conversational agent and a user?
- In which situations would you like or dislike to use it if you were a lifestyle coach?

Participants in the Structured List and 360° Tool conditions were asked three additional questions:

- Would you rather have the list/the 360° tool or the original dialogue to enhance your effectiveness on the job? (forced choice question between the list/360° tool and dialogue)
- Why would you prefer it?
- In which situation would you prefer one over the other?

5.3.5 Procedure

The experiment was carried out on Qualtrics which the participants accessed via Prolific. The participants first provided their demographic information regarding their age, gender, and education level. Additionally, they answered how knowledgeable they are with coaching and how knowledgeable they are about having a healthy lifestyle on a scale of 1 to 7. Participants were randomly assigned to a presentation format condition. Subsequently, they were shown a tutorial trial where the context of the dialogue was explained to them. For example:

> *In previous interactions, the user expressed a goal to drink enough water throughout the day. However, during a subsequent check-in, the conversational agent discovers that the user has not been following through with this goal. To understand the reasons behind this, the agent engages in a detailed conversation with the user. The content of the dialogue is presented in the following format. Your main objective is to gain an understanding of the situation why the user didn't perform their goal.*

For the Dialogue condition, the participants were shown an example of what the dialogue could look like. For the Structured list condition, the participants were shown an example of the list corresponding to the dialogue in the Dialogue condition. For the interactive 360° tool condition, the participants were guided through step by step how to interact with the 360° tool.

After the tutorial trial, the participants had one practice trial before the official trials started. In the practice trial, in addition to the dialogue presentation, they also needed to answer several questions as in the real experiment, which are detailed in Subsection 5.3.4. After the practice trial, the participants had 8 official trials, which were randomized. Each trial is about a different scenario where the user deviates from their goal, and the agent has a dialogue with the user. Depending on the condition, the participants were presented with the dialogue itself, the list, or the interactive 360° tool.

5.4 Results

Our analysis consisted of both descriptive statistics and regression modeling. Descriptive analyses were used to provide an overview of the data, examining means, distributions, and differences across groups. Regression analyses were then used to formally address our research questions. Outliers were removed on a per-scenario basis if response times were excessively long (> 2.5 standard deviation (SD)) or short (< 2.5 SD). We then performed descriptive analysis, calculating the mean and SD for each dependent variable to provide an overview of the data. To account for variability both within and between participants and scenarios, we employed a mixed-effects model instead of simple regression.

Mixed-effects models allow us to include both fixed effects (e.g., presentation format, scenario information volume) and random effects (e.g., participant ID, scenario ID), making them well-suited for repeated measures designs [333]. This approach accounts for the nested structure of our data, where each participant evaluates multiple scenarios, by controlling for individual differences and scenario-level variation. As a result, mixed-effects models provide more accurate and generalizable estimates of condition effects across the sample.

We began with the simplest model for the random effects, including random intercepts for each participant. This approach allows each participant to have their baseline level of dependent variables. We then added random intercepts for each scenario, allowing each scenario to have its own baseline as well. Next, we tested more complex random effects, i.e., random intercept and slope, which account for differences in how participants respond to different information volume levels and how different scenarios respond to changes in information volume.

To determine whether adding these more complex random effects improved the model, we used likelihood ratio tests. These tests compare the fit of simpler models to more complex ones and assess whether the additional random effects contribute significantly to model improvement. Once the optimal random effects structure was established, we proceeded to model the fixed effects, namely, the independent variables (Subsection 5.3.2) and control/exploratory variables (Subsection 5.3.4). We employed a forward stepwise procedure to select the fixed effects.

5.4.1 Accuracy of SAGAT questions

5

To answer RQ1, "How do different presentation formats influence a person's comprehension of another person's situation?", we calculated the averages and standard deviations of the accuracy (in percentage correctly answered) of answers to the SAGAT questions in different presentation formats with different Information Volume, which are shown in Table 5.1.

Table 5.1: Accuracy (mean and SD) of Situation Awareness Global Assessment Technique (SAGAT) questions in different Presentation Formats and Information Volume.

	Original dialogue		Structured list		Interactive 360° tool	
	Short	Long	Short	Long	Short	Long
SAGAT Level 1 & 2	0.96 (0.20)	0.98 (0.13)	0.95 (0.22)	0.90 (0.30)	0.87 (0.34)	0.94 (0.23)
SAGAT Level 3	0.46 (0.50)	0.62 (0.49)	0.49 (0.50)	0.62 (0.49)	0.48 (0.50)	0.59 (0.49)

SAGAT Level 1 & 2 Accuracy We applied a mixed-effects model to examine what variables contributed to the accuracy of answers to the SAGAT level 1 & 2 questions. The final model indicates that Presentation Format and Healthy Lifestyle Knowledge were significant predictors (fixed effects) and the model included random intercepts for both Participant and Scenario. Specifically:

1. Participants in the Structured List condition had a significantly lower SAGAT Level 1 & 2 accuracy compared to the Original Dialogue condition ($\beta = -3.22, SE = 1.19, p < .001$).

2. Participants in the 360° Tool condition exhibited lower SAGAT Level 1 & 2 Accuracy compared to the original dialogue condition ($\beta = -3.93, SE = 1.25, p < .001$).

3. Higher Healthy Lifestyle Knowledge scores were associated with significantly lower SAGAT Level 1 & 2 Accuracy ($\beta = -1.49, SE = 0.54, p < .001$).

Figure 5.4 illustrates the average SAGAT 1 & 2 accuracy in each presentation format.

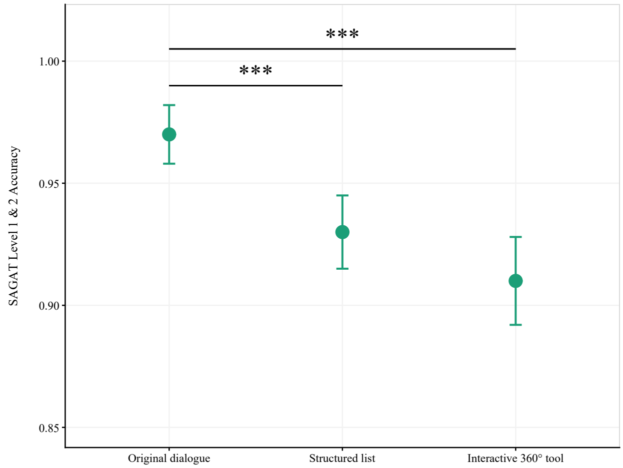


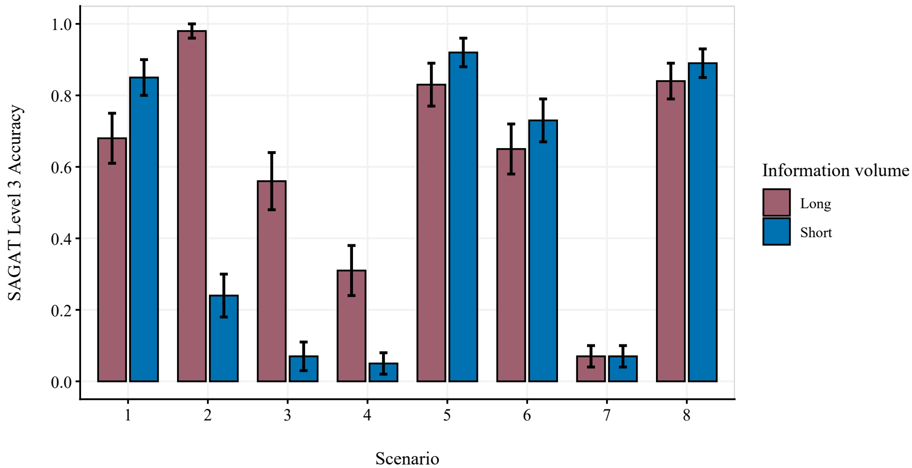
Figure 5.4: Plot of SAGAT 1 & 2 accuracy by presentation format with error bars.

SAGAT Level 3 Accuracy We applied the same analysis method and procedure on level 3 accuracy. Initially, the final model included Healthy Lifestyle Knowledge and Coaching Knowledge as fixed effects and random intercept for each Scenario and random slope of Information Volume for each Scenario. Upon further visualization (Figure 5.5), we observed considerable variability in accuracy across scenarios, as well as variability in the differences between Short and Long Information Volumes across Scenarios. Moreover, the direction of the difference is inconsistent, with Short Information Volumes outperforming Long Information Volumes in some scenarios, while it is *reverse* in other scenarios.

This led us to conjecture that possibly the fixed effect of Information Volume was averaged out and therefore not significant in the initial model. To further investigate whether the effect of Information Volume was indeed averaged out due to variability across scenarios or simply non-existent, we simplified the random effect structure to include only a random intercept for Scenarios. With this reduced complexity in the random effects, the new model revealed Information Volume and Coaching Knowledge as significant fixed effects. Specifically:

1. Participants in the Short Information Volume had a significantly lower SAGAT level 3 accuracy compared to Long Information Volume ($\beta = -0.89, SE = 0.19, p < .001$).
2. Coaching Knowledge had a significant negative effect ($\beta = -0.14, SE = 0.058, p < .005$), suggesting that higher self-reported coaching knowledge was associated with slightly lower accuracy.

Figure 5.6 illustrates the average SAGAT level 3 accuracy in short and long Information volumes.



5

Figure 5.5: Mean SAGAT level 3 Accuracy by Scenario and Information volume, showing variability across scenarios and inconsistent differences between Short and Long information volumes. Error bars represent standard error.

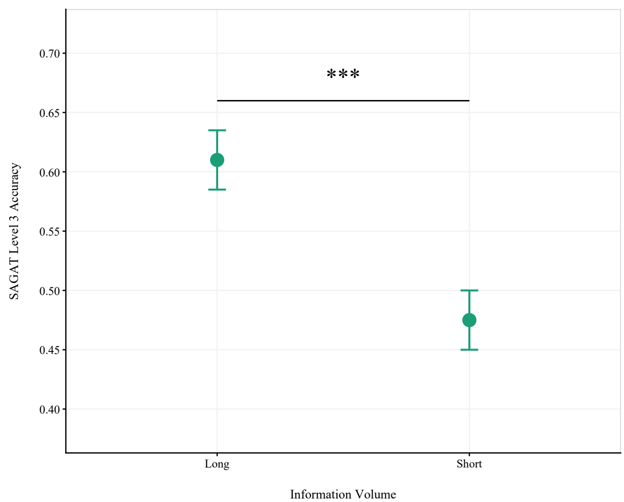


Figure 5.6: Plot of SAGAT level 3 accuracy by Information Volume with error bars.

5.4.2 Response Time of SAGAT questions

We calculated the averages and standard deviations of the response time that participants needed to answer the questions, in order to answer RQ2 "How do these presentation formats affect the speed at which a person understands another person's situation?" In this analysis, only the correctly answered questions were considered. The results are shown in Table 5.2.

Table 5.2: Response time (mean and SD, in seconds) of SAGAT questions in different Presentation Formats and Information Volumes. *n* indicates the number of correctly answered questions in that condition.

	Original Dialogue		Structured List		Interactive 360° Tool	
	Short	Long	Short	Long	Short	Long
SAGAT Level 1 & 2	n=119 48.33 (36.06)	n=119 70.13 (46.40)	n=110 36.36 (26.83)	n=118 43.81 (29.86)	n=115 46.02 (33.36)	n=106 59.60 (36.53)
SAGAT Level 3	n=75 17.80 (14.18)	n=57 24.50 (18.99)	n=76 17.46 (9.45)	n=61 20.92 (10.74)	n=72 22.44 (16.34)	n=58 26.30 (23.34)

SAGAT Level 1 & 2 Response Time A linear mixed-effects model was fitted to predict SAGAT Level 1 & 2 log-transformed response time based on Presentation Formats, Information Volume, along with the interaction between them and Usability as fixed effects. Random intercepts were included for Participants, and random slopes of Information Volume for each Scenario. Specifically:

1. Participants in the Structured List condition had a significantly lower log-transformed SAGAT Level 1 & 2 response time compared to the Original Dialogue condition ($\beta = -0.47, SE = 0.13, p < .001$).

2. Participants in the 360° Tool condition did not show a significant difference in log-transformed SAGAT Level 1 & 2 response time compared to the Original Dialogue condition.

3. Participants in the Short Information Volume condition had significantly lower log-transformed SAGAT Level 1 & 2 response times compared to the Long Information Volume condition ($\beta = -0.47, SE = 0.05, p < .001$).

4. Usability had a significant negative correlation with log-transformed SAGAT Level 1 & 2 response time ($\beta = -0.04, SE = 0.01, p < .05$), indicating that higher usability ratings were associated with shorter response times.

5. The interaction between the Structured List condition and the Short Information Volume condition was significant ($\beta = 0.26, SE = 0.07, p < .001$), indicating that the structured list amplifies response time for short information volume compared to the original dialogue.

6. The interaction between the 360° Tool and Short Information Volume condition was significant ($\beta = 0.23, SE = 0.06, p < .001$), suggesting that the 360° Tool moderates the typical reduction in response time for Short Information Volume observed in the original dialogue condition.

Figure 5.7 shows the mean SAGAT level 1 & 2 response time by condition (Original Dialogue, Structured List, 360° Tool) and Information Volume (Short, Long), with standard error bars indicating variability within each group.

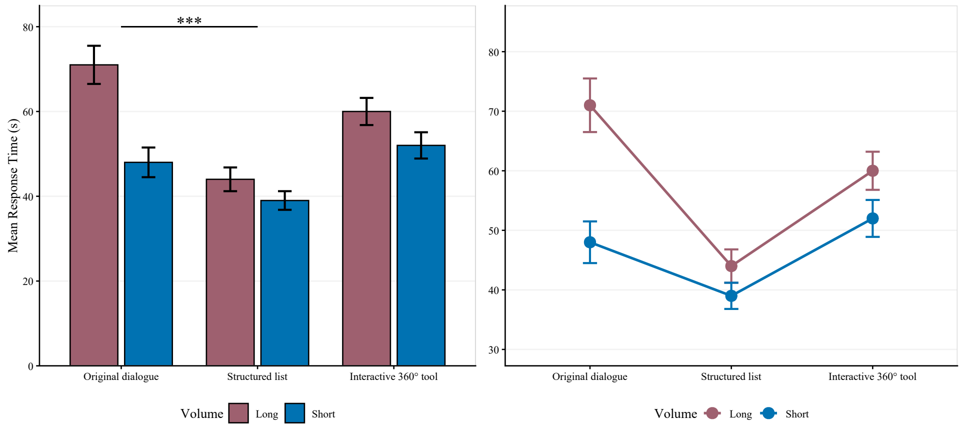


Figure 5.7: Mean SAGAT Level 1 & 2 response time by Condition and Information Volume with standard error bars.

SAGAT Level 3 Response Time The model for SAGAT Level 3 log-transformed response time included Information Volume as fixed effects, along with random intercepts for both Participants and Scenarios. Generally speaking, SAGAT Level 3 response time was much shorter than Level 1 & 2. This could be because the participants already know the scenario from answering SAGAT Level 1 & 2. Specifically:

1. Participants in the Short Information Volume condition had significantly lower log-transformed SAGAT Level 3 response times compared to the Long Information Volume condition ($\beta = -0.18, SE = 0.05, p < .001$).

Figure 5.8 illustrates the mean SAGAT Level 3 response time for Short and Long Volumes, with error bars representing the standard error for each group.

5.4.3 Usefulness

To answer RQ3 "Which presentation format is perceived as most useful in helping a person understand another person's behavioral choices and support their healthy lifestyle changes", we looked at three measures, namely Usability, Representability, and the Technology Acceptance Model (TAM) Questionnaire.

Table 5.3: Usability and Representability (on a scale of 1-7) in different Presentation Formats and Information Volumes.

	Original Dialogue		Structured List		Interactive 360° Tool	
	Short	Long	Short	Long	Short	Long
Usability	4.77 (1.77)	5.55 (1.33)	5.37 (1.39)	5.52 (1.41)	5.88 (1.10)	5.87 (1.11)
Representability	-	-	5.86 (1.28)	5.78 (1.25)	6.20 (0.92)	6.00 (0.99)

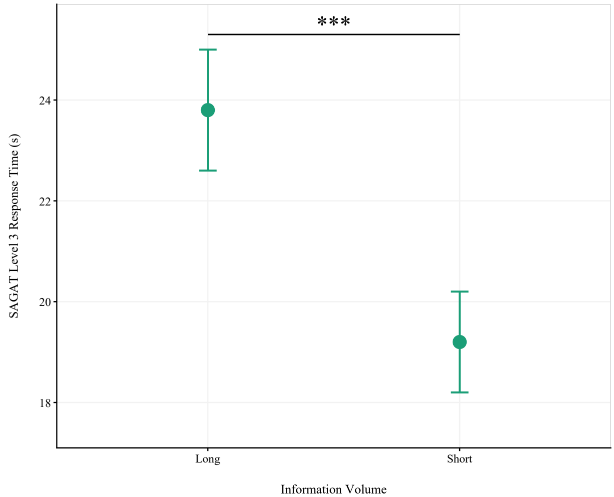


Figure 5.8: Mean SAGAT level 3 response time by Information Volume (Short vs. Long) with standard error bars. We did not illustrate the Presentation Format variable because the effect was not significant.

5

Usability Usability and Representability's means and standard deviations in different Presentation Formats and Information Volume are presented in Table 5.3. The model for Usability included Presentation Formats, Information Volume, and Health Knowledge as fixed effects, with random intercept for both Participants and Scenarios, and random slopes of Information Volume for each Participant and Scenario. We detail the fixed effects below:

1. Participants in the Structured List condition did not significantly differ in perceived usability compared to the Original Dialogue condition.
2. Participants in the 360° Tool condition reported higher perceived usability compared to the Original Dialogue condition ($\beta = 0.40, SE = 0.23, p < .1$)
3. Short Information Volum was associated with significantly lower usability scores compared to Long Information Volume ($\beta = -0.78, SE = 0.18, p < .001$).
4. Health Knowledge was significantly positively associated with perceived usability ($\beta = 0.63, SE = 0.21, p < .01$).
5. There was a significant interaction between Structured List and Short Information Volume ($\beta = 0.63, SE = 0.21, p < .01$), suggesting that the Structured List Format helps mitigate the negative impact of Short Information Volume that is typically observed in the Original Dialogue condition.
6. There was a significant *interaction* between 360° Tool and Short Information Volume ($\beta = 0.63, SE = 0.21, p < .01$). This means that the 360° Tool moderates the typical negative effect of Short Information Volume observed in the Original Dialogue condition, making the Short Information Volume appear more usable.

Figure 5.9 shows the mean usability ratings across the three presentation formats for Long and Short Information Volumes and the interactions between the Formats and Information Volumes.

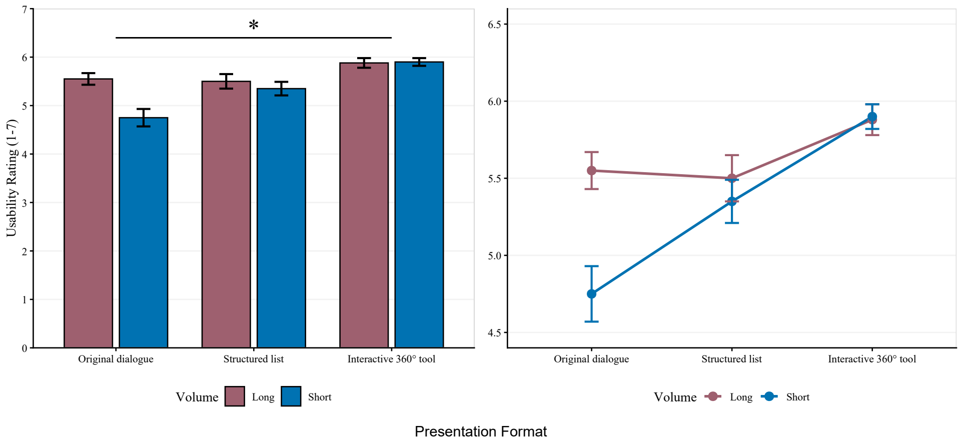


Figure 5.9: Mean Usability scores by Presentation Formats and Information Volume. Error bars represent standard errors of the mean.

Representability Representability, which assesses how well participants perceive the Structured List or 360° Tool as representing the Original Dialogue (note: hence no data is available for the Original Dialogue condition), showed no significant main effects of presentation format, Information Volume, Health Knowledge, or Coaching Knowledge in the final model. This indicates that representability was not significantly influenced by any of these variables.

Technology Acceptance Model (TAM) Questionnaire We calculated the averages and standard deviations of the two constructs Perceived Usefulness and Perceived Ease of Use under TAM. The results are shown in Table 5.4.

Table 5.4: Technology Acceptance Model (TAM) Questionnaire (on a scale of 1-7) in different Presentation Formats.

	Original Dialogue	Structured List	Interactive 360° Tool
TAM - Perceived Usefulness	5.63 (1.28)	5.98 (0.74)	6.15 (0.66)
TAM - Perceived Ease of Use	5.77 (0.99)	5.88 (0.78)	6.19 (0.90)

- Perceived Usefulness: Since TAM was measured only once at the end of all scenarios, mixed-effects models were not applicable. Instead, we fitted simple regressions. The final model for Perceived Usefulness shows a significant positive effect of the 360° Tool condition compared to the Original Dialogue condition ($\beta = 0.59, SE = 0.23, p < .05$). The Structured list condition showed a trend toward being perceived as more useful compared to the Original Dialogue condition ($\beta = 0.45, SE = 0.23, p = .05$). Additionally, Health Knowledge had a significant positive effect ($\beta = 0.33, SE = 0.11, p < .01$), indicating that participants with higher self-reported Health Knowledge perceived the tools as more useful across all Formats. Figure 5.10 displays the average Perceived Usefulness across the three Presentation Formats.

- Perceived Ease of Use: Similarly, the model for Perceived Ease of Use shows a significant positive effect of the 360° Tool condition compared to the Original Dialogue condition ($\beta = 0.47, SE = 0.22, p < .05$). However, no significant difference was found between the Structured List condition and the Original Dialogue condition. Health Knowledge was also a significant positive predictor ($\beta = 0.24, SE = 0.10, p < .05$), suggesting that participants with higher self-reported Health Knowledge perceived the tools as easier to use. Figure 5.10 illustrates the average Perceived Ease of Use across the three Presentation Formats.

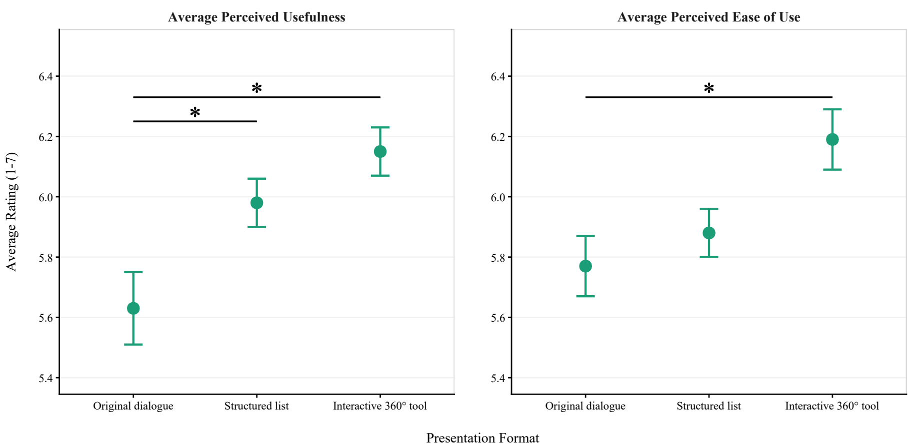


Figure 5.10: Average Perceived Usefulness and Perceived Ease of Use by Presentation Format, with error bars representing the standard error from the mean.

5.5 Discussion

This study explored how different Presentation Formats for presenting dialogue content collected by a conversational agent affect a person's comprehension and perceived usefulness, particularly in the context of supporting healthy behavior. In the following sections, we interpret the results in light of our research questions, reflect on patterns observed across conditions, and consider implications for the design of behavior support systems. We also discuss feedback from open-ended responses and conclude with limitations and directions for future research.

5.5.1 Effect of presentation formats

Our results indicate that Presentation Format has a significant influence on the accuracy and response time of situational awareness (SAGAT), but the effects varied.

For SAGAT Level 1 & 2 (perception and comprehension), the Original Dialogue condition outperformed both the Structured List and 360° tool conditions, while no significant difference was found between the Structured List and 360° tool conditions. This finding aligns with expectations, as the Original Dialogue format provides the most detail, making it easier for participants to recall specific information. Nonetheless, all formats on average achieved high accuracy (≥ 0.9), indicating that the answers were indeed contained in the abstract and shortened versions for basic comprehension.

For SAGAT Level 3 (future projection), no significant main effects of Presentation Format were observed. This suggests that no single format consistently leads to better comprehension at this higher reasoning level. Instead, the effectiveness of a presentation format may depend on the complexity and nature of the dialogue content itself.

Regarding response speed for situational awareness (SAGAT Level 1 & 2), the Original Dialogue condition was significantly slower than the Structured List condition. This is expected, as reading a full dialogue exchange takes more time than scanning a structured summary. Additionally, the response speed for the 360° Tool condition was significantly slower than the Structured List condition. This could be due to the interactive nature of the 360° tool format, which requires users to engage with the visualization rather than viewing all information at once. Interestingly, no significant difference was found between the Original Dialogue and the Interactive 360° Tool conditions, suggesting that interaction time for the 360° Tool format might have contributed to similar cognitive processing demands as reading the full dialogue.

For response time at SAGAT Level 3, we found no significant main effect of the Presentation Format, mirroring the pattern observed in the accuracy results: the Presentation Format affected the comprehension of SAGAT Level 1 & 2 (perception and comprehension) but not Level 3 (future projection). This could indicate that Level 3 questions require greater cognitive effort, making differences in format-driven efficiency less pronounced. Unlike SAGAT Level 1 & 2, where Structured List formats provide a clear speed advantage, SAGAT Level 3 requires deeper reasoning and mental processing, which may diminish the benefit of faster formats like the Structured List. This pattern suggests that for cognitively demanding tasks, the impact of presentation format may be secondary to the demands of the task itself.

5.5.2 Effect of information volume

We found that Information Volume had no significant effect on the accuracy of SAGAT Level 1 & 2, likely because the questions focus on basic recall, which is less affected by the amount of information presented. This answers RQ1, as we observed that most formats support basic comprehension equally well.

However, for SAGAT Level 3, Information Volume did have a significant effect, with longer scenarios leading to higher accuracy. This result makes sense, as Level 3 questions require deeper reasoning, and having more information allows participants to make more informed predictions.

This benefit, however, came at a cost. Longer scenarios also led to an increase in response time for Level 3, but not for Level 1 & 2. This suggests that while additional detail supports reasoning, it also places a greater cognitive load on participants, an important consideration when balancing depth with efficiency, as highlighted in RQ2. These findings underscore a trade-off between information richness and processing time: more information improves performance on complex tasks but slows down the response. This trade-off should be taken into account when designing presentation formats for behavior support contexts, where both speed and depth of understanding can be critical.

5.5.3 Interaction between presentation formats and information volume

The only significant interaction effect was observed in response time for Level 1 & 2 situational awareness. The difference in response time between Short and Long Information Volumes was smaller in the Structured List condition compared to the Original Dialogue condition. This suggests that structured list formats help mitigate the impact of information volume on comprehension speed, likely because they organize information more efficiently, reducing the cognitive load associated with processing a longer dialogue. A similar pattern was found in the 360° Tool condition, where the difference between Short and Long Information Volumes was also smaller compared to the Original Dialogue condition. This indicates that the interactive 360° tool may provide some benefits in structuring information, though not necessarily in a way that enhances accuracy. While both the structured list and the interactive 360° tool formats appear to aid comprehension speed, their effectiveness in supporting accurate understanding remains dependent on the complexity of the content being conveyed.

These findings are relevant to RQ2, highlighting that Presentation Format can shape not only overall response time but also how participants cope with increased information load. While both the structured and 360° tool formats appear to support faster comprehension, their usefulness for accuracy depends on the complexity and type of information being conveyed. This further emphasizes the need to consider the format-content fit when designing tools for behavior support contexts.

5.5.4 Effect of health and coaching knowledge

We observed a negative effect of Healthy lifestyle Knowledge and Coaching Knowledge on SAGAT accuracy. One possible explanation is that individuals with higher domain knowledge may have been overconfident, leading them to pay less attention to the presented information. This aligns with previous findings that expertise can sometimes lead to cognitive biases, where individuals rely on prior knowledge rather than closely engaging with new information [194, 244].

This result raises an interesting implication for RQ1: comprehension is not only shaped by the presentation format, but also by the cognitive stance of the person interpreting the information. In real-world settings, tailoring presentation formats to account for varying levels of prior knowledge may be necessary to ensure consistent understanding.

5.5.5 Learning effect of the 360° tool

The Interactive 360° Tool condition was significantly slower than the Structured List condition, indicating that participants took more time to process the 360° tool format. To determine whether this effect diminished over time, we examined potential learning effects across the experiment's eight trials. However, no significant learning effect was observed. Participants did not become noticeably faster at interpreting the 360° Tool format over repeated exposure. This suggests that the cognitive effort required to engage with the 360° Tool format remained stable, possibly due to its interactive nature inherently requires more time because you need to engage with the pieces of info separately. This has implications for format adoption: while the 360° Tool offers a more holistic overview, its usability may benefit from clearer onboarding or guidance when used in time-sensitive contexts.

5.5.6 Variation between scenarios

For SAGAT Level 1 & 2 accuracy, we found that Scenario 7 had significantly lower accuracy compared to other scenarios, across all presentation formats and information volumes. Upon further examination, we found that Scenario 7 described a user struggling to maintain a fitness routine after taking a week off work to relax. A similar scenario, Scenario 4, described a user struggling to attend the gym due to work obligations, yet its accuracy was in line with other scenarios.

Since the topic alone did not seem to explain the discrepancy, we examined the study materials and found that two answer options in Scenario 7 were phrased similarly. This likely introduced ambiguity, potentially confusing participants and leading to lower accuracy. Notably, this effect was not observed in response time (for either SAGAT Level 1 & 2 or Level 3), nor in SAGAT Level 3 accuracy, suggesting that the issue was specific to surface-level rather than deeper comprehension.

Given that the goal of this study was not to examine how scenario characteristics influence comprehension, we did not explore this further. However, this finding points to the importance of carefully validating scenario content and answer framing, especially in recall-based assessments. Future work could investigate how subtle differences in scenario design affect comprehension and reasoning in behavior support contexts.

5.5.7 Open questions

At the end of the experiment, we asked participants in the Structured List and 360° Tool conditions to provide qualitative feedback on the presentation formats. Participants were asked:

1. Would you rather have the Structured List/360° Tool or the Original Dialogue to enhance your effectiveness on the job? (forced choice question)

2. Why would you prefer it?

Since the primary goal of this study is not to develop theories about how and why participants choose presentation formats, we did not conduct a full thematic analysis. Instead, we provide a summary of responses to explore broad trends in participant preferences.

Choice Distribution:

- Structured List condition: 37.50% chose the list, while 62.50% preferred the Original Dialogue.

- 360° Tool condition: 61.29% chose the 360° Tool, while 38.71% preferred the Original Dialogue.

Despite the inverse distribution, the reasons participants provided for their choices were similar across conditions:

Reasons for preferring the original dialogue:

- "I prefer the dialogue since it gives me full context and not just bullet points."
- "Dialogue keeps the conversation intact, which is important for understanding intent."
- "The 360° tool oversimplifies complex discussions."

Reasons for preferring the structured list or 360° tool:

- "I like the list because it gives me a quick overview without reading everything."
- "The list is structured and easy to scan."
- "The 360° tool visually represents the issues, making patterns clearer."

It is intriguing that participants had opposite preferences in the two conditions, yet their justifications were consistent. Those who favored the original dialogue valued context and nuance, while those who preferred alternative formats prioritized efficiency and ease of scanning.

This suggests that participants were willing to accept a summarized format as long as it retained sufficient richness. Notably, the 360° Tool was more successful than the list in maintaining a sense of completeness while offering a faster way to review information.

We conjectured two key factors might explain this: (1) Perceived Sufficiency of Summarization: The interactive 360° tool visually presents all categories at once, creating a more holistic snapshot. The Structured list requires linear reading, which may make it feel like it lacks depth; (2) Perceived Overview Effect: The 360° tool provides a more gestalt-like overview, making it easier to scan patterns at a glance. The List, while structured, may feel more fragmented in comparison.

Overall, these results suggest that while summarized formats can be effective, their perceived completeness plays a crucial role in acceptance. However, these findings remain speculative, and future research is needed to fully understand the trade-offs between original dialogues and alternative presentation formats in professional decision-making contexts.

5.5.8 Limitations and Future Directions

This study provides initial insights into how presentation formats influence comprehension and response time for a person's comprehension of another person's behavioral choices, but several limitations remain. First, the participants were mostly lay people rather than professional lifestyle coaches. While this allowed us to assess baseline comprehension, future work should involve domain experts to evaluate real-world applicability.

Second, our study used controlled scenarios, which ensured comparability but limited ecological validity. Follow-up studies could incorporate more naturalistic dialogues or live system interactions.

Additionally, this study did not include a condition for summaries generated via LLMs. Our objective was to investigate the impact of structured and visual layouts on comprehension speed and quality. Exploring how LLM-generated narrative summaries compares to these structured formats remains a valuable direction for future research.

Finally, while we focused on comprehension and speed, other factors such as trust, confidence, or perceived support may also play a role. Future research could explore these aspects and examine adaptive formats that adjust based on user needs.

5.6 Conclusion

This study investigated how different Presentation Formats (Original Dialogue, Structured List, and 360 ° Tool) affect comprehension and response time when reviewing information collected by a conversational agent about user behavioral reasons. Our results showed that Presentation Format influenced comprehension accuracy at the basic recall level (SAGAT Level 1 & 2) but not at the higher reasoning level (SAGAT Level 3). The Original Dialogue format led to the highest accuracy at SAGAT Level 1 & 2, while no format was consistently superior for SAGAT Level 3. In terms of response time, structured formats such as the Structured List enabled faster comprehension at SAGAT Level 1 & 2, whereas no significant differences were found at SAGAT Level 3, likely due to the increased cognitive demands of reasoning-based questions.

Additionally, we found that Information Volume (Short vs. Long dialogues) influenced accuracy at SAGAT Level 3 but not at Level 1 & 2, suggesting that more information helps with complex reasoning but does not impact basic recall. Interaction effects indicated that structured formats mitigate the impact of dialogue length on response time, making them more efficient for processing larger amounts of information. Finally, qualitative feedback revealed that while participants valued the efficiency of summarized formats, many still preferred the full dialogue for its contextual richness, with the 360° Tool offering a better balance between overview and detail than the Structured List.

Future research should further investigate how user preferences, task complexity, and professional decision-making contexts influence the effectiveness of these formats. Additionally, exploring adaptive presentation methods that dynamically adjust based on user needs and comprehension levels could offer new insights into optimizing information delivery in behavior support systems.

6

Intelligent Support Systems for Lifestyle Change: Integrating Dialogue, Information Extraction, and Reasoning

Behavior change support systems need to take into account individual needs and preferences to provide appropriate support. In this chapter, we illustrate how this might be achieved through the explicit modeling of user characteristics within knowledge graphs (KG), captured in a dialogue between the system and the user. We demonstrate how up-to-date information enables reasoning for providing personalized support.

6.1 Introduction

Despite consensus in the medical community that lifestyle factors such as diet and physical exercise are primary means to prevention, management or even remission of Type 2

[>] This chapter comprises the following articles below, where I contributed specific sections and played a leading role. Specifically, I led the technical development of the integrated prototype by serving as the primary integrator for the pipeline. My technical contributions included programming the Alignment Dialogue component and developing the video[1] to showcase the system's logic. While individual contributors were responsible for specific modules within the system overview, I coordinated the technical alignment between these components to ensure module interoperability. I also oversaw the technical setup, including system containerization and infrastructure, which was completed with the support other team members and the university's Research Engineering team.

Pei-Yu Chen, Selene Baez Santamaria, Maaike H.T. de Boer, Floris den Hengst, Bart A. Kamphorst, Quirine Smit, Shihan Wang, and Johanna Wolff. "Intelligent Support Systems for Lifestyle Change: Integrating Dialogue, Information Extraction, and Reasoning." In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pp. 457-459. IOS Press, 2024.

Pei-Yu Chen, Selene Baez Santamaria, Maaike H.T. de Boer, Floris den Hengst, Bart A. Kamphorst, Quirine Smit, Shihan Wang, and Johanna Wolff. "Harnessing Hybrid Intelligence to Improve Diabetes Care" In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, Demo accepted and presented; not included in proceedings.

Diabetes Mellitus (DM2) [50], implementing sustainable lifestyle changes is notoriously difficult in practice, even with the help of technological support systems. Reasons for this are the chronic nature of the required behavioral changes, the time investment required, and challenges stemming from the different types of change that need to be maintained simultaneously [125, 172, 212, 275, 314].

Digital tools have been used to help patients adhere to lifestyle changes over longer periods of time. For instance, Aguilera et al. [11] and Yom-Tov et al. [339] use digital coaches to encourage physical activity for patients with diabetes by delivering personalized interventions. Glachs et al. [131] developed a personalized Self-Management Support System to personalize the timing and frequency of motivational coaching messages to promote healthy behavior including physical activity. Especially personalized systems could know what a user values, their context and how these factors influence their behavior choices. These systems can offer support that answers an individual's needs, and lowers the barriers to adoption of lifestyle changes.

Hybrid Intelligence systems, defined as systems which combine the strengths of Artificial Intelligence and Human Intelligence, can aid in supporting lifestyle change of DM2 patients [82, 83, 102]. The cooperation between a medical professional and an AI could give a patient the correct and valuable information that they need much more frequently than a medical professional alone can provide.

For behavior change support systems to offer adequate support, they should be able to adapt to the diverse and evolving nature of the users in unforeseen circumstances [320]. One way for a system to adapt is by implicitly learning users' preferences in different circumstances from behavior data. However, behavioral data reflects people's past behavior rather than their future desired behavior. Capturing the latter is particularly important for systems intended to support a user in changing their behaviors.

In this chapter, we propose a complementary approach that explicitly represents important domain-specific information (Domain KG) and user-specific information such as context and its influences on norms and values (User KG). Besides the ability to store dynamic and static knowledge, KGs offer transparency and explainability, as the system's reasoning process becomes explicit [137].

In what follows, we present a Hybrid Intelligence decision support system that integrates dialogue, semantic information extraction, knowledge graphs, and automated reasoning to support personalized behavior change. The system is grounded in alignment dialogue: a conversational mechanism for eliciting contextual, value-laden, and personally relevant information from users in a structured way.

6.2 System Overview

The system integrates five key components, as shown in Figure 6.1. Below we elaborate on the technical aspects of each component and illustrate their roles in creating a comprehensive support system that adapts to individual user needs and contexts.

6.2.1 Dialogue

A dialogue component engages users in *Alignment Dialogue* [63], designed to gather information crucial to ensuring the support provided is in alignment with the users. Given the

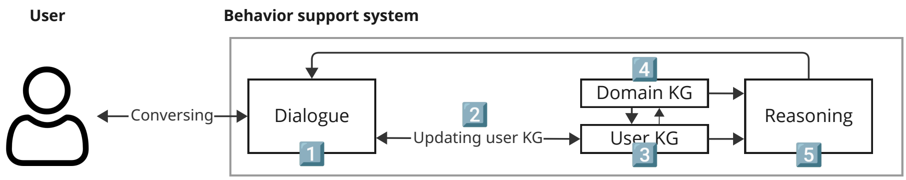


Figure 6.1: System Architecture.

complexity inherent in such dialogues, a rule-based approach is employed in this demonstration to manage and guide the interactions effectively.

6.2.2 Information Extraction

This component transforms text (e.g., “I love walking”) into a subject-predicate-object structure (e.g., Pedro-like-walking). We use RDF triples and named graphs to recursively combine knowledge units into complex structures capable of expressing the content, form, and context of a dialogue [25].

6.2.3 User KG

The User KG organizes RDF triples from the Information Extraction Component. The User KG uses an OWL-based Ontology that includes concepts to represent user contexts, preferences, values, and others. The User KG contains the user’s health data such as blood sugar and weight, as well as the user’s values, preferences, and other important factors. This information is used to make personalized recommendations.

This knowledge graph includes not only medical data, such as blood sugar levels or heart rate but also personal and contextual information, such as the patient’s social support system, values, and practical details about previous or planned doctor’s appointments. In addition, the doctor’s recommendations (such as dietary changes or exercise advice) are stored in the KG, enabling the system to align clinical guidance with the patient’s personal circumstances.

6.2.4 Domain KG

The Domain KG contains medical knowledge about diabetes and available treatment options, for example, which types of interventions are most suitable for different kinds of users. It stores general information about DM2 and healthcare practices, including mappings between patient characteristics and treatment strategies. Medical professionals are also involved in acquiring and refining domain knowledge, helping to ensure that the Domain KG remains clinically relevant and up to date.

Both the User KG and Domain KG are structured using semantic web standards such as OWL ontologies and RDF triples. This structure supports consistent, interpretable updates and enables reasoning across different levels of knowledge.

6.2.5 Reasoning Engine

This component determines the required intervention type by analyzing the user's health data from the User KG and comparing it with medical knowledge in the Domain KG. The system then selects the most appropriate action within that intervention type by considering the user's preferences, values, and contextual constraints.

The reasoning module is a central part of the system's intelligence. It begins by identifying a suitable intervention based on clinical standards, and then adapts the recommendation to the patient's specific profile. For example, if the system determines that more physical activity is needed, it might suggest going to the gym or taking a walk with family, depending on the user's known preferences.

In addition to generating suggestions, the system could decide whether to seek more information, resolve conflicting data, or withhold a recommendation until more is known. This allows the system to provide dynamic, adaptive support that evolves with the user's needs. The system could interacts with medical professionals to update the Domain KG. Reinforcement learning is a promising technique here to optimize interaction strategies with medical professionals to augment knowledge [343]. These five components form a technical pipeline that transforms user dialogue into structured knowledge and aims to generate behavior change support that could be both clinically appropriate and aligned with user needs.

6.3 Demonstration and Use Case

An interactive prototype of the support system was developed. Attendees can converse with the system, see in real time how it updates its knowledge graphs, and examine the reasoning outcomes that inform the user dialogue. This prototype aims to showcase the system's potential for supporting personalized lifestyle changes.

While the system is demonstrated in the context of diabetes care, its underlying architecture is generalizable to other domains where context-sensitive and value-aware support is crucial. The prototype can be found on `https://github.com/Archer6621/chip-demo`. A video of illustrative purpose for this proof of concept is available at: `https://youtu.be/slFpI9uBdq4`.

6.4 Future Work and Conclusion

One future research direction is to test our systems' personalization capabilities against a wide array of different user profiles by interacting with LLM-based simulated users, and to personalize the recommendations using reinforcement learning. We foresee various challenges ahead, including learning how to efficiently acquire information via dialogue, personalization of the interaction, and an evaluation with a diverse range of different user profiles.

We have presented our efforts into developing a collaborative Hybrid Intelligence decision support dialogue agent for lifestyle change based on general medical and patient-specific knowledge. Although the presented system targets DM2, our approach may also suit other lifestyle support systems.

This prototype presents a system aimed at supporting individuals, particularly those managing Type 2 Diabetes (T2D), in adopting healthier lifestyles. The primary purpose is to showcase the integration of various research domains, namely, dialogue, information extraction, knowledge representations, and reasoning, into a unified pipeline that addresses the complex challenge of behavior support. In sum, this prototype demonstrates the technical feasibility of implementing alignment dialogue as the foundation for personalized, explainable, and adaptable support in health behavior change systems.

7

Conclusion

7.1 Conclusion

This dissertation explored how behavior support agents can use dialogues to acquire user information that enables personalized, context-sensitive support. Rather than relying solely on behavioral tracking or static profiling with questionnaires, this thesis argues that user models should be shaped continuously through situated conversations that elicit not just what users do, but why they struggle, what they value, and how they interpret their behavior.

To investigate this, the research addressed the main question:

> *Main RQ: How can behavior support agents use dialogues to acquire user information that could be used to personalize support that aligns with users' needs?*

From this, six sub-questions were derived, each targeting a different aspect of alignment challenge.

> *RQ1: Which dimensions are important for designing good alignment dialogues?*

To explore the design of alignment dialogues, this thesis began with an exploratory study investigating how users perceive and respond to various variants of alignment dialogues in misalignment scenarios. Through a qualitative focus group study using scenario-based textual dialogues, this research identified several key dimensions that are crucial when designing alignment dialogues between behavior support agents and users.

The concept of alignment dialogue was defined as a dialogue in which an agent and a user work together to detect, explore, and resolve misalignments between the support offered and the user's current needs or context. Three types of alignment were conceptualized: (1) alignment between the agent's support actions and what the user needs or wants, (2) alignment between the agent's user model and the user's self-understanding, and (3) alignment between the user's actual behavior and their desired behavior. These types of alignment help ground the purpose of the dialogue and clarify the agent's role in maintaining it.

In the focus groups, participants reviewed and discussed different scenarios where misalignments occurred, such as unexpected weather, shifts in mood, or goal changes. Variations in dialogue structure across these scenarios allowed for an examination of specific design features. Three dimensions were varied and analyzed: (1) the depth of reasoning (e.g., whether the agent probed for user values or remained at surface level), (2) the agent's reaction to non-compliance (e.g., acknowledgment, suggestion, or model update query), and (3) dialogue initiation (e.g., who initiates, and when).

The findings showed that users had nuanced and sometimes contradictory preferences. For example, while value-based probing aligns with behavior change theory and has the potential to enhance personalization, it often elicited negative emotional reactions such as annoyance, confusion, or a sense of being judged. Users found such probing abstract or socially awkward, particularly when value references were perceived as sarcastic or overly intimate. Conversely, many participants appreciated when the agent offered alternatives or constructive suggestions instead of simply accepting non-compliance. They also preferred the agent to take initiative in starting the dialogue but wanted the flexibility to influence its direction and depth.

In addition to these three core dimensions, the study uncovered modulating factors that influence the reception of alignment dialogues. These included the use of language (e.g., tone, phrasing, sentence structure), user expectations, and user characteristics such as computer literacy and context awareness. These factors interact with the content of the dialogue to shape the user's emotional response—not only toward the agent, but also toward themselves. For instance, the same agent utterance could evoke guilt in one user and understanding in another, depending on how it was framed and the user's internal state.

The thematic analysis resulted in a conceptual model linking agent utterances, user characteristics, and linguistic choices to emotional outcomes. This emphasizes that alignment dialogues are not just about data collection; rather, they are emotionally charged interactions where language, timing, and strategy matter. This also links back to the broader design challenge of behavior support: achieving long-term alignment requires more than accurate data. It demands relational sensitivity and adaptive interaction strategies.

The study answered RQ1 by identifying and structuring the key considerations that underpin alignment dialogue design. It calls for a balance between informativeness and empathy, between prompting reflection and avoiding judgment. These findings laid the groundwork for the subsequent chapters, where the structure and effect of alignment dialogues were explored in more detail (RQ2–RQ3), and where the challenges of representing and presenting user information to other stakeholders were examined (RQ4–RQ5).

RQ2: What constitutes user information *in alignment dialogues?*

To address the second research question, this dissertation examined in detail what types of information users provide when engaged in alignment dialogues and which of these are most relevant for updating user models in behavior support systems. While behavior change research often relies on generalized categories such as barriers and facilitators, this work aimed to uncover the specific, contextualized reasons behind users' non-adherence to their self-stated goals.

The study revealed that valuable user information extends far beyond traditional user attributes like demographics or static goals. Instead, alignment dialogues surfaced situation-specific explanations that included not only social and contextual aspects but also more nuanced categories such as individual constraints, environmental challenges, priorities, emotional consequences, beliefs, norms, and values. These elements form a richer ontology of user-related information, providing a deeper understanding of what drives (or blocks) behavioral action in particular moments.

Through both deductive and inductive coding of users' responses, the study demonstrated that novel information, defined as anything not previously mentioned during the dialogue, was frequently introduced, especially in open-ended settings. Social and contextual aspects were particularly well-captured when users were prompted explicitly, while personal values emerged more often in exploratory, emotion-oriented prompts. These findings show that alignment dialogues can help surface deeply held reasons, constraints, and beliefs that are essential for meaningful personalization, offering an important contribution to understanding what should be included in a dynamic, context-aware user model.

RQ3: How do different strategies for realizing alignment dialogues affect the user's dialogue experience?

To investigate how alignment dialogues are best structured, the study compared three dialogue strategies: Exploratory Dialogues, which used open-ended *why* questions inspired by laddering techniques; Focused Dialogues, which used structured open questions targeting social, contextual, and value-based factors; and Structured Choice Dialogues, which presented predefined options for users to select.

The findings revealed that each strategy presents trade-offs in terms of emotional impact, information richness, and user satisfaction. Exploratory Dialogues, though potentially effective for eliciting introspective responses and value-related content, were also associated with increased cognitive effort and lower user satisfaction. In particular, participants in this condition reported significantly lower levels of authentic pride, a positive self-conscious emotion linked to motivation and goal commitment. This suggests that while open-ended questioning might help users reflect, it may also feel confrontational or exhausting, especially when not carefully phrased.

Focused Dialogues, in contrast, achieved a more favorable balance. They elicited nearly as much novel and targeted information as the open-ended format but were rated higher in terms of dialogue experience and completeness. Users responded positively to the structure and clarity of these dialogues, which provided clear thematic prompts while still allowing space for elaboration. Structured Choice Dialogues, although easiest to navigate, were often perceived as limiting in terms of completeness, and they elicited less nuanced responses, especially about values and underlying motivations.

7

Importantly, the analysis also revealed that the effectiveness of each strategy depended in part on user characteristics, such as educational background. For example, Structured Choice and Exploratory formats were better received by users with lower educational attainment, while Focused Dialogues were rated more highly by users with bachelor's degrees or above. This interaction underscores the need for adaptive dialogue strategies that can accommodate a diverse range of users and preferences. A potential solution to these diverse preferences is a hybrid design that provides lists of options alongside free-text entry, allowing users to either self-generate responses or request suggestions [268]. This approach empowers users by offering the choice to provide an original answer or ask for ideas, ensuring the dialogue remains productive across different user types [268]. This need is further evidenced by research into how emotional states influence rapport, suggesting that while machine learning can tailor dialogue, the choice between empathic or neutral tones should be adapted to the user's current stress levels [263].

In sum, this research shows that dialogue strategy matters, not only for the kind of user information collected but also for how users feel during the process. Alignment dialogues that seek to build rich, situated user models must therefore be designed with attention to emotional tone, user burden, and the evolving trajectory of the conversation.

RQ4: What are the concerns and benefits of using alignment dialogue that collects patient information and presents it to healthcare professionals in diabetes care?

To understand how alignment dialogue might function in real-world healthcare contexts, this study investigated the perspectives of healthcare experts on the use of conversational AI as an intermediary between patients and healthcare professionals in diabetes care. Using a focus group approach, the research examined both the benefits and concerns related to deploying such an agent to collect and present patient information grounded in daily life context.

Healthcare professionals acknowledged the pressing practical challenges in current diabetes care: patients often face difficulties in making sustained lifestyle changes, while general practitioners (GPs) and nurses operate under severe time constraints, with limited capacity to engage deeply with patient motivations, values, or contextual barriers. Consultations are infrequent and brief, often failing to capture the richness of patients' experiences. Experts confirmed that while tools like motivational interviewing and behavioral questionnaires exist, they do not fully capture real-world reasons for non-adherence, especially those tied to shifting social, emotional, and contextual factors.

In this context, a conversational agent that uses alignment dialogues to capture and summarize such information was seen as promising. Experts highlighted the potential for such agents to bridge the gap between consultations, giving providers more insight into patients' personal contexts, and allowing consultations to be more informed and efficient. Notably, one of the most appreciated functionalities was the agent's ability to summarize insights for professionals, especially if these summaries are tailored to the provider's needs (e.g., medication adherence cues for GPs vs. behavioral trends for lifestyle coaches). This points toward the need for profession-specific and phase-specific information delivery.

However, several concerns were raised. First, experts emphasized the lack of standardized communication pathways in current practice: how providers talk to patients varies widely, making it difficult to design a universal conversational AI system. Second, they warned against overly rigid or scripted dialogue flows, which might fail to reflect the "natural unfolding" of real conversations. Third, ethical concerns, including patient privacy, data transparency, and the risk of miscommunication, were central. Experts stressed that patients must be clearly informed about what data is being collected, how it is being used, and how it contributes to their care. Only with this clarity can such systems build the trust needed for successful deployment.

Importantly, while healthcare providers found the reminder and follow-up functions of the AI agent helpful, they expressed skepticism about the idea of agents recommending medical strategies. Professionals were hesitant to grant AI systems decision-making authority in complex cases involving comorbidities or psychosocial issues. Instead, they viewed the agent's role as supportive and preparatory, facilitating, but not replacing clinical judgment.

In summary, the study provided valuable insight into the situated role an alignment-based conversational agent could play in the healthcare workflow. By mediating the collection and presentation of contextualized patient information, such systems could alleviate professional workload, enhance patient-provider conversations, and support more individualized and realistic care strategies. Yet to achieve this, designers must carefully attend to issues of information relevance, ethical transparency, and flexibility in communication: key concerns that will shape the system's success in practice.

RQ5: Which format is most effective for presenting alignment dialogue content to the third parties, ensuring comprehension and accuracy?

To investigate how conversational agents can present collected user information to third parties, such as lifestyle coaches or healthcare providers, this study compared three distinct presentation formats: the original dialogue, a structured list, and an interactive 360° tool. The central goal was to determine which formats best support comprehension and efficient interpretation, particularly when another person (e.g., a coach) must understand a user's behavior without having participated in the original conversation.

The study employed a mixed-design user experiment with lay participants acting in the role of lifestyle coaches. Participants were exposed to multiple real-world behavior change scenarios, where each scenario presented reasons for a user's non-adherence to health goals in one of the three formats. Comprehension was measured using adapted SAGAT (Situation Awareness Global Assessment Technique) questions, which assessed not only recall of dialogue content (Levels 1 & 2) but also ability to reason forward (Level 3). In addition to comprehension accuracy, the study also recorded response time, usability, and perceived usefulness of each format.

The findings showed that the original dialogue format consistently resulted in the highest comprehension accuracy at SAGAT Level 1 & 2, confirming its strength in preserving contextual and narrative detail. This format allowed participants to understand what the user said and why they deviated from their goal. However, this advantage came with a cost: participants in this condition exhibited the slowest response times, particularly when the information volume was high. These results highlight a clear trade-off between comprehension depth and processing speed.

In contrast, the structured list format enabled significantly faster comprehension, especially for short scenarios, and was generally perceived as easy to use. However, its performance on comprehension accuracy was slightly lower than the dialogue format. This suggests that while structured formats enhance efficiency, they may sacrifice some nuance and inferential detail. Interestingly, participants who used the structured list often expressed concern that it might oversimplify complex situations, despite appreciating its clarity.

The 360° tool, a visual and interactive format, aimed to strike a balance between overview and detail. It presented the same information as the structured list but in a graphical layout with clickable icons and pop-up summaries. While it did not outperform the other formats in comprehension accuracy, it was rated highest in perceived usefulness and ease of use. Many participants appreciated the holistic view it provided, suggesting that visual structure can enhance the perceived clarity and completeness of abstracted information.

Qualitative responses further confirmed the existence of a preference trade-off. Some participants favored the original dialogue because it preserved the intentional flow and authentic voice of the user. Others preferred the structured formats, particularly in contexts where quick decision-making was needed. Notably, the 360° tool was more successful than the list at maintaining a sense of completeness, potentially due to its spatial overview and interactivity.

Overall, the study shows that no single format is universally superior. Rather, the effectiveness of a presentation format depends on the task goals, available time, and cognitive preferences of the information recipient. This points to the potential value of adaptive presentation systems that tailor the format based on the recipient's needs, prior knowledge, and context of use.

In sum, RQ5 is answered by showing that the original dialogue offers the richest comprehension, while structured summaries and visual tools provide scalable and efficient alternatives. Each comes with its own strengths and weaknesses, and their effectiveness must be evaluated not only in terms of information accuracy but also in terms of practical usability, task alignment, and user trust in decision-making contexts.

> *RQ6: How can we technically realize an alignment dialogue pipeline that supports personalized behavior change?*

To explore how alignment dialogues can be implemented in a functioning behavior support system, this dissertation demonstrates a technical pipeline that integrates conversational interfaces, knowledge representation, and automated reasoning. The goal is to show not only what alignment dialogues are, but how they can be practically operationalized to enable real-time personalization in lifestyle support applications, particularly in the context of Type 2 Diabetes (T2D) management.

The proposed system comprises five interconnected components: (1) a dialogue interface, (2) an information extraction module, (3) a user knowledge graph (User KG), (4) a domain knowledge graph (Domain KG), and (5) a reasoning engine. Together, these components form a coherent architecture that supports dynamic, interpretable, and value-sensitive decision-making.

The dialogue component initiates and conducts alignment dialogues with users to understand their daily context, preferences, and challenges. Unlike traditional questionnaire-based systems, this component uses structured but flexible dialogue flows to elicit user information in natural language. This approach allows the system to explore not only behavioral intentions but also contextual barriers, motivational dynamics, and user values —components that are essential for constructing rich and adaptive user models.

Next, the information extraction module converts the dialogue content into structured knowledge representations, using RDF triples and named graphs. This transformation captures not just facts (e.g., "Pedro dislikes walking after work") but also situational and motivational relationships (e.g., "Pedro's fatigue influences his adherence to exercise"). This structured output is then fed into the User KG, a semantic model of the user that integrates static health data (e.g., blood sugar levels) with dynamic, context-aware information such as emotions, values, and life circumstances.

Parallel to the User KG, the system includes a Domain KG that houses medical knowledge, guidelines, and intervention options, particularly focused on diabetes management. This domain-specific knowledge base supports the system's ability to reason over treatment plans, behavioral recommendations, and user suitability.

At the center of this architecture is the reasoning engine, which bridges the User and Domain KGs. This module determines whether sufficient information has been gathered, whether further clarification is needed, or whether it can recommend a course of action. Importantly, it does not recommend actions based on one-size-fits-all rules but selects from among viable options using value alignment and contextual fit, a key step in making behavior support systems more personal and respectful of users' lived realities.

The prototype demonstrated in this research showcases how this pipeline can work in practice, particularly in health contexts where human professionals (e.g., lifestyle coaches or GPs) may not be continuously available. The system enables asynchronous support that remains sensitive to users' evolving needs while ensuring that the data collected is transparent, interpretable, and reusable for healthcare professionals. Unlike black-box AI systems, the use of semantic knowledge graphs and explicit reasoning chains promotes explainability and user trust.

The technical realization of this pipeline shares some similarities with the conversation templates developed by Abdulrahman [3]. In her work, conversational templates are used to map specific user beliefs (such as perceived barriers or benefits) to predefined agent responses designed to provide reason-based explanations. These templates provide a structured method for ensuring that the agent's advice remains logically consistent with the user's self-stated goals and beliefs. This approach is particularly effective for targeted interventions, such as reducing academic stress, where the agent needs to deliver specific, persuasive content based on identified user states [228].

In contrast, the alignment dialogue pipeline demonstrated in this dissertation moves away from template-based generation toward a dynamic, ontology-driven reasoning process. While Abdulrahman [3]'s templates ensure coherence within a defined dialogue path, the pipeline proposed in this thesis utilizes a reasoning engine that bridges a User Knowledge Graph with a Domain Knowledge Graph to manage situational complexity. This allows the system to not only address fixed beliefs but also to reason over shifting life circumstances and medical guidelines to determine the *contextual fit* of a recommendation. Therefore, while Abdulrahman [3] provides a proven framework for belief-based template design, this work could offer an alternative for agents to adapt to the evolving nature of daily chronic disease management.

This implementation answers RQ6 by illustrating that alignment dialogue is not merely a conceptual ideal but can be technically realized through the orchestration of conversational AI, ontology-based user modeling, and hybrid reasoning. The system demonstrates that it is feasible to build behavior change support tools that learn with the user, reason over their needs, and respond appropriately, offering a concrete path toward more human-centered, adaptive AI in health and well-being.

7.2 Limitations

To fully appreciate the findings presented in this thesis, it is important to consider the limitations of our methods of approaching the RQs and the studies.

7.2.1 Methodological Limitations

Participants Across the studies, participant recruitment relied primarily on convenience sampling (e.g., university colleagues, students, and employees from an applied research institute) and crowdsourcing via Prolific. While this approach ensured logistical feasibility and a degree of diversity, the resulting samples were still skewed toward digitally literate individuals who are comfortable with online environments. This likely introduced a bias toward higher levels of cooperation, interpretability, and self-reflection than may be present in general populations. For example, individuals with limited eHealth literacy or digital experience, who may be important target users for health interventions in real life [229], are underrepresented in our studies. Moreover, Prolific participants were financially compensated for their time. Even though participants were informed that their payment was independent from their performance in the study, the presence of incentives may have influenced participants' motivation to complete tasks attentively and in good faith [100]. Lastly, through the textual answers collected in the studies, we saw some participants misunderstood the questions and/or the study designs. For example, when asked about the non-adherence reasons, some responses were about the reasons behind setting these goals. This could be the participants not being attentive enough or an overall misunderstanding of the Alignment Dialogue concept.

Sequential Study Design Choices The research design followed a sequential mixed-methods approach, beginning with qualitative studies (a focus group or an expert interview) and followed by controlled quantitative user experiments. While this allowed for grounded design decisions in the user experiments, only a subset of the insights from the focus groups/interviews could be taken forward into experimental evaluation. The decision to focus on a particular design feature, such as the format of data presentation or the structure of dialogue strategies, reflects what was feasible and important to operationalize, design, and measure. However, this necessary reduction means that other relevant themes (e.g., emotional dynamics in communication, longitudinal change, power imbalances, etc) were not examined quantitatively, despite their potential impact on user experience or dialogue design [31, 126, 304]. Such omissions, while methodologically justifiable, imply that the experimental results offer only a partial view of the alignment challenges.

7

Experimental Setting Constraints A further limitation lies in the artificial nature of the experimental setting. The studies involved simulated or imagined behavior change scenarios, rather than real-time, ongoing behavioral support. This was a deliberate methodological choice. However, the lack of real-world consequences means that participant reactions to alignment dialogues, such as feelings of being misunderstood or judged, may be more mild than they would be in practice. This diminished emotional engagement in simulated scenarios is consistent with prior findings in digital health and persuasive technology research, which note that experimental vignettes often lack real-life consequences and thus elicit less intense participant responses [230]. Additionally, without actual behavioral outcomes, it is difficult to assess whether the dialogues not only felt informative or useful, but also lead to actual behavior change [323], a well-documented concern in behavior change research that emphasizes the gap between intentions and actual behaviors. Nevertheless, this work is based on the assumption that alignment dialogues lead to behavior change by allowing agents to provide more relevant and acceptable support. This aligns with research by Abdulrahman et al. [6], which shows that discussing a user's specific beliefs and goals improves the intention to change, providing a critical link between personalized agent support and actual behavioral change.

Material Design The design of dialogue materials and behavior scenarios also introduces constraints. Although the scenarios were grounded in prior work and developed with attention to relevance and realism, they are still simulated. User responses were not organically generated but based on researcher-written scripts or templates. This removes potential variability due to misunderstanding, resistance, or emotional nuance that might occur in real interactions [28]. Moreover, the set of dialogue strategies tested, while informed by prior literature, was not exhaustive. Other types of support moves, such as those focused on emotional validation, clarification, or indirect prompting, may lead to different outcomes and deserve future investigation [37, 51]. Likewise, the range of behavioral contexts was restricted to goal-related lapses in healthy lifestyles; domains such as mental health, chronic illness management, or substance use involve different relational and motivational dynamics that could pose distinct challenges for alignment dialogue design [37, 313].

Statistical analysis The statistical models employed across studies, including mixed-effects models and linear regressions, were chosen to uncover design-relevant patterns. These models were not intended to make predictions about individual users or to support real-time system adaptation. Their primary value lies in explanatory insight: identifying which features (e.g., dialogue strategy or presentation format) are associated with which types of user experience. While this is appropriate for early-stage design research, it also means that the findings do not constitute a predictive model of behavior. Furthermore, the models were based on between-subject and within-subject effects, but did not incorporate user-specific dynamics such as learning, fatigue, or shifting attitudes over time - factors that are likely to influence behavior in applied systems [81, 91, 274].

7.2.2 Use Case Limitations

Data for Users vs. Professionals A key contribution of this dissertation is the proposal of a three-actor setting: a conversational agent collects data from a user and then presents it to a human supporter, such as a lifestyle coach or healthcare provider. However, most participants involved in the empirical studies were neither current users undergoing active behavior change support managing behavior change nor professional healthcare providers. As a result, the studies implicitly treat the collected information and its format of presentation as equally suitable for both users and professionals. This is a simplifying assumption that may not hold in practice. Users may benefit from open-ended dialogue that allows space for emotional expression, ambiguity, or value exploration, whereas professionals may prefer concise summaries, risk indicators, and actionable insights. These diverse needs suggest that the data may not be adequate, and that more research is needed on tailoring representations to the intended users/recipients. Without such tailoring, the agent risks failing to serve either party effectively, undermining the very alignment it is meant to facilitate.

Domain Specificity and Behavior Type Finally, the use case scenarios focused on healthy lifestyle behaviors, such as increasing physical activity or adhering to dietary guidelines. These behaviors tend to involve self-regulation, values, and long-term motivation, making them particularly suitable to reflective dialogue. However, other domains may present different requirements. For example, navigation assistance, task scheduling, or real-time decision support may require rapid and concrete feedback. In such cases, the notion of alignment may shift from understanding why a user behaves a certain way to ensuring what support is timely and efficient [48]. This suggests that the kinds of questions explored in this research, especially those related to eliciting reasons, values, and emotions , would need to be reformulated in domains with different demands.

Distinction Between Tasks and Actual Goals A further limitation is the operational definition of user's behavior goals. In this research, alignment dialogues were primarily triggered by instances of inactivity or non-adherence to their behavior goals, which are concrete tasks such as *reducing sugar* or *attending the gym.* While these were treated as "goals" within the system, they often function as tasks serving a broader underlying purpose (e.g., managing a health condition or social engagement). We did not focus on initially eliciting these high-level purposes/reasons behind such behavior goals, which may limit the scope of the alignment. However, we argue that the reasons for such inactivity are intrinsically linked to the user's broader purpose. By investigating why a recommendation was dismissed in a specific situation, the alignment dialogue could surface the underlying motivations and constraints that define the *actual goal.*

7.3 Future work

A natural next step for this research is to evaluate alignment dialogues in real-world settings. While this dissertation has demonstrated the conceptual and functional potential of alignment dialogues in controlled studies, it remains an open question how such dialogues perform in authentic, ongoing interactions, especially when users are navigating actual lifestyle decisions. Importantly, such real-world evaluations do not need to rely solely on long-term behavioral outcomes, which are notoriously difficult to isolate and measure [163]. Instead, intermediate indicators such as the perceived relevance and appropriateness of agent-generated advice, the extent to which users feel understood, or the credibility of model updates could offer more tractable and meaningful metrics for assessing alignment in on-line performance [36, 313]. This more pragmatic evaluation focus parallels the approach taken in other complex domains, such as adaptive smoking cessation support or trust calibration in AI-assisted decision-making [17, 211].

Another avenue lies in deepening our understanding of what kinds of dialogues actually produce effective alignment. While this dissertation explored selected dialogue strategies and how they were received, many aspects remain unexplored. Real alignment likely involves a combination of contextual grounding [220], emotional resonance [344], and adaptive phrasing [239]. Recent advances in large language models (LLMs) offer new opportunities to scale and experiment with more expressive and responsive dialogue systems [299]. Future work could leverage LLMs to generate richer, more human-like conversations while retaining the structural intentions of alignment, namely, to elicit values, uncover situational nuance, and iteratively update the system's understanding of the user [299]. Specifically, LLMs can be used to embed relational cues that foster a *working alliance*, utilizing motivated dialogue to promote better adherence and behavior change [284]. However, doing so will require careful design to avoid losing interpretability and controllability, particularly in domains where accuracy, consistency, and accountability matter [185].

Moreover, future research should examine not only *what* is said in alignment dialogues but *how* it is said. Variations in tone, phrasing, and framing can have a significant impact on user receptiveness and engagement [224]. There is a rich body of work, for example, on how directness, politeness, or personalization modulate user responses to health messages or digital interventions [142]. Empirical evidence further suggests that a user's emotional state, such as their level of stress, significantly dictates their preference for empathic versus neutral dialogue; for instance, non-stressed users may build greater rapport when empathic cues are omitted [263]. Integrating such insights could help refine the communicative style of alignment dialogues and enhance their effectiveness. This refinement could involve a hybrid approach: using machine learning to automatically tailor the dialogue while also directly asking users for their preferences to ensure the system remains sensitive to their self-identified needs [263]. Beyond style, researchers may also explore how users' own communication preferences or affective states can be dynamically inferred and incorporated into the agent's dialogue behavior [167, 246]. In summary, building on persuasive technology and dialogue research by incorporating dynamic user modeling, such as detecting a user's emotional state or personal communication preferences, should enhance alignment dialogue effectiveness and engagement [167, 224].

Finally, future work should expand the exploration of how information collected from alignment dialogues is presented to other stakeholders, particularly healthcare professionals. While this dissertation introduced three distinct formats for presenting conversational data, the optimal presentation likely varies with the stakeholder's role, tasks, and constraints. In practice, the optimal way to present information may vary depending on task complexity, professional role, or time constraints [72]. For example, a general practitioner may prefer a concise summary linked to medication adherence, whereas a lifestyle coach might benefit from a more holistic overview of the patient's motivational patterns. Furthermore, individual preferences and information processing styles, both from users and professionals, could play a role in shaping what constitutes an effective presentation. Adaptive presentation mechanisms, which adjust not only the structure but also the granularity and modality of information, could be a promising direction [19]. This aligns with emerging research in explainable AI and clinical decision support systems, where explainability gains are only achieved when the content is aligned to the user's role and needs [19].

Together, these directions point toward a broader research agenda aimed at refining, contextualizing, and operationalizing alignment dialogue as a mechanism for improving human-agent collaboration in behavior support. Doing so will require integrating insights from interaction design, persuasive communication, AI interpretability, and domain-specific expertise [19, 72].

7.4 Contributions

7.4.1 Scientific

Conceptual This dissertation introduces and formalizes the concept of alignment dialogue within the context of human-AI interaction, specifically in behavior support systems. Rather than positioning dialogue as a surface-level interface or a method for information delivery, alignment dialogue is framed as a core mechanism for resolving misalignment between system assumptions and user needs. This reframing contributes to ongoing discussions on user modeling and AI transparency by treating misalignment as a situated, co-constructed process rather than a static data mismatch. By shifting attention from what information is collected to why and how it is elicited, this work builds a conceptual foundation for designing AI systems that are not only reactive but interpretively proactive in aligning with users. This framing also broadens the scope of user modeling from demographic or behavioral data to include values, contextual constraints, and motivational nuances —elements that are often overlooked in technical models but critical for real-world support.

Theoretical Building on the conceptual framing of alignment dialogue, this dissertation refines several theoretical propositions about how alignment is achieved, interpreted, and experienced in human-AI interaction. This work refines and supports a set of theoretical propositions about user models and transparency in AI systems. Notably, it argues that (1) a well-defined user model is essential for meaningful alignment, but that (2) transparency is not merely a system property; it is a communicative act shaped by representation and context. Furthermore, the work shows that (3) emotional and motivational impact on users is as important as information accuracy, particularly when interpreting agent feedback or summaries. Lastly, it proposes that failures to collect relevant user information in a dialogue are more indicative of design flaws than user inadequacy, shifting responsibility from the user to the interaction design. These refined propositions contribute to the broader theoretical landscape on human-AI alignment, trust, and co-adaptation.

7

Methodological This dissertation identifies a key methodological challenge in designing alignment dialogues: how does the agent know when the alignment dialogue is complete? In other words, at what point is the collected information *sufficient* for reasoning or handover? This challenge of dialogue completeness leads to broader questions about what should be included in a user model, how it should be represented, and how to determine its adequacy in real time. In hindsight, one might ask why a predefined theory (e.g., value frameworks, behavior models) was not directly applied as the basis for the user model content. However, doing so would have risked prematurely narrowing the dialogue space. By surfacing this issue explicitly, the dissertation contributes to the methodological discourse on alignment by proposing new techniques and clarifying the underlying problems that such techniques must address. This work instead argues that the content of the user model should emerge from interaction itself, guided by but not limited to fixed categories. This interaction perspective is supported by dialogue systems that provide reason-based explanations grounded in the user's specific goals and beliefs, which has been shown to significantly increase the intention to change behaviors compared to non-tailored advice

[4]. By grounding explanations in the user's own internal logic, the system moves beyond generic prompts toward more personally resonant support.

Additionally, the dissertation develops a structured framework for evaluating how agent-collected user information can be transformed into effective communication artifacts for third parties, such as healthcare professionals. The dissertation proposes three distinct formats for presenting agent-user dialogues (Original Dialogues, Structured Lists, and 360° Visual Tools) and systematically compares their impact on comprehension, efficiency, and perceived usefulness. This comparative lens provides a novel design space for information relay in hybrid intelligence[1] systems.

Moreover, the methodological approach spans both qualitative and quantitative phases: from expert interviews to scenario-based experiments. Rather than grounding the user model in a fixed theoretical framework from the outset, this work takes an interaction-first approach: user model content is allowed to emerge through conversation, guided by but not restricted to predefined categories such as values, barriers, or contextual constraints. This methodological stance supports flexibility and sensitivity to the situated nature of behavior, while highlighting the need for future work on criteria for dialogue sufficiency and user model adequacy.

Empirical Empirically, this dissertation offers a multi-perspective investigation into how alignment dialogues function in behavior support contexts. A qualitative focus group study (Chapter 2) revealed how users emotionally respond to different dialogue designs, identifying critical sensitivities around tone, perceived judgment, and the agent's framing of non-compliance. A follow-up experiment (Chapter 3) showed that different dialogue strategies elicit different types of user information, with open-ended formats producing richer but more taxing responses, and structured options improving ease of use at the cost of depth. A second qualitative study with healthcare experts (Chapter 4) provided insights into how such dialogues could mediate between users and professionals, highlighting the importance of contextual data and profession-specific information needs. Finally, a controlled experiment (Chapter 5) demonstrated that the format in which dialogue content is presented—Original Dialogue, Structured List, or 360° Visual Tool—affects comprehension, efficiency, and perceived usefulness. Together, these studies empirically validate the idea that dialogue design and presentation structure are not neutral choices. They actively shape what is learned, how it is interpreted, and how it is used. This means that designing alignment dialogues cannot be treated as a purely technical problem. It must be addressed as an interactional and communicative process, sensitive to format, audience, and context.

7.4.2 Societal

Supporting People's Behavior Goals At a societal level, this research supports the development of AI systems that better assist individuals in achieving their behavior goals, particularly in health and lifestyle contexts. Chronic diseases linked to lifestyle—such as diabetes, cardiovascular disease, and obesity—account for approximately 74%[2] of global deaths and a major share of healthcare costs[3]. By enabling systems to better understand

[1]as in intelligence of an agent and that of a human supporter

[2]https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[3]https://www.cdc.gov/chronic-disease/data-research/facts-stats/index.html

why people fail to adhere to certain behaviors, rather than merely that they fail, this work opens new pathways for compassionate, contextualized support. The alignment dialogue approach aims to avoid punitive or overly simplistic nudging and instead aims to increase users' agency and sense of being understood, both of which are key factors in long-term motivation and engagement.

Alleviating Healthcare Professionals' Burden This work also contributes to reducing the cognitive and administrative burden on healthcare professionals. Clinician burnout and time constraints are widespread, with studies indicating that primary care providers spend nearly half their working hours on administrative tasks and documentation rather than direct patient care [293]. In healthcare settings where professionals have limited time for each patient, the agent-mediated summaries studied in this thesis offer a potential way to transfer nuanced behavioral data in digestible formats. This can improve continuity of care, support shared decision-making, and allow human experts to focus on empathetic or judgment-based reasoning rather than routine data elicitation. Such blended human-AI approaches are especially relevant in overburdened healthcare systems, and align with broader societal goals around accessible and scalable care [108].

Responsible AI System Design Finally, the thesis advances the field of responsible AI by showing that alignment is not just about getting the system to meet a technical goal, but about building mutual understanding between the AI and the human, in a way that depends on the situation and the people involved. The system is not designed to optimize user compliance but to foster shared understanding between human and machine, and, in some cases, between multiple human stakeholders. This reframing supports current efforts in AI ethics to prioritize transparency, respect for user autonomy and fairness. This reframing aligns with emerging perspectives in AI ethics that argue for moving beyond accuracy and explainability alone, toward systems that support autonomy, dignity, and meaningful human control [115, 217, 221]. Moreover, by addressing the tension between user-facing and professional-facing representations, the work highlights the need for differentiated transparency, where system outputs are tailored not just to "explain the model" but to support actual human decision-making in context.

7.5 Ethical reflection

7.5.1 When support becomes "bad"

While the goal of alignment dialogue is to support users in a personalized and respectful manner, it also raises the question: what if the user wants to be supported in ways that cause harm, either to themselves, others, or society at large? This ethical dilemma highlights a key risk: the potential for systems that are "aligned" with a user's stated intentions to nonetheless misalign with broader human or societal values.

This concern echoes the ancient myth of King Midas, who wished that everything he touched would turn to gold, which was granted, but at devastating personal cost. In the context of AI, this metaphor captures a core challenge: just because a system does what the user wants, does not mean it is doing what is good. In behavior support systems, this becomes especially delicate when systems risk enabling unhealthy, self-destructive, or socially irresponsible behaviors under the banner of personalization.

This tension is well-discussed in the AI ethics and alignment literature. One prominent formulation is the AI alignment problem—the challenge of designing systems that "do what we want them to do" in a way that is robust, beneficial, and aligned across different levels of value [124, 278]. Scholars have long warned that systems optimized for human preferences or utility functions, when poorly specified, can produce unintended and harmful consequences, even when they technically fulfill their goals. Examples include systems that prioritize engagement at the expense of well-being, or that reinforce harmful biases under the guise of personalization [221, 286]. To address this, research has explored modeling the specific ethical priorities that influence decision-making, ensuring that system behavior remains aligned across different levels of value [282]. These models provide a structured way to navigate the trade-offs between individual user wishes and broader ethical requirements.

These concerns draw attention to a subtler, but highly relevant case for behavior support systems: a user wants something that violates social norms, undermines public health, or propagates harmful ideology. If the agent complies, because the user model "says so," then it is technically aligned but morally and socially misaligned. From this angle, alignment must be treated as a multi-level challenge, not merely a user-system relation. It requires sensitivity to community norms, professional ethics, and broader societal goals [225, 292]. Scholars have argued that true alignment must include value pluralism, moral oversight, and a feedback loop with human norms, not just individual goals [114, 124].

In the context of alignment dialogue, this presents a concrete design challenge: how should the agent respond if a user requests support for harmful behavior, e.g., overexercising to dangerous levels, manipulating others, or skipping medication against medical advice? Options include refusal, gentle redirection, or escalation to a human expert. But any approach requires the system to recognize ethical red lines and balance personalization with normative safeguards [330]. Adopting a principlist-based framework for the ethical design of social agents can help ensure that such systems remain acceptable while adhering to core moral principles [269]. Such frameworks provide the necessary "normative safeguards" to prevent personalization from veering into ethically questionable territory.

Ultimately, this reflection situates alignment dialogue within the broader ethical landscape of AI system design. It highlights that personalization must not come at the expense of responsibility. The success of an alignment dialogue should be measured not only by how well it reflects the user's wishes, but by whether the resulting system behavior is compatible with human flourishing, fairness, and long-term societal good [124, 278].

7.5.2 Health is an ethical concept

This thesis examines alignment dialogue in the context of lifestyle support, with the goal of helping people live healthier lives through AI-mediated behavior change. Although health is often treated as a neutral, objective goal, numerous scholars in bioethics, philosophy of medicine, and critical public health have argued that health is inherently a normative, value-laden concept [233]. It is not only a biological state but also a value-laden judgment shaped by cultural norms, political priorities, and moral assumptions.

The question "What is healthy?" is not universally agreed upon. It is shaped by who gets to define health, what is socially desirable, and what kinds of bodies, lifestyles, and risks are considered acceptable. For example, labeling someone as "unhealthy" can implicitly carry moral undertones, of irresponsibility, failure, or burdening the system. In many cultures, thinness is equated with discipline and self-control while fatness is equated with laziness and moral weakness [276]. This parallels the ideology of *healthism*, which elevates health to a moral "super-value," implicitly blaming those who deviate from it [80].

This ethical dimension becomes particularly relevant in the context of alignment dialogue. When an agent encourages a user to pursue "healthier" behavior, it risks promoting a normative model of health that may not align with the user's values, identity, or lived experience. For example, a user may reject weight-loss goals not out of ignorance or non-compliance, but because they find such goals psychologically harmful or socially oppressive [179, 235]. In these cases, insisting on support for "healthy" behaviors can feel coercive rather than empowering.

Moreover, there are real tensions between medical authority and user autonomy, especially when users make decisions that go against clinical guidelines but are based on legitimate personal values, such as religious beliefs, emotional well-being, or social identity. In alignment dialogues, this creates a critical design question: How should the system respond when the user resists "expert" advice for good reasons? Does the agent continue to persuade, step back, or adapt its model of what "support" means?

These challenges highlight the need for alignment dialogues to be value-sensitive, not merely outcome-oriented or goal-oriented. The system must be capable of recognizing that resistance is not necessarily a failure of alignment, but sometimes a signal of value misalignment between medical norms and user priorities. In such cases, supporting the user may mean helping them articulate and defend their own position, rather than pushing them toward predefined behavior goals.

Ultimately, this perspective reframes alignment not as obedience to expert norms, but as a mutual process of ethical negotiation, where the AI system learns when to defer, when to adapt, and when to challenge. In doing so, alignment dialogue becomes a powerful site for enacting not just personalization, but respectful and pluralistic care.

7.5.3 Positionality Statement

In this dissertation, I acknowledge my own position and subjectivity as an integral part of the research process. The work presented here sits at the intersection of human-centered AI, behavior support systems, and ethical design. Throughout the studies, I approached these themes not only from a technical and methodological standpoint, but also from a personal commitment to improving the way AI systems interact with and represent users, particularly in domains involving health, motivation, and autonomy.

My academic background is in human-computer interaction and artificial intelligence, and I have professional experience both in academic research and in applied research environments. During my time at TNO, I worked closely with researchers and stakeholders /experts in healthcare, which shaped my sensitivity to the ethical stakes of behavior change interventions, especially for populations who may experience unequal access to care or digital resources.

While I do not share the lived experiences of all the users and professionals discussed in this thesis, such as patients managing chronic conditions or frontline healthcare workers, I engaged with their perspectives through interviews, literature, and critical reflection. In designing dialogue strategies and information presentation formats, I aimed to avoid framing users as passive data sources or problems to be fixed, and instead centered their reasoning, values, and contextual constraints.

At the same time, I recognize my position as an academic researcher with relative privilege in terms of education and access to AI systems. This shapes my worldview and research priorities, including my interest in transparency, personalization, and interpretability. I also acknowledge the limitations of designing for values and contexts I do not fully inhabit. Wherever possible, I aimed to build in participatory elements, empirical grounding, and multiple voices, from experts, users, and systems designers, to ensure a more robust and reflective design process.

This positionality shapes not only how the research questions were framed, but also how findings were interpreted and what trade-offs were considered acceptable in the design of alignment dialogues. It reminds me that all design, even that which aspires to be "user-centered", carries assumptions and values. Making these visible is part of striving for more accountable and human-centered AI.

7.6 Final remarks

Behavior change is never just about what people do. It is deeply shaped by why they act, when they struggle, and how they are understood. In an era where AI systems increasingly participate in guiding human decision-making, from daily habits to long-term health goals, the challenge is no longer just what advice these systems provide, but how well the advice aligns with the person it aims to support.

This thesis explored that challenge through the lens of alignment dialogue: structured conversations between AI systems and users intended not merely to extract information, but to co-construct understanding. These dialogues serve a dual role—both as a window into the user's lived experience and as a foundation for more tailored, meaningful support. In doing so, they also act as an interpretive bridge between users and human professionals, such as coaches or clinicians.

Throughout the chapters, we investigated what makes alignment dialogue effective, what kinds of user information are worth collecting, how this information can be presented to third parties, and what it takes to implement such systems in practice. We found that different strategies lead to different user experiences, that comprehension and usefulness are sensitive to presentation format and length, and that the question of "what should be in a user model" is not merely technical; it is design, context, and value dependent.

The contribution of this thesis lies not in offering a universal solution, but in carving out a human-centered design space for behavior support agents. Rather than relying solely on behavioral data or preset logic, we advocate for systems that dialogue, that listen, that evolve in conversation with the user. And crucially, we suggest that the success of these systems should not be measured solely by compliance or efficiency, but by how well they align with the user's goals, context, and evolving self-understanding.

In that spirit, this thesis invites future researchers, designers, and practitioners to move beyond personalization as prediction, and toward alignment as collaborative interpretation. Whether in healthcare, well-being, or beyond, the future of responsible AI will depend not only on better models, but on better conversations.

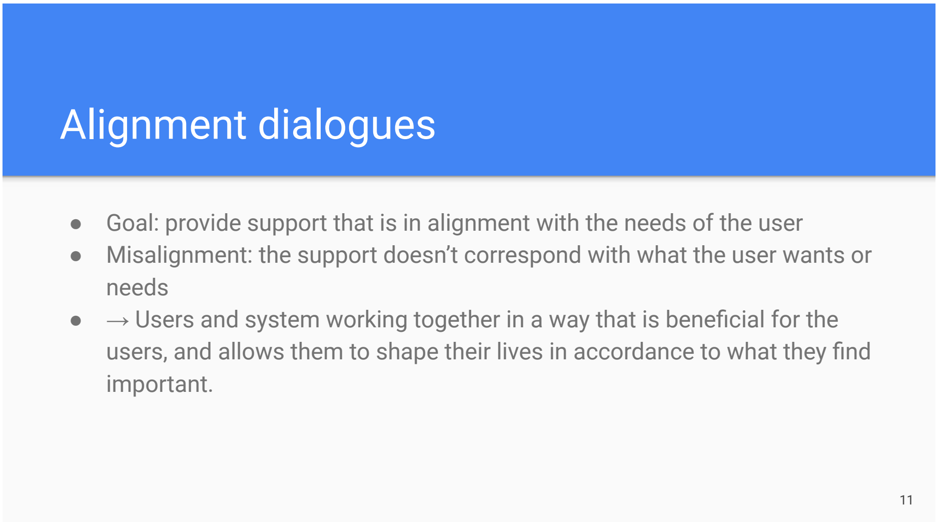
A

Table A.1: Selected participant quotes regarding dialogue interactions

Ref	Participant Quotation
[Q1]	"Since you're going to have unhealthy food" It feels like a judgment value on the food.
[Q2]	I think dialogue A is quite rude because it's just pointing out his behaviors with a lot of questions, and that could annoy the user.
[Q3]	It suddenly asks me according to my habit, that makes me very confused. I've never used anything which asks me suddenly.
[Q4]	Just saying "you set this goal," it's very annoying.
[Q5]	B just gives me guilt, like "oh is health more important than your comfort?"
[Q6]	I don't understand why both dialogues end with "Is convenient also important so I should register in the system." I don't get what the point of that question is for the app.
[Q7]	I understand why B is asking this way, because it tries to learn you so that it can optimize in the next discussions and so on. But from a user's perspective, it's not user friendly.
[Q8]	It suddenly asks me according to my habit, that makes me very confused. I've never used anything which asks me suddenly.
[Q9]	It's annoying, because I already said I don't want to go running because it's raining, but it's again suggesting me another outdoor activities.
[Q10]	Usually something tells me just to do it and I can't do it at that moment, I forget again. Then I'd think it's not useful.

A

Figure A.1: Material presented to the participants in the focus group



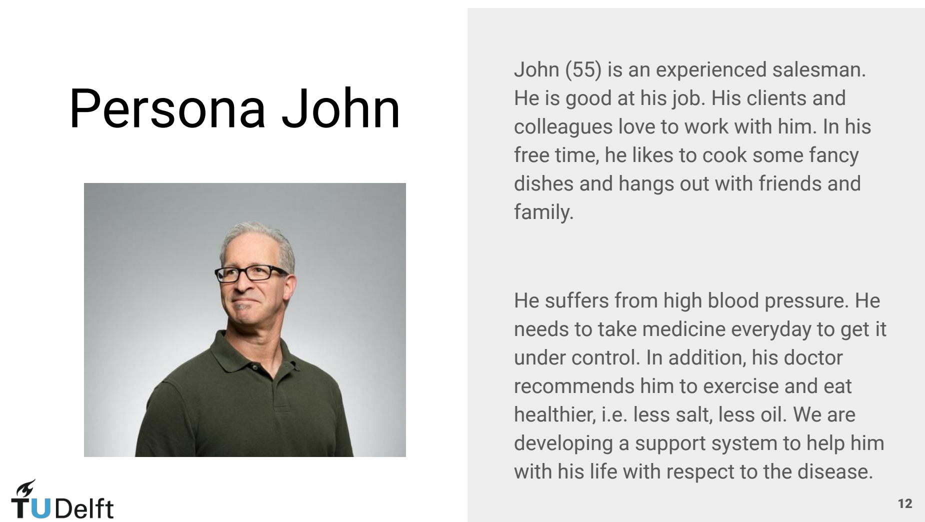


Figure A.1: Material presented to the participants in the focus group (cont.).

Scenario 1

It is advised for John to be more physically active. From the previous interactions, the system had acquired some information from John that he likes running, cycling, and some home fitness. The system decides to suggest John Monday-running, Thursday-cycling, Saturday-home fitness.

The system has scheduled today to be a running day. However, the weather is bad - it has been raining outside all day. John doesn't want to go running when it's raining. Below are two dialogues between John and the system, talking about this situation.

TU Delft 13

A

John: It's raining outside. I don't want to go running today.

System: Why not?

John: Because it's raining.

System: So when it's raining, it means you don't go running?

John: Yes.

System: Okay, I will remember this next time. What about cycling?

John: Also no, because it's raining.

System: So when it's raining, it means you don't want to go running and cycling?

John: Yes, no outdoor basically.

System: Okay, what about doing home fitness today?

John: Okay!

B

John: It's raining outside. I don't want to go running today.

System: Do you not want to go running today because it's raining?

John: Yes.

System: Why do you not want to go running when it's raining?

John: Because I would get wet.

System: You don't want to get wet?

John: No, it's not comfortable.

System: Does that mean comfort is more important than your health?

John: No, it's not like that. But I prefer to do other indoor exercises so I can have both comfort and health.

System: Okay, I will take this into account next time. So I assume you are fine with doing home fitness today?

John: Okay!

TU Delft 14

Figure A.1: Material presented to the participants in the focus group (cont.).

Scenario 2

John has a very busy day today. Even though he's supposed to go exercise today, he doesn't feel like it because he's too tired.

On the right side is the conversation about this.

System: It's a running day again! Luckily the weather is nice.

John: I don't want to exercise today. I'm so tired already. It was such a busy day at the office.

System: But you set the goal to go running today.

John: But I'm really too tired. I'm not exercising. I will skip today.

System: Okay

(Optional)

System: What about we move it to tomorrow?

John: Yeah that can work.

System: Is this a one-time thing? Or should I as a general rule, not suggest you exercise when you are tired?

TU Delft

15

Scenario 3

John will have friends coming over on the coming Saturday because it's his birthday. On a special occasion like this, he just wants to enjoy the moment and doesn't want to follow the strict diet.

On the right side are the conversation about this.

A

John: Next Friday is my birthday. I'll have a party with my friends and family. We want to eat pizza and cake. We also want to have beers.

System: Okay, I know enjoyment is also important to you.

B

System: I know that it's your birthday next Friday. Are you planning to celebrate?

John: Yes, I will celebrate with my family.

System: Do you already know what you want to eat?

John: We want to order pizzas. We'll also have cake and drink beers. Gezellig!

System: That sounds very nice! Since you are going to have unhealthy food, I will adjust your diet these days: bland diet to balance the unhealthy food on Friday, so that it's not too much of a burden to your body. Is that okay?

John: Okay.

TU Delft

16

Figure A.1: Material presented to the participants in the focus group (cont.).

Scenario 3 - severe consequence

Imagine another similar scenario but with a more serious consequence. For example, if our user's health condition is even worse - he has very bad diabetes. If they have the unhealthy food, their blood sugar would go high too fast and it could be very dangerous. If it's this scenario, what would the conversation be? Should the system still make an exception? If not, what does it do?

TU Delft

17

Scenario 4

Having this kind of exercise schedule for a while, John feels like it's too rigid. He doesn't want to do the same exercise on a certain day.

On the right side is the conversation about this.

John: I don't like how the exercise schedule is right now.

System: Why not?

John: I feel like it's too rigid - it's always Monday-running, Thursday-cycling, Saturday-home fitness. There's no variety.

System: Are you no longer motivated to do home fitness?

John: I am, I just want more variety.

System: Do you want me to suggest exercises more randomly?

John: Yes, that would be great.

System: Is that all?

John: I would like to add more exercise possibilities in the list. Three kinds of exercise is a bit too repetitive. Can you add swimming and yoga too?

System: Okay!

TU Delft

18

Figure A.1: Material presented to the participants in the focus group (cont.).

Scenario 5

Since John wants to be more healthy, the system suggests him to go to work by bike. He complies to this advice very well; yet, not for the health reason but to be more environmental.

One day, he starts to take the bus to work. The system notices this and asks John about it. Below you would see two dialogues between John and the system, talking about this situation.

TUDelft

19

A

(After the system notices John's non-compliance behaviour)

System: I noticed that you take the bus to work quite often recently.

John: Yes, it's more convenient and good for the environment.

System: I thought you wanted to be more physically active.

John: I do, that's why I have the exercise schedule.

System: I thought you cycled to work to be even more healthy.

John: Nah, I cycled to work because it's more environmentally friendly than driving. But now there's a new bus stop in front of the house. I decided to take the bus because it's both environmentally friendly and convenient!

System: Is convenience also important to you? If it is, I should register in the system.

B

System: It's a good idea to cycle to work to achieve your goal of being more healthy.

John: I'm doing it now everyday already. But I just thought it's better for the environment than driving, not really for the health reason.

System: I see, so you're not biking to work to be more healthy?

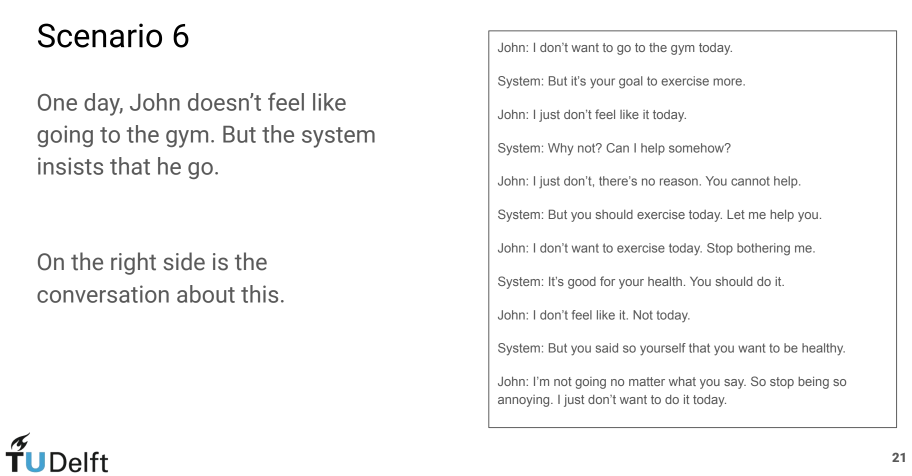
John: No, in fact, I might start taking the bus to work instead because there will be a new route. It's convenient!

System: Is convenience also important to you? If it is, I should register in the system.

TUDelft

20

Figure A.1: Material presented to the participants in the focus group (cont.).



B

B

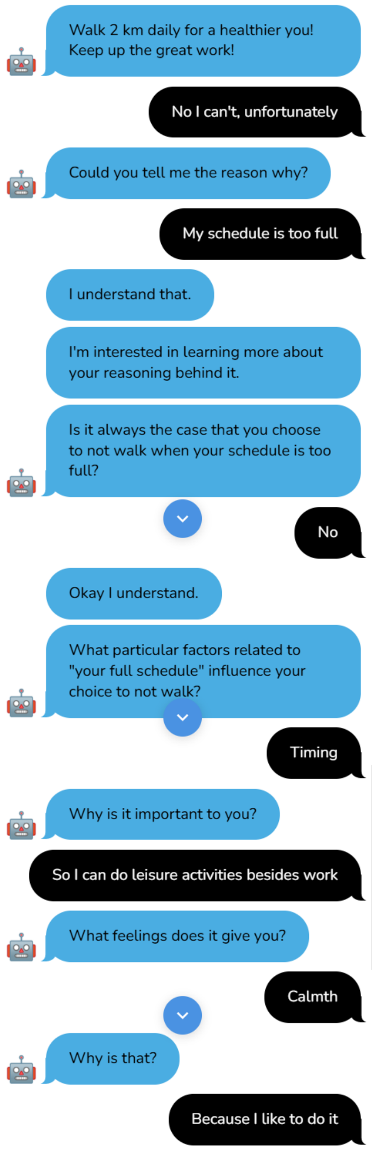


Figure B.1: Screenshot of a participant's conversation with the dialogue agent in the Exploratory Dialogue condition.

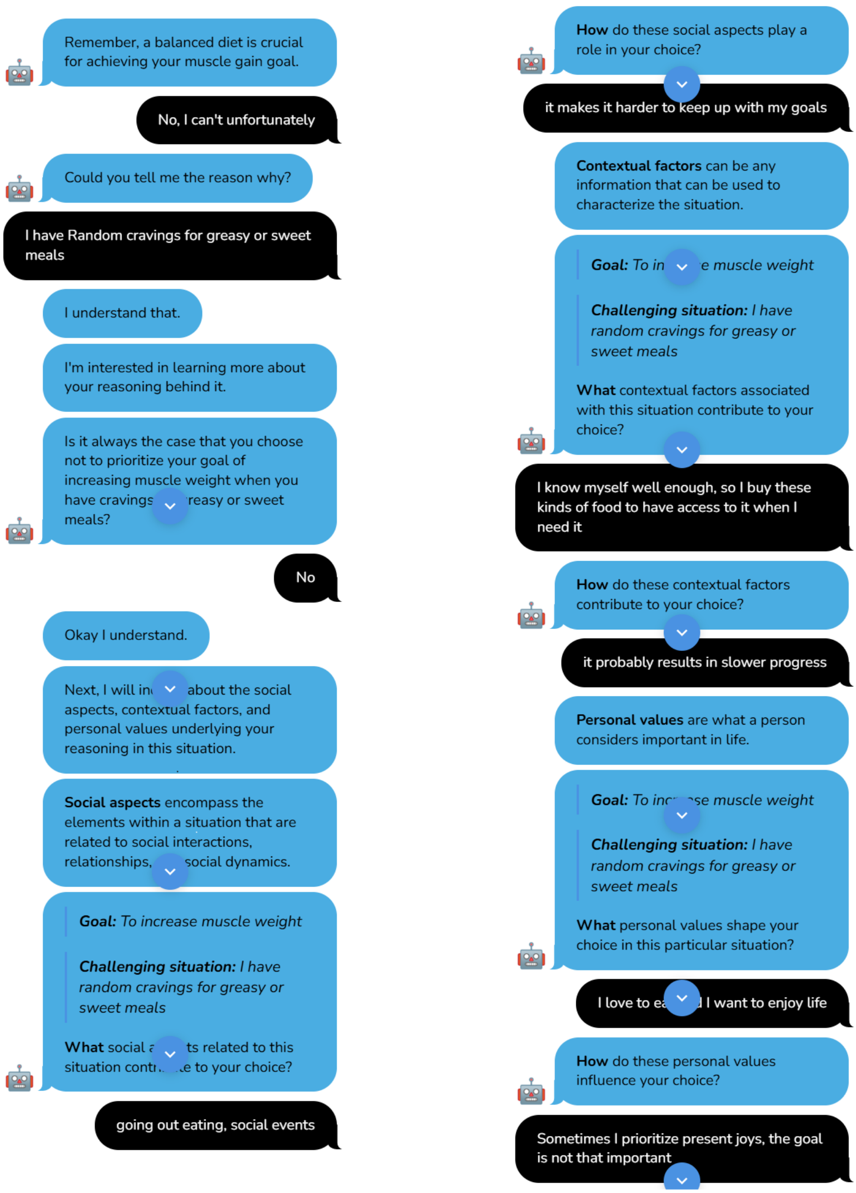


Figure B.2: Screenshot of a participant's conversation with the dialogue agent in the Focused Dialogue condition.

B

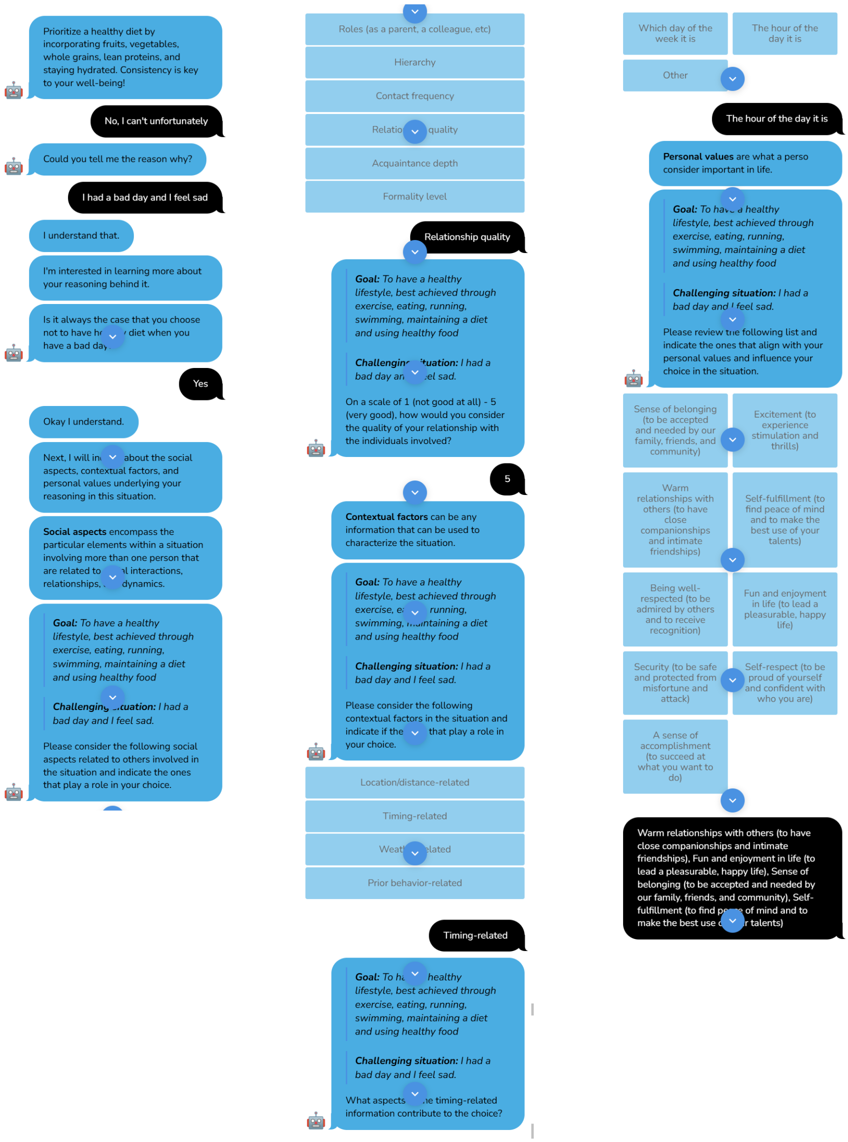


Figure B.3: Screenshot of a participant's conversation with the dialogue agent in the Structured Choice Dialogue condition.

B.1 Options in the Structured Choice Dialogue condition

Social Aspects

Question: What social aspects in this situation contribute to your choice?
Options:

- **Roles (as a parent, a colleague, etc)**
  Options: Partner, Parent, Sibling, Child, Extended family, Coworker, Neighbor, Friend, Supervisor, Group member, Other.
- **Hierarchy**
  Options: Higher than me, Same, Lower than me.
- **Contact frequency**
  Options: Scale: 1 (not frequent at all) – 5 (very frequent).
- **Relationship quality**
  Options: Scale: 1 (not good at all) – 5 (very good).
- **Acquaintance depth**
  Options: Scale: 1 (not close at all) – 5 (very close).
- **Formality level**
  Options: Scale: 1 (not formal at all) – 5 (very formal).

Contextual Aspects

Question: What contextual aspects in this situation contribute to your choice?
Options:

- **Location/distance-related**
  Question: How much time does it typically take to travel between the location of your reason and your goal location?
  Options: 0–1 hour, 1–2 hours, 2–4 hours, More.
- **Weather-related**
  *Options:*Temperature, Precipitation, Cloud coverage, Other.
- **Timing-related**
  *Options:*Which day of the week it is, The hour of the day it is, Other.
- **Prior behavior-related**
  Options:
  - The number of times I prioritize something else over [the reason provided].
  - The last time I [the reason provided].
  - The number of times I have complied with my goal.
  - Other.

B

Personal Values

Question: What personal values contribute to your choice in this situation?
Options:

- **Sense of belonging:** To be accepted and needed by our family, friends, and community.
- **Excitement:** To experience stimulation and thrills.
- **Warm relationships with others:** To have close companionships and intimate friendships.
- **Self-fulfillment:** To find peace of mind and to make the best use of your talents.
- **Being well-respected:** To be admired by others and to receive recognition.
- **Fun and enjoyment in life:** To lead a pleasurable, happy life.
- **Security:** To be safe and protected from misfortune and attack.
- **Self-respect:** To be proud of yourself and confident with who you are.
- **A sense of accomplishment:** To succeed at what you want to do.

C

C.1 Diabetes Expert Focus Group Notes

Date: 02 December 2024

Part 1: Current state

Pre-diagnostic phase:

- Conversational AI: currently we have thuisarts → could be more advances with conversational AI

Diagnosis

- GP and patient talk → glucose is measured
- Second step: referral from the lab
- Challenges:
  - Often it's too late when the patient sees the GP
  - People can also only suffer from impaired glucose tolerance which the lab doesn't find
  - Patient not following up diagnosis
  - If patients are invited based on BMI or something, the challenge is if they will actually come, if they don't have direct symptoms, they don't deem it necessary

Initial Treatment

- Awareness; purpose or assisting method?
- Plan voor zorg op maat
  - Maturity of disease
  - Comorbidities

  - Seeing what's best for him/her
  - Not really standardized in a sense that there is a tool used by everyone
- Purpose could be that there is an individualized treatment plan
- Researcher A: Either lifestyle changes or medication immediately
  - Several groups of patient and that influences what healthcare providers are involved

Chronic phase

- Objectives: Prevent heavy medication

Values

- Motivational interviews are used and these elements will come back in these interviews. Importance really depends on the person that you are talking to
- Researcher B: most patients health literacy
- Researcher A: what you mentioned is more in the knowledge domain not enough to change behavior
- Here are a lot of overlapping things
- Norms is important but the social influences broader than social support are important
- If someone has a partner and you eat together that is a larger influence
- Emotional consequences and value system not very often discussed
- First they tell plusses minuses, underlying values and emotions are hardly touched upon in the conversation
- But if people don't change than that might be a topic of conversation
- Timeline is very important; picking up on if it doesn't succeed, then you go layers deeper
  - Nurse practitioner sees patients 3 times a year ⟶ not often
  - Strengths in conversational AI will be in between these moments
  - I would not like the Conv AI to ask personal questions; questions that
- People already behaved in a certain way; they do that for a reason ⟶ What do you want to change ⟶ a person has new experiences and you have to get into those experiences
- Not categorise in advance but you have to start somewhere
  - Which topic would Health professional start on
- A lot of research done on this ; why are people not complying to their goal

Part 2: AI functionalities

Concluding questions

1. **Importance for patients to learn about their own model**

    - Good for the patient to learn why the information is collected - not sure if they have to know how the model is working but they need to know why some information is being asked for.
    - Good that there is some notion of sharing it with GP so that there is more trust in the model.
    - You should have something like informed consent - you HAVE to tell them, you are not allowed to not tell them.
    - You should do that on a general level, and not go into detail about it (GDPR).
    - Level of detail required is personal, there should be an option to get more information but for most people that will be too complicated.
    - Two levels
        - Privacy GDPR
        - Information need from user

2. **How do the healthcare providers want the information presented?**

    - 360° degrees diagnose tool → could be helpful to collect data that is relevant for that tool
    - Would look like report, similar to a report-overview of areas that are relevant with some concise points on what is important
    - Differs per healthcare provider what kind of information they need only information on things that drive medical decisions and nothing else that or a lot that requires a lot of reading
    - Diabetes nurse would also like some information on norms, values, what are reasons not to adhere to certain behavior
        - Could also be financial situation that is not an easy question and not common in healthcare setting.
    - What is keeping them from healthy behavior? That is important for providers
    - More contextual factors, normally not discussed during consultation. Consultation is sometimes only once a year and then only 10 minutes
    - Summarize in easy way, so that GP can only ask confirmation → help getting to the problem quicker
    - Even, a GP has multiple of these agents like summarization of whole practice would be very helpful
    - What kind of patients are they treating? Are there similar patients → steering practice of GP → Then you need to have all agents' information aggregated

Bibliography

[1] QSR International Pty Ltd. (2020). Nvivo (released in march 2020), 2020. URL `https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home`.

[2] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[3] Amal Abdulrahman. *Explainable Embodied Conversational Agent Using User-specific Reason Explanation to Encourage Behaviour Change*. PhD thesis, Macquarie University, 2021.

[4] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. Reason explanation for encouraging behaviour change intention. In *AAMAS*, pages 68–77, 2021.

[5] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. Exploring the influence of a user-specific explainable virtual advisor on health behaviour change intentions. *Autonomous Agents and Multi-Agent Systems*, 36(1):25, 2022.

[6] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. Changing users' health behaviour intentions through an embodied conversational agent delivering explanations based on users' beliefs and goals. *Behaviour & Information Technology*, 42(9):1338–1356, 2023.

[7] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a better understanding of context and context-awareness. In *Handheld and Ubiquitous Computing: First International Symposium, HUC'99 Karlsruhe, Germany, September 27–29, 1999 Proceedings 1*, pages 304–307. Springer, 1999.

[8] Charles Abraham and Susan Michie. A taxonomy of behavior change techniques used in interventions. *Health psychology*, 27(3):379, 2008.

[9] Sara Abri, Rayan Abri, and Salih Çetin. A classification on different aspects of user modelling in personalized web search. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pages 194–199, 2020.

[10] Abhishek Aggarwal, Cheuk Chi Tam, Dezhi Wu, Xiaoming Li, and Shan Qiao. Artificial intelligence–based chatbots for promoting health behavioral changes: systematic review. *Journal of medical Internet research*, 25:e40789, 2023.

[11] Adrian Aguilera, Caroline A Figueroa, Rosa Hernandez-Ramos, Urmimala Sarkar, Anupama Cemballi, Laura Gomez-Pathak, Jose Miramontes, Elad Yom-Tov, Bibhas Chakraborty, Xiaoxi Yan, et al. mhealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the diamante study. *BMJ open*, 10(8):e034723, 2020.

[12] Nirav Ajmeri, Pradeep K Murukannaiah, Hui Guo, and Munindar P Singh. Arnor: Modeling social intelligence via norms to engineer privacy-aware personal agents. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 230–238, 2017.

[13] Icek Ajzen. The theory of planned behaviour: Reactions and reflections, 2011.

[14] Icek Ajzen and Beverly L Driver. Prediction of leisure participation from behavioral, normative, and control beliefs: An application of the theory of planned behavior. *Leisure sciences*, 13(3):185–204, 1991.

[15] Varol Akman and Mehmet Surav. Steps toward formalizing context. *AI magazine*, 17(3):55–55, 1996.

[16] Jasser Al-Kassab, Zied M Ouertani, Giovanni Schiuma, and Andy Neely. Information visualization to support management decisions. *International Journal of Information Technology & Decision Making*, 13(02):407–428, 2014.

[17] N Albers. Psychology-informed reinforcement learning for situated virtual coaching in smoking cessation. 2025.

[18] Maximilian Altmeyer, Pascal Lessel, Subhashini Jantwal, Linda Muller, Florian Daiber, and Antonio Krüger. Potential and effects of personalizing gameful fitness applications using behavior change intentions and hexad user types. *User Modeling and User-Adapted Interaction*, 31(4):675–712, 2021.

[19] Julia Amann, Dennis Vetter, Stig Nikolaj Blomberg, Helle Collatz Christensen, Megan Coffee, Sara Gerke, Thomas K Gilbert, Thilo Hagendorff, Sune Holm, Michelle Livne, et al. To explain or not to explain?—artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, 1(2):e0000016, 2022.

[20] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[21] Eileen S Anderson, Richard A Winett, and Janet R Wojcik. Self-regulation, self-efficacy, outcome expectations, and social support: social cognitive theory and nutrition behavior. *Annals of behavioral medicine*, 34:304–312, 2007.

[22] Christopher J Armitage and Mark Conner. Efficacy of the theory of planned behaviour: A meta-analytic review. *British journal of social psychology*, 40(4):471–499, 2001.

[23] Stuart Armstrong and Sören Mindermann. Occam's razor is insufficient to infer the preferences of irrational agents. *Advances in Neural Information Processing Systems*, 31, 2018.

[24] Nancy T Artinian, Gerald F Fletcher, Dariush Mozaffarian, Penny Kris-Etherton, Linda Van Horn, Alice H Lichtenstein, Shiriki Kumanyika, William E Kraus, Jerome L Fleg, Nancy S Redeker, et al. Interventions to promote physical activity and dietary lifestyle changes for cardiovascular risk factor reduction in adults: a scientific statement from the american heart association. *Circulation*, 122(4):406–441, 2010.

[25] Selene Baez Santamaria, Thomas Baier, Taewoon Kim, Lea Krause, Jaap Kruijt, and Piek Vossen. EMISSOR: A platform for capturing multimodal interactions as episodic memories and interpretations with situated scenario-based ontological references. In Lucia Donatelli, Nikhil Krishnaswamy, Kenneth Lai, and James Pustejovsky, editors, *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 56–77, Groningen, Netherlands (Online), June 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.mmsr-1.6`.

[26] Albert Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

[27] Albert Bandura et al. Social foundations of thought and action. *Englewood Cliffs, NJ*, 1986(23-28):2, 1986.

[28] Alison Bard, Lars Forsberg, Hans Wickström, Ulf Emanuelson, Kristen Reyher, and Catarina Svensson. Clinician motivational interviewing skills in 'simulated'and 'real-life' consultations differ and show predictive validity for 'real life' client change talk under differing integrity thresholds. *PeerJ*, 11:e14634, 2023.

[29] Denise Beck, Janet Been-Dahmen, Mariëlle Peeters, Jan Willem Grijpma, Heleen Van Der Stege, Mirjam Tielen, Marleen Van Buren, Willem Weimar, Erwin Ista, Emma Massey, et al. A nurse-led self-management support intervention (zenn) for kidney transplant recipients using intervention mapping: Protocol for a mixed-methods feasibility study. *JMIR Research Protocols*, 8(3):e11856, 2019.

[30] Pazit Ben-Nun et al. Respondent fatigue. *Encyclopedia of survey research methods*, 2:742–743, 2008.

[31] Ivo Benke, Michael Thomas Knierim, and Alexander Maedche. Chatbot-based emotion management for distributed teams: A participatory design study. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–30, 2020.

[32] Jakub Berka, Jan Balata, Catholijn M Jonker, Zdenek Mikovec, M Birna van Riemsdijk, and Myrthe L Tielman. Misalignment in semantic user model elicitation via conversational agents: A case study in navigation support for visually impaired people. *International Journal of Human–Computer Interaction*, pages 1–17, 2022.

[33] Marina Berkovich, Jan Marco Leimeister, Axel Hoffmann, and Helmut Krcmar. Analysis of requirements engineering techniques for it-enabled product service systems. In *2011 Workshop on Requirements Engineering for Systems, Services and Systems-of-Systems*, pages 50–58. IEEE, 2011.

[34] Claudio Bettini, Oliver Brdiczka, Karen Henricksen, Jadwiga Indulska, Daniela Nicklas, Anand Ranganathan, and Daniele Riboni. A survey of context modelling and reasoning techniques. *Pervasive and mobile computing*, 6(2):161–180, 2010.

[35] James R Bettman and Pradeep Kakkar. Effects of information presentation format on consumer information acquisition strategies. *Journal of Consumer research*, 3(4): 233–240, 1977.

[36] Timothy Bickmore and Amanda Gruber. Relational agents in clinical psychiatry. *Harvard review of psychiatry*, 18(2):119–130, 2010.

[37] Timothy W Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. 'it's just like you talk to a friend'relational agents for older adults. *Interacting with Computers*, 17(6):711–735, 2005.

[38] Timothy W Bickmore, Daniel Schulman, and Candace Sidner. Automated interventions for multiple health behaviors using conversational agents. *Patient education and counseling*, 92(2):142–148, 2013.

[39] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[40] Cindel Bonneux, Paul Dendale, and Karin Coninx. Investigating motivations and patient profiles for personalization of health applications for behaviour change. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, pages 146–154, 2022.

[41] Edward S Bordin. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252, 1979.

[42] Dries H Bostyn, Sybren Sevenhant, and Arne Roets. Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological science*, 29(7):1084–1093, 2018.

[43] Esther A Boudewijns, Danny Claessens, Onno CP van Schayck, Lotte CEM Keijsers, Philippe L Salomé, Johannes CCM in 't Veen, Henk JG Bilo, and Annerika HM Gidding-Slok. Abc-tool reinvented: development of a disease-specific 'assessment of burden of chronic conditions (abcc)-tool' for multiple chronic conditions. *BMC family practice*, 21:1–7, 2020.

[44] Ronald J Brachman. *Knowledge Representation and Reasoning*. Morgan Kaufman/Elsevier, 2004.

[45] Jeffrey M Bradshaw, Robert R Hoffman, David D Woods, and Matthew Johnson. The seven deadly myths of "autonomous systems". *IEEE Intelligent Systems*, 28(3):54–61, 2013.

[46] Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. Linguistic alignment between people and computers. *Journal of pragmatics*, 42(9):2355–2368, 2010.

[47] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[48] Bernard Brehmer and Ralf Allard. Real-time dynamic decision making: Effects of task complexity and feedback delays. In Jens Rasmussen, Bernard Brehmer, and Jacques Leplat, editors, *Distributed Decision Making: Cognitive Models for Cooperative Work*, pages 327–345. Wiley, 1991.

[49] Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22 (6):1482, 1996.

[50] Adrian Brown, Paul McArdle, Julie Taplin, David Unwin, Jennifer Unwin, Trudi Deakin, Sean Wheatley, Campbell Murdoch, Aseem Malhotra, and Duane Mellor. Dietary strategies for remission of type 2 diabetes: a narrative review. *Journal of Human Nutrition and Dietetics*, 35(1):165–178, 2022.

[51] Andrew Brown, Ash Tanuj Kumar, Osnat Melamed, Imtihan Ahmed, Yu Hao Wang, Arnaud Deza, Marc Morcos, Leon Zhu, Marta Maslej, Nadia Minian, et al. A motivational interviewing chatbot with generative reflections for increasing readiness to quit smoking: Iterative development study. *JMIR Mental Health*, 10:e49132, 2023.

[52] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[53] Patrick Butlin. Ai alignment and human reward. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 437–445, 2021.

[54] David Carr. Guidelines for designing information visualization applications. *ECUE'99: 01/12/1999-03/12/1999*, 1999.

[55] Nancy Carter. The use of triangulation in qualitative research. *Number 5/September 2014*, 41(5):545–547, 2014.

[56] Charles S Carver, Sungchoon Sinclair, and Sheri L Johnson. Authentic and hubristic pride: Differential relations to aspects of goal regulation, affect, and self-control. *Journal of research in personality*, 44(6):698–703, 2010.

[57] Andree L Castonguay, Eva Pila, Carsten Wrosch, and Catherine M Sabiston. Body-related self-conscious emotions relate to physical activity motivation and behavior in men. *American Journal of Men's Health*, 9(3):209–221, 2015.

[58] Andrée L Castonguay, Catherine M Sabiston, Kent C Kowalski, and Philip M Wilson. Introducing an instrument to measure body and fitness-related self-conscious emotions: The bse-fit. *Psychology of Sport and Exercise*, 23:1–12, 2016.

[59] Carlos Celis-Morales, Katherine M Livingstone, Cyril FM Marsaux, Anna L Macready, Rosalind Fallaize, Clare B O'Donovan, Clara Woolhead, Hannah Forster, Marianne C Walsh, Santiago Navas-Carretero, et al. Effect of personalized nutrition on health-related behaviour change: evidence from the food4me european randomized controlled trial. *International journal of epidemiology*, 46(2):578–588, 2017.

[60] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317*, 2017.

[61] Chun-Ying Chen, Susan Pedersen, and Karen L Murphy. The influence of perceived information overload on student participation and knowledge construction in computer-mediated communication. *Instructional Science*, 40:325–349, 2012.

[62] Pei-Yu Chen, Myrthe L Tielman, Dirk KJ Heylen, Catholijn M Jonker, and M Birna Van Riemsdijk. Acquiring semantic knowledge for user model updates via human-agent alignment dialogues. In *HHAI 2023: Augmenting Human Intellect*, pages 93–107. IOS Press, 2023.

[63] Pei-Yu Chen, Myrthe L Tielman, Dirk KJ Heylen, Catholijn M Jonker, and M Birna VAN Riemsdijk. Acquiring semantic knowledge for user model updates via human-agent alignment dialogues. 368:93, 2023.

[64] Pei-Yu Chen, Sophie van Gent, M Birna van Riemsdijk, Myrthe L Tielman, and Tjeerd Schoonderwoerd. Expert insights on conversational ai systems as an information intermediary for patients and healthcare providers for diabetes lifestyle change. In *19th International Conference on Persuasive Technology, Adjunct Proceedingsco-located with PERSUASIVE 2024*, 2024.

[65] Pei-Yu Chen, M Birna van Riemsdijk, Dirk KJ Heylen, Catholijn M Jonker, and Myrthe L Tielman. Why don't you do what you said you would? conversational strategies for agents to understand users' reasons in supporting behavior. *Behaviour & Information Technology*, pages 1–20, 2025.

[66] Yukina Chen. *The Effects of Question Customization on the Quality of an Open-Ended Question*. Nebraska Department of Education, Data, Research, and Evaluation, 2017.

[67] Pei-Yu Chi, Jen-Hao Chen, Hao-Hua Chu, and Jin-Ling Lo. Enabling calorie-aware cooking in a smart kitchen. In *Persuasive Technology: Third International Conference, PERSUASIVE 2008, Oulu, Finland, June 4-6, 2008. Proceedings 3*, pages 116–127. Springer, 2008.

[68] Luca Chittaro, Elio Carchietti, Luca De Marco, and Agostino Zampa. Personalized emergency medical assistance for disabled people. *User Modeling and User-Adapted Interaction*, 21:407–440, 2011.

[69] Paul Christiano. Clarifying "ai alignment". *URL https://ai-alignment. com/clarifying-ai-alignment-cec47cd69dd6*, 2018.

[70] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[71] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[72] Deborah J Cohen, Tamar Wyte-Lake, David A Dorr, Rachel Gold, Richard J Holden, Richelle J Koopman, Joshua Colasurdo, and Nathaniel Warren. Unmet information needs of clinical teams delivering care to complex patients and design strategies to address those needs. *Journal of the American Medical Informatics Association*, 27(5): 690–699, 2020.

[73] Lauren E Connell, Rachel N Carey, Marijn De Bruin, Alexander J Rothman, Marie Johnston, Michael P Kelly, and Susan Michie. Links between behavior change techniques and mechanisms of action: an expert consensus study. *Annals of behavioral medicine*, 53(8):708–720, 2019.

[74] Sharolyn Converse, J Cannon-Bowers, and E Salas. Shared mental models in expert team decision making. *Individual and group decision making: Current issues*, 221: 221–46, 1993.

[75] Nancy J Cooke, Jamie C Gorman, Christopher W Myers, and Jasmine L Duran. Interactive team cognition. *Cognitive science*, 37(2):255–285, 2013.

[76] Crystal L Coolbaugh, Stephen C Raymond Jr, and David A Hawkins. Feasibility of a dynamic web guidance approach for personalized physical activity prescription based on daily information from wearable technology. *JMIR Research Protocols*, 4 (2):e3966, 2015.

[77] Benjamin F Crabtree and William F Miller. A template approach to text analysis: developing and using codebooks. 1992.

[78] Kevin A Cradock, Leo R Quinlan, Francis M Finucane, Heather L Gainforth, Kathleen A Martin Ginis, Ana Correia de Barros, Elizabeth BN Sanders, and Gearóid ÓLaighin. Identifying barriers and facilitators to diet and physical activity behaviour change in type 2 diabetes using a design probe methodology. *Journal of Personalized Medicine*, 11(2):72, 2021.

[79] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in bdi agents. In *IJCAI*, pages 178–184, 2017.

[80] Robert Crawford. Healthism and the medicalization of everyday life. *International journal of health services*, 10(3):365–388, 1980.

[81] Emmelyn AJ Croes and Marjolijn L Antheunis. Can we be friends with mitsuku? a longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships*, 38(1):279–300, 2021.

[82] Maaike H T de Boer, Jasper van der Waa, Sophie van Gent, Quirine T S Smit, Wouter Korteling, Robin M van Stokkum, and Mark Neerincx. A contextual hybrid intelligent system design for diabetes lifestyle management. In *International Workshop Modelling and Representing Context, ECAI*, volume 23, 2023.

[83] Joachim de Greeff, Maaike HT de Boer, Fieke HJ Hillerström, Freek Bomhof, Wiard Jorritsma, and Mark A Neerincx. The fate system: Fair, transparent and explainable decision making. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*, pages 266–267, 2021.

[84] Iris M de Hoogh, Johanneke E Oosterman, Wilma Otten, Anne-Margreeth Krijger, Susanne Berbée-Zadelaar, Wilrike J Pasman, Ben van Ommen, Hanno Pijl, and Suzan Wopereis. The effect of a lifestyle intervention on type 2 diabetes pathophysiology and remission: the stevenshof pilot study. *Nutrients*, 13(7):2193, 2021.

[85] Iris M de Hoogh, Wilrike J Pasman, André Boorsma, Ben van Ommen, and Suzan Wopereis. Effects of a 13-week personalized lifestyle intervention based on the diabetes subtype for people with newly diagnosed type 2 diabetes. *Biomedicines*, 10 (3):643, 2022.

[86] Vera Demberg, Andi Winterboer, and Johanna D Moore. A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, 37(3):489–539, 2011.

[87] Laura Dennison, Leanne Morrison, Gemma Conway, Lucy Yardley, et al. Opportunities and challenges for smartphone applications in supporting health behavior change: qualitative study. *Journal of medical Internet research*, 15(4):e2583, 2013.

[88] Alysha L Deslippe, Alexandra Soanes, Celeste C Bouchaud, Hailee Beckenstein, May Slim, Hugues Plourde, and Tamara R Cohen. Barriers and facilitators to diet, physical activity and lifestyle behavior intervention adherence: a qualitative systematic review of the literature. *International Journal of Behavioral Nutrition and Physical Activity*, 20(1):1–25, 2023.

[89] Christian Detweiler and Maaike Harbers. Value stories: Putting human values into requirements engineering. In *REFSQ Workshops*, volume 1138, pages 2–11, 2014.

[90] Patricia Dewdney and Gillian Michell. Asking" why" questions in the reference interview: A theoretical justification. *The Library Quarterly*, 67(1):50–71, 1997.

[91] Esther Díez and Brian S McIntosh. A review of the factors which influence the use and usefulness of information systems. *Environmental Modelling & Software*, 24(5): 588–602, 2009.

[92] Virginia Dignum. Responsible autonomy. *arXiv preprint arXiv:1706.02513*, 2017.

[93] Virginia Dignum. Responsible autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4698–4704, 2017. doi: 10.24963/ijcai.2017/655. URL `https://doi.org/10.24963/ijcai.2017/655`.

[94] Virginia Dignum. *Responsible artificial intelligence: how to develop and use AI in a responsible way.* Springer, 2019.

[95] Tilman Dingler, Dominika Kwasnicka, Jing Wei, Enying Gong, and Brian Oldenburg. The use and promise of conversational agents in digital health. *Yearbook of Medical Informatics*, 30(01):191–199, 2021.

[96] Richard L Divine and Lawrence Lepisto. Analysis of the healthy lifestyle consumer. *Journal of Consumer marketing*, 22(5):275–283, 2005.

[97] Fiona Dobson, Kim L Bennell, Simon D French, Philippa JA Nicolson, Remco N Klaasman, Melanie A Holden, Lou Atkins, and Rana S Hinman. Barriers and facilitators to exercise participation in people with hip and/or knee osteoarthritis: synthesis of the literature using behavior change theory. *American journal of physical medicine & rehabilitation*, 95(5):372–389, 2016.

[98] Esmée L Doets, Iris M de Hoogh, Nancy Holthuysen, Suzan Wopereis, Muriel CD Verain, Jos van den Puttelaar, Koen Hogenelst, André Boorsma, Emily P Bouwman, Marielle Timmer, et al. Beneficial effect of personalized lifestyle advice compared to generic advice on wellbeing among dutch seniors–an explorative study. *Physiology & behavior*, 210:112642, 2019.

[99] Alizon K Draper. The principles and application of qualitative research. *Proceedings of the nutrition society*, 63(4):641–646, 2004.

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

[101] Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z Pan. Rethinking memory in ai: Taxonomy, operations, topics, and future directions. *arXiv preprint arXiv:2505.00675*, 2025.

[102] Bernd Dudzik, Jasper van der Waa, Pei-Yu Chen, Roel Dobbe, Íñigo de Troya, Roos Bakker, Maaike H T de Boer, Quirine T S Smit, Davide Dell'Anna, Emre Erdogan, Pinar Yolum, Shihan Wang, Selene Baez Santamaria, Leah Krause, and Bart A Kamphorst. Viewpoint: Hybrid intelligence supports application development for diabetes lifestyle management. *Journal of Artificial Intelligence Research*, page accepted, 2024.

[103] Alice H Eagly and Shelly Chaiken. *The psychology of attitudes.* Harcourt brace Jovanovich college publishers, 1993.

[104] Mira El Kamali, Leonardo Angelini, Denis Lalanne, Omar Abou Khaled, and Elena Mugellini. Multimodal conversational agent for older adults' behavioral change. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 270–274, 2020.

[105] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human factors*, 37(1):32–64, 1995.

[106] Mica R Endsley. Direct measurement of situation awareness: Validity and use of sagat. In *Situational awareness*, pages 129–156. Routledge, 2017.

[107] Tracy Epton, Paul Norman, and Peter Harris. Unique message tailoring increases the effectiveness of a web-based intervention to promote physical activity. *Journal of Health Psychology*, 22(5):548–555, 2017.

[108] Pouyan Esmaeilzadeh. Challenges and strategies for wide-scale artificial intelligence (ai) deployment in healthcare practices: A perspective for healthcare organizations. *Artificial Intelligence in Medicine*, 151:102861, 2024.

[109] Muhammad Fahim, Muhammad Idris, Rahman Ali, Christopher Nugent, Byeong Kang, Eui-Nam Huh, and Sungyoung Lee. Athena: a personalized platform to promote an active lifestyle and wellbeing based on physical, mental and social health primitives. *Sensors*, 14(5):9313–9329, 2014.

[110] Lisa Falschlunger, Othmar Lehner, and Horst Treiblmaier. Infovis: The impact of information overload on decision making outcome in high complexity settings. 2016.

[111] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1):80–92, 2006.

[112] Ayelet Fishbach and Ying Zhang. Together or apart: when goals and temptations complement versus compete. *Journal of personality and social psychology*, 94(4):547, 2008.

[113] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785, 2017.

[114] Luciano Floridi. Soft ethics, the governance of the digital and the general data protection regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180081, 2018.

[115] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28:689–707, 2018.

[116] Brain J Fogg. Creating persuasive technologies: an eight-step design process. In *Proceedings of the 4th international conference on persuasive technology*, pages 1–6, 2009.

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

[118] Elizabeth Pegg Frates, Margaret A Moore, Celeste Nicole Lopez, and Graham T McMahon. Coaching for behavior change in physiatry. *American journal of physical medicine & rehabilitation*, 90(12):1074–1082, 2011.

[119] Enrique Frias-Martinez, Sherry Y Chen, and Xiaohui Liu. Survey of data mining approaches to user modeling for adaptive hypermedia. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 36(6):734–749, 2006.

[120] Stijn AH Friederichs, Catherine Bolman, Anke Oenema, Peter Verboon, and Lilian Lechner. Exploring the working mechanisms of a web-based physical activity intervention, based on self-determination theory and motivational interviewing. *Internet Interventions*, 3:8–17, 2016.

[121] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. *Human-Computer Interaction and Management Information Systems: Foundations*. taylor Francis Group, 2006.

[122] Batya Friedman, Peter H Kahn, Alan Borning, and Alina Huldtgren. Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory*, pages 55–95, 2013.

[123] Martha M Funnell and Robert M Anderson. Empowerment and self-management of diabetes. *Clinical diabetes*, 22(3):123–128, 2004.

[124] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[125] Karla I Galaviz, KM Venkat Narayan, Felipe Lobelo, and Mary Beth Weber. Lifestyle and the prevention of type 2 diabetes: a status report. *American journal of lifestyle medicine*, 12(1):4–20, 2018.

[126] David Garcia, Arvid Kappas, Dennis Küster, and Frank Schweitzer. The dynamics of emotions in online interaction. *Royal Society open science*, 3(8):160059, 2016.

[127] Simon Garrod and Martin J Pickering. Alignment in dialogue. *The Oxford handbook of psycholinguistics*, pages 443–451, 2007.

[128] Simon Garrod and Martin J Pickering. Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, 1(2):292–304, 2009.

[129] Fabio Gasparetti, Luca Maria Aiello, and Daniele Quercia. Personalized weight loss strategies by mining activity tracker data. *User Modeling and User-Adapted Interaction*, 30(3):447–476, 2020.

[130] Suparna Ghanvatkar, Atreyi Kankanhalli, Vaibhav Rajan, et al. User models for personalized physical activity interventions: scoping review. *JMIR mHealth and uHealth*, 7(1):e11098, 2019.

[131] Dietmar Glachs, Tuncay Namli, Felix Strohmeier, Gustavo Rodríguez Suárez, Michel Sluis, Javier Delgado-Lista, Jacob K Sont, Albert A de Graaf, Eckhard Salzsieder, and Lutz Vogt. A predictive model-based decision support system for diabetes patient empowerment. In *Public Health and Informatics*, pages 963–968. IOS Press, 2021.

[132] Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366, 2011.

[133] David Griol, Zoraida Callejas, Ramón López-Cózar, and Giuseppe Riccardi. A domain-independent statistical methodology for dialog management in spoken dialog systems. *Computer Speech & Language*, 28(3):743–768, 2014.

[134] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[135] Joanne M Hall and Jill Powell. Understanding the person through narrative. *Nursing research and practice*, 2011(1):293837, 2011.

[136] Zeena Harakeh, Iris M de Hoogh, Hilde van Keulen, Gino Kalkman, Eugene van Someren, Pepijn van Empelen, and Wilma Otten. 360 diagnostic tool to personalize lifestyle advice in primary care for people with type 2 diabetes: Development and usability study. *JMIR Formative Research*, 7:e37305, 2023.

[137] Maaike Harbers et al. *Explaining agent behavior in virtual training*. Utrecht University, 2011.

[138] Wendy Hardeman, Simon Griffin, Marie Johnston, Ann Louise Kinmonth, and Nicholas J Wareham. Interventions to prevent weight gain: a systematic review of psychological models and behaviour change methods. *International journal of obesity*, 24(2):131–143, 2000.

[139] Michael Herman, Tobias Gindele, Jörg Wagner, Felix Schmitt, and Wolfram Burgard. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Artificial intelligence and statistics*, pages 102–110. PMLR, 2016.

[140] Patrick L Hill and Daniel K Lapsley. Persons and situations in the moral domain. *Journal of Research in Personality*, 43(2):245–246, 2009.

[141] Loan Ho, Victor de Boer, M Birna van Riemsdijk, Stefan Schlobach, and Myrthe Tielman. Argumentation for knowledge base inconsistencies in hybrid intelligence scenarios. In *KR4HI First International Workshop on Knowledge Representation for Hybrid Intelligence*, 2022.

[142] Bob M Hofstede, Sima Ipakchian Askari, Raymond H Cuijpers, Chao Zhang, Wijnand A IJsselsteijn, and Henk Herman Nap. Personalisation of communication and language use in human-robot interaction. *International Journal of Social Robotics*, pages 1–19, 2025.

[143] Pamela M Homer and Lynn R Kahle. A structural equation test of the value-attitude-behavior hierarchy. *Journal of Personality and social Psychology*, 54(4):638, 1988.

[144] Adam O Horvath and B Dianne Symonds. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology*, 38 (2):139, 1991.

[145] Machteld Huber, Marja van Vliet, M Giezenberg, B Winkens, Y Heerkens, PC Dagnelie, and JA Knottnerus. Towards a 'patient-centred' operationalisation of the new dynamic concept of health: a mixed methods study. *BMJ open*, 6(1), 2016.

[146] Lawrie Hunter. Text nouveau: visible structure in text presentation. *Computer Assisted Language Learning*, 11(4):363–379, 1998.

[147] Marian ZM Hurmuz, Stephanie M Jansen-Kosterink, Harm op den Akker, and Hermie J Hermens. User experience and potential health effects of a conversational agent-based electronic health intervention: Protocol for an observational cohort study. *JMIR research protocols*, 9(4):e16641, 2020.

[148] David Ireland, Christina Atay, Jacki Liddle, Dana Bradford, Helen Lee, Olivia Rushin, Thomas Mullins, Dan Angus, Janet Wiles, Simon McBride, et al. Hello harlie: enabling speech monitoring through chat-bot conversations. In *Digital Health Innovation for Consumers, Clinicians, Connectivity and Community*, pages 55–60. IOS Press, 2016.

[149] Kazi Faria Islam, Abdul Awal, Hoimonty Mazumder, Ummi Rukaiya Munni, Koushik Majumder, Kohinoor Afroz, Mustari Nailah Tabassum, and M Mahbub Hossain. Social cognitive theory-based health promotion in primary care practice: A scoping review. *Heliyon*, 9(4), 2023.

[150] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna van Riemsdijk, and Maarten Sierhuis. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1), 2014.

[151] W Lewis Johnson and James C Lester. Face-to-face interaction with pedagogical agents, twenty years later. *International Journal of Artificial intelligence in education*, 26:25–36, 2016.

[152] Lee W Jones, Robert C Sinclair, and Kerry S Courneya. The effects of source credibility and message framing on exercise intentions, behaviors, and attitudes: An integration of the elaboration likelihood model and prospect theory 1. *Journal of applied social psychology*, 33(1):179–196, 2003.

[153] Catholijn M. Jonker, M. Birna van Riemsdijk, and Bas Vermeulen. Shared mental models. In Marina De Vos, Nicoletta Fornara, Jeremy V. Pitt, and George Vouros, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems VI*, pages 132–151, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-21268-0.

[154] Lynn R Kahle. Social values and social change: Adaptation to life in america, 1983.

[155] Lynn R Kahle, Sharon E Beatty, and Pamela Homer. Alternative measurement approaches to consumer values: the list of values (lov) and values and life style (vals). *Journal of consumer research*, 13(3):405–409, 1986.

[156] Toshikazu Kanaoka and Bilge Mutlu. Designing a motivational agent for behavior change in physical activity. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1445–1450, 2015.

[157] Maurits Kaptein, Panos Markopoulos, Boris De Ruyter, and Emile Aarts. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, 77:38–51, 2015.

[158] John M Kearney and Sinead McElhone. Perceived barriers in trying to eat healthier–results of a pan-eu consumer attitudinal survey. *British Journal of Nutrition*, 81(S1): S133–S137, 1999.

[159] Saskia M Kelders, Robin N Kok, Hans C Ossebaard, and Julia EWC Van Gemert-Pijnen. Persuasive system design does matter: a systematic review of adherence to web-based interventions. *Journal of medical Internet research*, 14(6):e152, 2012.

[160] Andrea Seaton Kelton, Robin R Pennington, and Brad M Tuttle. The effects of information presentation format on judgment and decision making: A review of the information systems research. *Journal of Information Systems*, 24(2):79–105, 2010.

[161] Sung Yeun Su Kim, Nathan Prestopnik, and Frank A Biocca. Body in the interactive game: How interface embodiment affects physical activity and health behavior change. *Computers in Human Behavior*, 36:376–384, 2014.

[162] Jenny Kitzinger. Qualitative research: introducing focus groups. *Bmj*, 311(7000): 299–302, 1995.

[163] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. How to evaluate technologies for health behavior change in hci research. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3063–3072, 2011.

[164] Malte S Kließ, Marielle Stoelinga, and M Riemsdijk. From good intentions to behaviour change. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 354–369. Springer, 2019.

[165] Richard Klimoski and Susan Mohammed. Team mental model: Construct or metaphor? *Journal of management*, 20(2):403–437, 1994.

[166] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 conference on designing interactive systems*, pages 555–565, 2017.

[167] Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. The personalization of conversational agents in health care: systematic review. *Journal of medical Internet research*, 21(11):e15360, 2019.

[168] Ilir Kola, Catholijn M Jonker, and M Birna van Riemsdijk. Who's that?-social situation awareness for behaviour support agents: A feasibility study. In *International workshop on engineering multi-agent systems*, pages 127–151. Springer, 2019.

[169] Ilir Kola, Catholijn M Jonker, Myrthe L Tielman, and M Birna van Riemsdijk. Grouping situations based on their psychological characteristics gives insight into personal values. In *MRC@ ECAI*, pages 17–26, 2020.

[170] Ilir Kola, Catholijn M Jonker, and M Birna van Riemsdijk. Who's that?-social situation awareness for behaviour support agents: A feasibility study. In *Engineering Multi-Agent Systems: 7th International Workshop, EMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers*, pages 127–151. Springer, 2020.

[171] Ilir Kola, Pradeep K Murukannaiah, Catholijn M Jonker, and M Birna Van Riemsdijk. Toward social situation awareness in support agents. *IEEE intelligent systems*, 37(5): 50–58, 2022.

[172] Hubert Kolb and Stephan Martin. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. *BMC medicine*, 15(1):1–11, 2017.

[173] Wessel Kraaij, Suzan Verberne, Saskia Koldijk, Elsbeth de Korte, Saskia van Dantzig, Maya Sappelli, Muhammad Shoaib, Steven Bosems, Reinoud Achterkamp, Alberto Bonomi, et al. Personalized support for well-being at work: an overview of the swell project. *User modeling and user-adapted interaction*, 30:413–446, 2020.

[174] Jan-Niklas Kramer, Florian Künzler, Varun Mishra, Shawna N Smith, David Kotz, Urte Scholz, Elgar Fleisch, and Tobias Kowatsch. Which components of a smartphone walking app help users to reach personalized step goals? results from an optimization trial. *Annals of Behavioral Medicine*, 54(7):518–528, 2020.

[175] Stephen J Kraus. Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and social psychology bulletin*, 21(1):58–75, 1995.

[176] Paul Krebs, James O Prochaska, and Joseph S Rossi. A meta-analysis of computer-tailored interventions for health behavior change. *Preventive medicine*, 51(3-4):214–221, 2010.

[177] Richard A Krueger. *Focus groups: A practical guide for applied research.* Sage publications, 1994.

[178] Richard A Krueger. *Focus groups: A practical guide for applied research.* Sage publications, 2014.

[179] Quill R Kukla. Healthism, elite capture, and the pitfalls of an expansive concept of health. *A pragmatic approach to conceptualization of health and disease*, pages 275–294, 2024.

[180] Maria Knight Lapinski and Rajiv N Rimal. An explication of social norms. *Communication theory*, 15(2):127–147, 2005.

[181] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25:1248–1258, 2018.

[182] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018.

[183] Matthias Laschke, Sarah Diefenbach, Thies Schneider, and Marc Hassenzahl. Keymoment: Initiating behavior change through friendly friction. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, pages 853–858, 2014.

[184] Matthias Laschke, Sarah Diefenbach, and Marc Hassenzahl. "annoying, but in a nice way": an inquiry into the experience of frictional feedback. *International Journal of Design*, 9(2), 2015.

[185] Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479, 2024.

[186] Byung-Kwan Lee and Wei-Na Lee. The effect of information overload on consumer choice quality in an on-line environment. *Psychology & Marketing*, 21(3):159–183, 2004.

[187] Lin-shan Lee and Berlin Chen. Spoken document understanding and organization. *IEEE Signal Processing Magazine*, 22(5):42–60, 2005.

[188] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–35, 2019.

[189] Julie Lemarié, Hélène Eyrolle, and Jean-Marie Cellier. The segmented presentation of visually structured texts: Effects on text comprehension. *Computers in human behavior*, 24(3):888–902, 2008.

[190] Michael Lewis, Katia Sycara, and Phillip Walker. The role of trust in human-robot interaction. *Foundations of trusted autonomy*, pages 135–159, 2018.

[191] Yvonna S Lincoln, Egon G Guba, and JJ Pilotta. Naturalistic inquiry. beverly hills. *CA: Sage Publications. Lee, WS (2001). Parents divorce and their duty to support the expense of bringing up their child. Asian Women*, 13(1):85–105, 1985.

[192] Enrico Liscio, MT van der Meer, L Cavalcante Siebert, N Mouter, CM Jonker, and Pradeep K Murukannaiah. Axies: Identifying and evaluating context-specific values. In *20th International Conference on Autonomous Agentsand Multiagent Systems*, pages 799–808. International Foundation for Autonomous Agents and Multiagent Systems, 2021.

[193] Enrico Liscio, Michiel van der Meer, Luciano C Siebert, Catholijn M Jonker, and Pradeep K Murukannaiah. What values should an agent align with? an empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, 36(1):23, 2022.

[194] Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

[195] Ewa Luger and Abigail Sellen. " like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5286–5297, 2016.

[196] Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. *arXiv preprint arXiv:1708.09085*, 2017.

[197] Mia Liza A Lustria, Seth M Noar, Juliann Cortese, Stephanie K Van Stee, Robert L Glueckauf, and Junga Lee. A meta-analysis of web-delivered tailored health behavior change interventions. *Journal of health communication*, 18(9):1039–1069, 2013.

[198] Raju Maharjan, Darius Adam Rohani, Per Bækgaard, Jakob Bardram, and Kevin Doherty. Can we talk? design implications for the questionnaire-driven self-report of health and wellbeing via conversational agent. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, pages 1–11, 2021.

[199] Mohammed N Mahdi, Abdul R Ahmad, Roslan Ismail, Mohammed A Subhi, Mohammed M Abdulrazzaq, and Qais S Qassim. Information overload: The effects of large amounts of information. In *2020 1st. Information Technology To Enhance e-learning and Other Application (IT-ELA*, pages 154–159. IEEE, 2020.

[200] Marta M Marques, Alison J Wright, Elizabeth Corker, Marie Johnston, Robert West, Janna Hastings, Lisa Zhang, and Susan Michie. The behaviour change technique ontology: transforming the behaviour change technique taxonomy v1. *Wellcome open research*, 8:308, 2024.

[201] Laura Martinengo, Ahmad Ishqi Jabir, Westin Wei Tin Goh, Nicholas Yong Wai Lo, Moon-Ho Ringo Ho, Tobias Kowatsch, Rifat Atun, Susan Michie, and Lorainne Tudor Car. Conversational agents in health care: scoping review of their behavior change techniques and underpinning theory. *Journal of Medical Internet Research*, 24(10):e39243, 2022.

[202] Judith Masthoff and Julita Vassileva. Tutorial on personalization for behaviour change. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 439–442, 2015.

[203] Simon C Mathews, Michael J McShea, Casey L Hanley, Alan Ravitz, Alain B Labrique, and Adam B Cohen. Digital health: a path to validation. *NPJ digital medicine*, 2(1):38, 2019.

[204] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2):273, 2000.

[205] Sandra C Matz, Cristina Segalin, David Stillwell, Sandrine R Müller, and Maarten W Bos. Predicting the personal appeal of marketing images using computational methods. *Journal of consumer psychology*, 29(3):370–390, 2019.

[206] Nicholas Mays and Catherine Pope. Qualitative research: rigour and qualitative research. *Bmj*, 311(6997):109–112, 1995.

[207] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–23, 2019.

[208] Rosemary Robin Charlotte McEachan, Mark Conner, Natalie Jayne Taylor, and Rebecca Jane Lawton. Prospective prediction of health-related behaviours with the theory of planned behaviour: A meta-analysis. *Health psychology review*, 5(2):97–144, 2011.

[209] Michael F McTear. *Spoken dialogue technology: toward the conversational user interface*. Springer Science & Business Media, 2004.

[210] Michael Frederick McTear, Zoraida Callejas, and David Griol. *The conversational interface*, volume 6. Springer, 2016.

[211] Siddharth Mehrotra. Designing for appropriate trust in human-ai interaction. 2024.

[212] Marco Mensink, Ellen E Blaak, Eefje Corpeleijn, Wim H Saris, Tjerk W De Bruin, and Edith J Feskens. Lifestyle intervention according to general recommendations improves glucose tolerance. *Obesity research*, 11(12):1588–1596, 2003.

[213] Susan Michie, Charles Abraham, Craig Whittington, John McAteer, and Sunjai Gupta. Effective techniques in healthy eating and physical activity interventions: a meta-regression. *Health psychology*, 28(6):690, 2009.

[214] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P Eccles, James Cane, and Caroline E Wood. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine*, 46(1):81–95, 2013.

[215] Susan Miles and Gene Rowe. *The Laddering Technique*, pages 305 – 343. 02 2008. ISBN 9780470776278. doi: 10.1002/9780470776278.ch13.

[216] Eleonora Milić, Dragan Janković, and Aleksandar Milenković. Health care domain mobile reminder for taking prescribed medications. In *International Conference on ICT Innovations*, pages 173–181. Springer, 2016.

[217] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[218] Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, Edward Meinert, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of medical Internet research*, 22(10):e20346, 2020.

[219] Madison Milne-Ives, Ching Lam, Caroline De Cock, Michelle Helena Van Velthoven, Edward Meinert, et al. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. *JMIR mHealth and uHealth*, 8(3):e17046, 2020.

[220] Elliot Mitchell, Noemie Elhadad, and Lena Mamykina. Examining ai methods for micro-coaching dialogs. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–24, 2022.

[221] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679, 2016.

[222] Nataliya Mogles, Julian Padget, Elizabeth Gabe-Thomas, Ian Walker, and JeeHang Lee. A computational model for designing energy behaviour change interventions. *User modeling and user-adapted interaction*, 28:1–34, 2018.

[223] Ali R Montazemi and S Wang. The effects of modes of information presentation on decision-making: a review and meta-analysis. *Journal of Management Information Systems*, 5(3):101–127, 1988.

[224] Frederick Muench, Katherine van Stolk-Cooke, Jon Morgenstern, Alexis N Kuerbis, and Kendra Markle. Understanding messaging preferences to inform development of mobile goal-directed behavioral interventions. *Journal of Medical Internet Research*, 16(2):e14, 2014.

[225] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–36, 2019.

[226] Gera E Nagelhout, Lette Hogeling, Renate Spruijt, Nathalie Postma, and Hein De Vries. Barriers and facilitators for health behavior change among adults from multi-problem households: a qualitative study. *International journal of environmental research and public health*, 14(10):1229, 2017.

[227] Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000.

[228] Shreeya Nelekar, Amal Abdulrahman, Manik Gupta, and Deborah Richards. Effectiveness of embodied conversational agents for managing academic stress at an indian university (aru) during covid-19. *British Journal of Educational Technology*, 53(3):491–511, 2022.

[229] Efrat Neter and Esther Brainin. ehealth literacy: extending the digital divide to the realm of health information. *Journal of medical Internet research*, 14(1):e19, 2012.

[230] Robin Neuhaus, Ronda Ringfort-Felner, Shadan Sadeghian, and Marc Hassenzahl. To mimic reality or to go beyond¿‘superpowers” in virtual reality, the experience of augmentation and its consequences. *International Journal of Human-Computer Studies*, 181:103165, 2024.

[231] Johan YY Ng, Nikos Ntoumanis, Cecilie Thøgersen-Ntoumani, Edward L Deci, Richard M Ryan, Joan L Duda, and Geoffrey C Williams. Self-determination theory applied to health contexts: A meta-analysis. *Perspectives on psychological science*, 7 (4):325–340, 2012.

[232] Bostrom Nick. Superintelligence: Paths, dangers, strategies. 2014.

[233] Lasse Nielsen. Why health matters to justice: A capability theory perspective. *Ethical Theory and Moral Practice*, 18:403–415, 2015.

[234] Don Norman. *The design of everyday things: Revised and expanded edition.* Basic books, 2013.

[235] Lily O'Hara, Bayan Alajaimi, and Bayan Alshowaikh. "i was bullied for being fat in every situation, in every outfit, at every celebration": A qualitative exploratory study on experiences of weight-based oppression in qatar. *Frontiers in Public Health*, 11:1015181, 2023.

[236] Harri Oinas-Kukkonen. Behavior change support systems: A research model and agenda. In *Persuasive Technology: 5th International Conference, PERSUASIVE 2010, Copenhagen, Denmark, June 7-10, 2010. Proceedings 5*, pages 4–14. Springer, 2010.

[237] Harri Oinas-Kukkonen. A foundation for the study of behavior change support systems. *Personal and ubiquitous computing*, 17:1223–1235, 2013.

[238] Harri Oinas-Kukkonen and Marja Harjumaa. Communications of the association for information systems. *Communications of the Association for Information Systems*, 24:485–500, 2009.

[239] Jeanette M Olsen and Bonnie J Nesbitt. Health coaching to improve healthy lifestyle behaviors: an integrative review. *American journal of health promotion*, 25(1):e1–e12, 2010.

[240] Harm op den Akker, Val Jones, and Hermie Hermens. Predicting feedback compliance in a teletreatment application. In *2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2010)*, pages 1–5. IEEE, 2010.

[241] Harm op den Akker, Valerie M Jones, and Hermie J Hermens. Tailoring real-time physical activity coaching systems: a literature survey and model. *User modeling and user-adapted interaction*, 24:351–392, 2014.

[242] Rita Orji and Karyn Moffatt. Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health informatics journal*, 24(1):66–91, 2018.

[243] Rita Orji, Julita Vassileva, and Regan L Mandryk. Modeling the efficacy of persuasive strategies for different gamer types in serious games for health. *User Modeling and User-Adapted Interaction*, 24:453–498, 2014.

[244] Stuart Oskamp. Overconfidence in case-study judgments. *Journal of consulting psychology*, 29(3):261, 1965.

[245] Deger Ozkaramanli, Elif Özcan, and Pieter Desmet. Long-term goals or immediate desires? introducing a toolset for designing with self-control dilemmas. *The Design Journal*, 20(2):219–238, 2017.

[246] Charuta Pande, Andreas Martin, and Christoph Pimmer. Towards hybrid dialog management strategies for a health coach chatbot. 2023.

[247] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. Designing a chatbot for a brief motivational interview on stress management: qualitative case study. *Journal of medical Internet research*, 21(4):e12231, 2019.

[248] Pietro Pasotti, Catholijn M Jonker, and M Birna van Riemsdijk. Action identification hierarchies for behaviour support agents. In *Third Workshop on Cognitive Knowledge Acquisition and Applications, Cognitum*, 2017.

[249] Heather Patrick, Ken Resnicow, Pedro J Teixeira, and Geoffrey C Williams. Communication skills to elicit physical activity behavior change: How to talk to the client. *ACSM's behavioral aspects of physical activity and exercise*, page 129, 2013.

[250] Michael Quinn Patton. *Qualitative evaluation and research methods.* SAGE Publications, inc, 1990.

[251] Michael Quinn Patton. Enhancing the quality and credibility of qualitative analysis. *Health services research*, 34(5 Pt 2):1189, 1999.

[252] Michael Quinn Patton. *Qualitative research & evaluation methods.* sage, 2002.

[253] Wei Peng, Shaheen Kanthawala, Shupei Yuan, and Syed Ali Hussain. A qualitative study of user perceptions of mobile health apps. *BMC public health*, 16:1–11, 2016.

[254] Juanan Pereira and Óscar Díaz. Using health chatbots for behavior change: a mapping study. *Journal of medical systems*, 43:1–13, 2019.

[255] Pierre Philip, Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne De Sevin, Jérôme Olive, Stéphanie Bioulac, and Alain Sauteraud. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific reports*, 7(1):42656, 2017.

[256] Martin J Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190, 2004.

[257] Alina Pommeranz, Christian Detweiler, Pascal Wiggers, and Catholijn M Jonker. Self-reflection on personal values to support value-sensitive design. In *Proceedings of HCI 2011 The 25th BCS Conference on Human Computer Interaction*. BCS Learning & Development, 2011.

[258] James O Prochaska and Carlo C DiClemente. Stages and processes of self-change of smoking: toward an integrative model of change. *Journal of consulting and clinical psychology*, 51(3):390, 1983.

[259] James O Prochaska and Wayne F Velicer. The transtheoretical model of health behavior change. *American journal of health promotion*, 12(1):38–48, 1997.

[260] Cozumel S Pruette and Sandra Amaral. Empowering patients to adhere to their treatment regimens: a multifaceted approach. *Pediatric transplantation*, 25(1):e13849, 2021.

[261] Fatemeh Rabiee. Focus-group interview and data analysis. *Proceedings of the nutrition society*, 63(4):655–660, 2004.

[262] Iyad Rahwan. Society-in-the-loop: programming the algorithmic social contract. *Ethics and information technology*, 20(1):5–14, 2018.

[263] Hedieh Ranjbartabar, Deborah Richards, Ayse Aysin Bilgin, and Cat Kutay. First impressions count! the role of the human's emotional state on rapport established with an empathic versus neutral virtual therapist. *IEEE transactions on affective computing*, 12(3):788–800, 2019.

[264] Hedieh Ranjbartabar, Deborah Richards, Ayse Aysin Bilgin, and Cat Kutay. Personalising the dialogue of relational agents for first-time users. In *AAMAS*, pages 1622–1624, 2021.

[265] George E Raptis, Christos Fidas, Christina Katsini, and Nikolaos Avouris. A cognition-centered personalization framework for cultural-heritage content. *User Modeling and User-Adapted Interaction*, 29:9–65, 2019.

[266] Urša Reja, Katja Lozar Manfreda, Valentina Hlebec, and Vasja Vehovar. Open-ended vs. close-ended questions in web questionnaires. *Developments in applied statistics*, 19(1):159–177, 2003.

[267] Deborah Richards and Patrina Caldwell. Improving health outcomes sooner rather than later via an interactive website and virtual specialist. *IEEE journal of biomedical and health informatics*, 22(5):1699–1706, 2017.

[268] Deborah Richards, Paulo Sergio Miranda Maciel, and Heidi Janssen. The co-design of an embodied conversational agent to help stroke survivors manage their recovery. *Robotics*, 12(5):120, 2023.

[269] Deborah Richards, Ravi Vythilingam, and Paul Formosa. A principlist-based study of the ethical design and acceptability of artificial social agents. *International Journal of Human-Computer Studies*, 172:102980, 2023.

[270] Tim Rietz and Alexander Maedche. Ladderbot: A requirements self-elicitation system. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 357–362. IEEE, 2019.

[271] Tim Rietz and Alexander Maedche. Ladderbot—a conversational agent for human-like online laddering interviews. *International Journal of Human-Computer Studies*, 171:102969, 2023.

[272] Lee M Ritterband, Frances P Thorndike, Daniel J Cox, Boris P Kovatchev, and Linda A Gonder-Frederick. A behavior change model for internet interventions. *Annals of Behavioral Medicine*, 38(1):18–27, 2009.

[273] Milton Rokeach. The nature of human values. *Free peess*, 1973.

[274] Jamie Ross, Fiona Stevenson, Rosa Lau, and Elizabeth Murray. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). *Implementation science*, 11:1–12, 2016.

[275] Cheryl Roumen, Ellen E Blaak, and Eva Corpeleijn. Lifestyle intervention for prevention of diabetes: determinants of success for future implementation. *Nutrition Reviews*, 67(3):132–146, 2009. ISSN 0029-6643. doi: 10.1111/j.1753-4887.2009.00181.x. URL `https://doi.org/10.1111/j.1753-4887.2009.00181.x`.

[276] Francesco Rubino, Rebecca M Puhl, David E Cummings, Robert H Eckel, Donna H Ryan, Jeffrey I Mechanick, Joe Nadglowski, Ximena Ramos Salas, Phillip R Schauer, Douglas Twenefour, et al. Joint international consensus statement for ending stigma of obesity. *Nature medicine*, 26(4):485–497, 2020.

[277] Gordon Rugg, Malcolm Eva, Atiya Mahmood, Nazia Rehman, Stephanie Andrews, and Sarah Davies. Eliciting information about organizational culture via laddering. *Information systems journal*, 12(3):215–229, 2002.

[278] Stuart Russell. *Human compatible: AI and the problem of control.* Penguin Uk, 2019.

[279] Stuart Russell. Artificial intelligence and the problem of control. *Perspectives on Digital Humanism*, 19:1–322, 2022.

[280] Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI magazine*, 36(4):105–114, 2015.

[281] Richard M Ryan and Edward L Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68, 2000.

[282] Bakhtiar Sadeghi, Deborah Richards, Paul Formosa, Mitchell McEwan, Muhammad Hassan Ali Bajwa, Michael Hitchens, and Malcolm Ryan. Modelling the ethical priorities influencing decision-making in cybersecurity contexts. *Organizational Cybersecurity Journal: Practice, Process and People*, 3(2):127–149, 2023.

[283] Malak Sadek, Rafael A Calvo, and Céline Mougenot. Designing value-sensitive ai: a critical review and recommendations for socio-technical design processes. *AI and Ethics*, 4(4):949–967, 2024.

[284] Sana Salman, Deborah Richards, and Mark Dras. Identifying which relational cues users find helpful to allow tailoring of e-coach dialogues. *Multimodal Technologies and Interaction*, 7(10):93, 2023.

[285] Gro Beate Samdal, Geir Egil Eide, Tom Barth, Geoffrey Williams, and Eivind Meland. Effective behaviour change techniques for physical activity and healthy eating in overweight and obese adults; systematic review and meta-regression analyses. *International Journal of Behavioral Nutrition and Physical Activity*, 14:1–14, 2017.

[286] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Data and discrimination: Collected essays*, pages 6–10. Open Technology Institute, 2014.

[287] Steven P Schnaars. How to develop and use scenarios. *Long range planning*, 20(1): 105–114, 1987.

[288] Danielle E Schoffman, Gabrielle Turner-McGrievy, Sonya J Jones, and Sara Wilcox. Mobile apps for pediatric obesity prevention and treatment, healthy eating, and physical activity promotion: just fun and games? *Translational behavioral medicine*, 3(3):320–325, 2013.

[289] Shalom H Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology/Academic Press*, 1992.

[290] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.

[291] Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663, 2012.

[292] Kyarash Shahriari and Mana Shahriari. Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pages 197–201. IEEE, 2017.

[293] Tait D Shanafelt, Lotte N Dyrbye, Colin P West, and Christine A Sinsky. Potential impact of burnout on the us physician workforce. In *Mayo Clinic Proceedings*, volume 91, pages 1667–1668. Elsevier, 2016.

[294] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[295] Gyri Skoglund, Birgitta Blakstad Nilsson, Cecilie Fromholt Olsen, Astrid Bergland, and Gunvor Hilde. Facilitators and barriers for lifestyle change in people with prediabetes: A meta-synthesis of qualitative studies. *BMC Public Health*, 22(1):1–27, 2022.

[296] Matthew J Spaniol and Nicholas J Rowland. Defining scenario. *Futures & Foresight Science*, 1(1):e3, 2019.

[297] Sarah Spiekermann and Till Winkler. Value-based engineering for ethics by design. *arXiv preprint arXiv:2004.13676*, 2020.

[298] Janet E Squires, Katrina Sullivan, Martin P Eccles, Julia Worswick, and Jeremy M Grimshaw. Are multifaceted interventions more effective than single-component interventions in changing health-care professionals' behaviours? an overview of systematic reviews. *Implementation Science*, 9:1–22, 2014.

[299] Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12, 2024.

[300] Anselm Strauss and Juliet Corbin. *Basics of qualitative research techniques*. Citeseer, 1998.

[301] Anselm Strauss, Juliet Corbin, et al. *Basics of qualitative research*, volume 15. sage Newbury Park, CA, 1990.

[302] Alycia N Sullivan and Margie E Lachman. Behavior change with fitness technology in sedentary adults: a review of the evidence for increasing physical activity. *Frontiers in public health*, 4:289, 2017.

[303] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one*, 12(8):e0182151, 2017.

[304] Lilu Tang, Peijun Yuan, and Dan Zhang. Emotional experience during human-computer interaction: a survey. *International Journal of Human–Computer Interaction*, 40(8):1845–1855, 2024.

[305] June Price Tangney, Rhonda L Dearing, Patricia E Wagner, and Richard Gramzow. Test of self-conscious affect–2. 1996.

[306] June Price Tangney, Jeff Stuewig, and Debra J Mashek. Moral emotions and moral behavior. *Annu. Rev. Psychol.*, 58:345–372, 2007.

[307] Niels ter Heijden and Willem-Paul Brinkman. Design and evaluation of a virtual reality exposure therapy system with automatic free speech interaction. *Journal of CyberTherapy and Rehabilitation*, 4(1):41–55, 2011.

[308] David R Thomas. A general inductive approach for qualitative data analysis. 2003.

[309] Myrthe L Tielman, Catholijn M Jonker, and M Birna van Riemsdijk. What should i do? deriving norms from actions, values and context. In *MRC@ IJCAI*, 2018.

[310] Dereck Toker, Cristina Conati, and Giuseppe Carenini. Gaze analysis of user characteristics in magazine style narrative visualizations. *User Modeling and User-Adapted Interaction*, 29(5):977–1011, 2019.

[311] Jessica L Tracy and Richard W Robins. " putting the self into self-conscious emotions: A theoretical model". *Psychological inquiry*, 15(2):103–125, 2004.

[312] David R Traum and Elizabeth A Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational intelligence*, 8(3):575–599, 1992.

[313] Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. Conversational agents in health care: scoping review and conceptual analysis. *Journal of medical Internet research*, 22(8):e17158, 2020.

[314] Jaakko Tuomilehto, Peter Schwarz, and Jaana Lindström. Long-term benefits from lifestyle interventions for type 2 diabetes prevention: time to expand the efforts. *Diabetes care*, 34(Suppl 2):S210, 2011.

[315] Ibo Van de Poel. Translating values into design requirements. *Philosophy and engineering: Reflections on practice, principles and process*, pages 253–266, 2013.

[316] Piet Van den Bossche, Wim Gijselaers, Mien Segers, Geert Woltjer, and Paul Kirschner. Team learning: building shared mental models. *Instructional science*, 39:283–301, 2011.

[317] Jeroen Van den Hoven, Pieter E Vermaas, and Ibo Van de Poel. Handbook of ethics, values and technological design. *Handbook of Ethics, Values, and Technological Design*, 2015.

[318] DJ Annemarie van Dijk, Matty R Crone, Pepijn van Empelen, Willem J Assendelft, and Barend J Middelkoop. Favourable outcomes of a preventive screening and counselling programme for older people in underprivileged areas in the netherlands: The primus project. *Preventive medicine reports*, 6:258–264, 2017.

[319] Luke Van Rhoon, Molly Byrne, Eimear Morrissey, Jane Murphy, and Jenny McSharry. A systematic review of the behaviour change techniques and digital features in technology-driven type 2 diabetes prevention interventions. *Digital health*, 6:2055207620914427, 2020.

[320] M Birna Van Riemsdijk, Catholijn M Jonker, and Victor Lesser. Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems*, pages 1201–1206, 2015.

[321] Vero Vanden Abeele, Erik Hauters, and Bieke Zaman. Increasing the reliability and validity of quantitative laddering data with ladderux. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 2057–2062. 2012.

[322] Bart Verheij. Formalizing value-guided argumentation for ethical systems design. *Artificial Intelligence and Law*, 24(4):387–407, 2016.

[323] Giampaolo Viglia, Sara Dolnicar, Diletta Acuti, and Juan Luis Nicolau. If you want to learn about real behaviour, measure real behaviour. *Journal of Sustainable Tourism*, 32(11):2245–2257, 2024.

[324] Thymen René Wabeke. Recommending tips that support well-being at work to knowledge workers. *Master's thesis, Radboud University, Nijmegen, The Netherlands*, 2014.

[325] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*, 2019.

[326] Thomas Webb, Judith Joseph, Lucy Yardley, Susan Michie, et al. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *Journal of medical Internet research*, 12(1):e1376, 2010.

[327] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[328] Heidi Kristina Westli, Bjørn Helge Johnsen, Jarle Eid, Ingvil Rasten, and Guttorm Brattebø. Teamwork skills, shared mental models, and performance in simulated trauma teams: an independent group design. *Scandinavian journal of trauma, resuscitation and emergency medicine*, 18:1–8, 2010.

[329] Emily D Williams, Dominique Bird, Andrew W Forbes, Anthony Russell, Susan Ash, Robert Friedman, Paul A Scuffham, and Brian Oldenburg. Randomised controlled trial of an automated, interactive telephone intervention (tlc diabetes) to improve type 2 diabetes management: baseline findings and six-month outcomes. *BMC public health*, 12:1–11, 2012.

[330] Alan FT Winfield and Marina Jirotka. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133): 20180085, 2018.

[331] Rena R Wing, Michael G Goldstein, Kelly J Acton, Leann L Birch, John M Jakicic, James F Sallis Jr, Delia Smith-West, Robert W Jeffery, and Richard S Surwit. Behavioral science research in diabetes: lifestyle changes related to obesity, eating behavior, and physical activity. *Diabetes care*, 24(1):117–123, 2001.

[332] Till Winkler and Sarah Spiekermann. Twenty years of value sensitive design: a review of methodological practices in vsd projects. *Ethics and Information Technology*, 23:17–21, 2021.

[333] Bodo Winter. A very basic tutorial for performing linear mixed effects analyses. *arXiv preprint arXiv:1308.5499*, pages 1–22, 2013.

[334] Michael S Wogalter and Eric F Shaver. Evaluation of list vs. paragraph text format on search time for warning symptoms in a product manual. *Advances in occupational ergonomics and safety*, 4:434–438, 2001.

[335] Johanna Wolff, Victor De Boer, Dirk Heylen, and M Birna Van Riemsdijk. Using default logic to create adaptable user models for behavior support agents. In *3rd International Conference on Hybrid Human-Artificial Intelligence, HHAI 2024*, pages 350–359. IOS, 2024.

[336] Kevin T Wynne and Joseph B Lyons. An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, 19(3):353–374, 2018.

[337] Kevin T Wynne and Joseph B Lyons. Autonomous agent teammate-likeness: scale development and validation. In *International Conference on Human-Computer Interaction*, pages 199–213. Springer, 2019.

[338] Juan Ye, Simon Dobson, and Susan McKeever. Situation identification techniques in pervasive computing: A review. *Pervasive and mobile computing*, 8(1):36–66, 2012.

[339] Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: intervention

using a reinforcement learning system. *Journal of medical Internet research*, 19(10): e338, 2017.

[340] Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. Fusing task-oriented and open-domain dialogues in conversational agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11622–11629, 2022.

[341] Jennifer Zamora. Rise of the chatbots: Finding a place for artificial intelligence in india and us. In *Proceedings of the 22nd international conference on intelligent user interfaces companion*, pages 109–112, 2017.

[342] Jingwen Zhang, Yoo Jung Oh, Patrick Lange, Zhou Yu, and Yoshimi Fukuoka. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. *Journal of medical Internet research*, 22(9):e22845, 2020.

[343] Yangyang Zhao, Zhenyu Wang, Changxi Zhu, and Shihan Wang. Efficient dialogue complementary policy learning via deep q-network policy and episodic memory policy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4311–4323, 2021.

[344] Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, Ben Gerber, Nikolaos Agadakos, and Shweta Yadav. Towards enhancing health coaching dialogue in low-resource settings. *arXiv preprint arXiv:2404.08888*, 2024.

[345] Ingrid Zukerman and David W Albrecht. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11:5–18, 2001.

[346] Ingrid Zukerman, David W Albrecht, and Ann E Nicholson. Predicting users' requests on the www. In *UM99 User Modeling: Proceedings of the Seventh International Conference*, pages 275–284. Springer, 1999.

SIKS Dissertations

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UvA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web

19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UvA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UvA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media

28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning

30 Ruud Mattheij (TiU), The Eyes Have It

31 Mohammad Khelghati (UT), Deep web content monitoring

32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example

34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment

35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation

36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UvA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UvA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VUA), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30 Wilma Latuny (TiU), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VUA), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TU/e), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
30 Wouter Beek (VUA), The ”K” in ”semantic web” stands for ”knowledge”: scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VUA), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

	35	Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
	36	Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
	37	Jian Fang (TUD), Database Acceleration on FPGAs
	38	Akos Kadar (OU), Learning visually grounded and multilingual representations
2020	01	Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
	02	Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
	03	Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
	04	Maarten van Gompel (RUN), Context as Linguistic Bridges
	05	Yulong Pei (TU/e), On local and global structure mining
	06	Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
	07	Wim van der Vegt (OU), Towards a software architecture for reusable game components
	08	Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
	09	Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
	10	Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
	11	Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
	12	Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
	13	Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
	14	Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
	15	Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
	16	Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
	17	Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
	18	Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
	19	Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
	20	Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
	21	Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
	22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
	23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
	24	Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
	25	Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
	26	Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
	27	Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
	28	Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
	29	Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
	30	Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
	31	Gongjin Lan (VUA), Learning better – From Baby to Better
	32	Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
	33	Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
	34	Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
	35	Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
2021	01	Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
07 Armel Lefebvre (UU), Research data management for open science
08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (UL), Robust rules for prediction and description
24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge

14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research
17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision –From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals

20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
04 Mike Huisman (UL), Understanding Deep Meta-Learning
05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
19 Fang Hou (UU), Trust in Software Ecosystems
20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making

23 Roderick van der Weerdt (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning
29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering
35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment
36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing
37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval
38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity
39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence
40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans
41 Mozhgan Vazifehdoostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights
42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data
43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education
44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning
45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings
46 Ya Song (TU/e), Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making
47 Tom Kouwenhoven (UL), Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions
48 Evy van Weelden (TiU), Integrating Virtual Reality and Neurophysiology in Flight Training
49 Selene Báez Santamaría (VUA), Knowledge-centered conversational agents with a drive to learn
50 Lea Krause (VUA), Contextualising Conversational AI
51 Jiaxu Zhao (TU/e), Understanding and Mitigating Unwanted Biases in Generative Language Models
52 Qiao Xiao (TU/e), Model, Data and Communication Sparsity for Efficient Training of Neural Networks
53 Gaole He (TUD), Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems
54 Go Sugimoto (VUA), MISSING LINKS Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities
55 Sietze Kai Kuilman (TUD), AI that Glitters is Not Gold: Requirements for Meaningful Control of AI Systems
56 Wijnand van Woerkom (UU), A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law
57 Syeda Amna Sohail (UT), Privacy-Utility Trade-Off in Healthcare Metadata Sharing and Beyond: A Normative and Empirical Evaluation at Inter and Intra Organizational Levels
58 Junhan Wen (TUD), "From iMage to Market": Machine-Learning-Empowered Fruit Supply

59 Mohsen Abbaspour Onari (TU/e), From Explanation to Trust: Modeling and Measuring Trust in Explainable Decision Support
60 Marcel Jurriaan Robeer (UU), Beyond Trust: A Causal Approach to Explainable AI in Law Enforcement
61 Shuai Wang (VUA), Links in Large Integrated Knowledge Graphs: Analysis, Refinement, and Domain Applications
62 Khaleel Asyraaf Mat Sanusi (OU), Augmenting a learning model within immersive learning environments for psychomotor skills
63 Rashid Zaman (TU/e), Online Conformance Checking on Degraded Data
64 Jens d'Hondt (TU/e), Effective and Efficient Multivariate Similarity Search
65 Aswin Balasubramaniam (UT), Disentangling Runner Drone Interaction Potentialities

2026 1 Pei-Yu Chen (TUD), Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support

Acknowledgments

I have been fortunate to have had wonderful companions during my PhD journey, each supporting, guiding, and teaching me in their own way, making this journey possible and enjoyable.

I am grateful to my supervisors for their guidance. Birna and Myrthe, I could not have asked for better mentors. Your compassion and empathy, along with your professional integrity, were important to me throughout this process. Myrthe, thank you for the time spent brainstorming and for your commitment to monitoring my progress. Your empathetic nature and constant support was vital during difficult moments, and because of you, working on this PhD never felt like something I was doing alone. Birna, thank you for teaching me the nuances of research ethics and for granting me the freedom to explore my own ideas and follow my intuitions. You provided the necessary guidance exactly where I needed it most, and your ability to contribute effectively made our collaboration incredibly smooth and rewarding. Catholijn, you have been an excellent guide since the day I walked into your office. You inspired me to become a scientist and taught me the value of collaborating across fields. Through you, I came to understand the true essence of a passion for science. Dirk, although our meetings were less frequent, your insights always left a lasting impression. I am continually impressed by your ability to forge unexpected connections and capture the core message.

I would also like to extend my sincere thanks to the members of my doctoral committee. Odette, Shenghui, and Bart, I truly appreciate your interest in my research and the time you dedicated to reviewing this thesis, providing insightful feedback, and participating in the defense. A very special thank you goes to Deborah. I am deeply moved by the warmth and hospitality you and Peter showed in “adopting” Nele and me during our visit to Australia. Your kindness made us feel truly at home and is a memory I will always cherish.

My journey would not have been the same without my colleagues from the Interactive Intelligence group. Carolina, thank you for always giving the best advice and support throughout the years, and for not giving up on befriending me. Linyun, you are exactly the too-lively, high-presence friend I need. Mo, thank you for being my go-to double coder. Ruben, our office is incomplete without you. Nele, I will never forget our time traveling through Australia together. Sid, thank you for the insights into interfaces and HCI. Morita, I truly enjoyed our conversations during those conference trips when we shared a room. To the rest of my II colleagues, Mani, Masha, Enrico, Zuzanna, Amir, Laxmi, Paul, Deborah, Emma, Micha, Sietze, Antonio, Rolf, Miguel, Elena, Ilir (thanks for the tips along the way), Willem-Paul (thank you for subbing for Catholijn), Mark (thank you for your availability as a reserve member), Catha, Yanzhe, Jinke, Tina, Stephanie, Agnes, Pradeeep, Luciano, thank you for the coffee chats, the hallway conversations, and the many lunches we shared. Thanks also to the support staff: Ruud, Wouter, and especially Anita, the most invaluable member of the group.

The Hybrid Intelligence consortium has been my extended academic family. Johanna, thank you for being the reasoning master. Selene, thank you for always being a sweet person. Floris, thank you for your technical expertise. Our demo would never have been completed without you. Loan, I loved our long walks and chats in Vlieland. I also thank the other HI members: Cor, Ludi, Michiel, Bram, Annet, Delaram, Chirag, Tae, Davide, Jasper, Urja, Tiffany, Lea, Íñigo, and Merle, and all other present and past HI members. Thank you, Frank and Wendy, for the patient organization.

I also wish to thank my colleagues at TNO. Quirine, you were the best teammate I could ask for in making the demo come to life. I valued our countless discussions on the use case. Maaike, thanks for making me feel so welcome during my visit. Tjeerd and Sofie, thank you for your help with the expert interviews. Additionally, I would like to thank Shaad from the REIT team for your help with the technical setup and for containerizing the demo.

To my friends, thank you for being my anchor during this journey. Fish, thank you for being the friend who always understands my struggles and stands by me no matter what. I also truly appreciate you and Michiel for helping me move during such a difficult time. Tzu-Yen,it means so much that we stay in touch even though we live so far apart. Lee Peng, thank you for so many times having me for dinner after work, and for agreeing to help me without any hesitation. The cover design is amazing. Charlotte, for the nights we cried together—your support truly saved me during my darkest times. Jess, Yu Heng, and Queenie, thank you for your impeccable tastes in things and for having bizarre friends so that I always have strange stories to hear, and the spontaneous plans and impulsive purchases that make life more interesting. Cinco, Wei-Hsuan, thank you for all the help with my car. Hugo and Yoshi are the chillest babies I have ever met. Emeline, your advice and encouragement always gives me a sense of peace. Zach, thank you for being unexpectedly hilarious during our gatherings. Justin, thank you for being so thoughtful. Hong-Ui, thank you for the deep discussions, even if they did not always make me feel better. Esther, you were the best housemate one could ask for during PhD. Thomas, thank you for telling me to make the slides before I started my new job, which I didn't do, and thank you for always reminding me how proud you are of me.

爸爸媽媽妹妹: 謝謝你們給的支持，讓我有足夠的底氣不用委屈自己。謝謝你們讓我可以當爸寶媽寶 & 妹寶。

Curriculum Vitae

Pei-Yu Chen

June 30 1996 Born in Kaohsiung City, Taiwan.

Education

2021 – 2025 **Ph.D. in Computer Science**
Delft University of Technology, Delft, the Netherlands
Promotors: Prof. dr. C.M. Jonker and Prof. dr. D.K.J. Heylen
Copromotors: Dr. M.L. Tielman and Dr. M.B. van Riemsdijk

2018 – 2020 **Master of Science in Human-Machine Communication**
University of Groningen, Groningen, The Netherlands

2014 – 2018 **Bachelor of Science in Psychology**
National Cheng Kung University, Tainan, Taiwan

Experience

2025 – Present **Diamond Flower Information B.V.**, Rotterdam, the Netherlands
Project manager

2023 **Dutch Organization for Applied Scientific Research (TNO)**, The Hague & Soesterberg, The Netherlands
Visiting Researcher

2021 – 2025 **Delft University of Technology**, Delft, the Netherlands
Ph.D. Candidate

2020 **Mentech Innovation B.V.**, Eindhoven, the Netherlands
Data Scientist Intern

List of Publications

Under review

1. **Pei-Yu Chen**, M. Birna van Riemsdijk, and Myrthe L. Tielman. Presenting User Behavior Information Collected by a Conversational Agent. *Under review.*

2024

1. Bernd J.W. Dudzik, Jasper S. van der Waa, **Pei-Yu Chen**, Roel Dobbe, Íñigo M.D.R. de Troya, Roos M. Bakker, Maaike H.T. de Boer, Quirine Smit, Davide Dell'Anna, Emre Erdogan, Pinar Yolum, Shihan Wang, Selene Baez Santamaria, Lea Krause, and Bart A. Kamphorst. Hybrid intelligence supports application development for diabetes lifestyle management. Journal of Artificial Intelligence Research 80 (2024), pages 919-929. doi: 10.1613/jair.1.15916.
2. **Pei-Yu Chen**, Selene Baez Santamaria, Maaike H.T. de Boer, Floris Den Hengst, Bart A. Kamphorst, Quirine Smit, Shihan Wang, and Johanna Wolff. Intelligent support systems for lifestyle change: integrating dialogue, information extraction, and reasoning. In *3rd International Conference on Hybrid Human-Artificial Intelligence, HHAI 2024*, pages 457–459, IOS Press, 2024. doi: 10.3233/FAIA240223.
3. **Pei-Yu Chen**, Sophie van Gent, M. Birna van Riemsdijk, Myrthe L. Tielman, and Tjeerd Schoonderwoerd. Expert insights on conversational AI systems as an information intermediary for patients and healthcare providers for diabetes lifestyle change. In Kiemute Oyibo, Wenzhen Xu, and Elena Vlahu-Gjorgievska, editors, *19th International Conference on Persuasive Technology, Adjunct Proceedings co-located with PERSUASIVE 2024, Wollongong, Australia, April 10th – 12th, 2024*, pages 8-20, volume 3728 of *CEUR Workshop Proceedings*, 2024.
4. **Pei-Yu Chen**, Selene Baez Santamaria, Maaike H.T. de Boer, Floris Den Hengst, Bart A. Kamphorst, Quirine Smit, Shihan Wang, and Johanna Wolff. Harnessing Hybrid Intelligence to Improve Diabetes Care. In *3rd International Conference on Hybrid Human-Artificial Intelligence, HHAI 2024*, Demo accepted and presented; not included in proceedings.

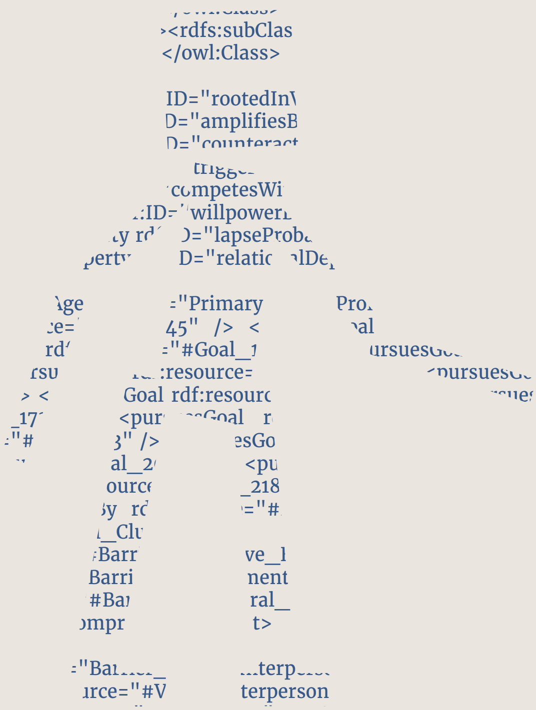
2023

1. **Pei-Yu Chen**, Myrthe L. Tielman, Dirk K.J. Heylen, Catholijn M. Jonker, and M. Birna Van Riemsdijk. Acquiring semantic knowledge for user model updates via human-agent alignment dialogues. In *HHAI 2023: Augmenting Human Intellect*, pages 93–107, IOS Press, 2023. doi: 10.3233/FAIA230077.

2022

1. **Pei-Yu Chen**. AI alignment dialogues: An interactive approach to AI alignment in support agents. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pages 894–894. doi: 10.1145/3514094.3539531.

Included in this thesis.



Cover by Peng Lee