Designing Engaging Conversational Interactions for Health & Behavior Change

Rafal Kocielnik

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Gary Hsieh, Chair

Daniel Avrahami

James Fogarty

Hannaneh Hajishirzi

Program Authorized to Offer Degree:
Human Centered Design & Engineering

University of Washington

Abstract

Designing Engaging Conversational Interactions for Health & Behavior Change

Rafal Kocielnik

Chair of the Supervisory Committee:
Associate Professor Gary Hsieh
Human Centered Design and Engineering

The recent popularity of chat and voice-based conversational interactions fueled by advances in natural language processing (NLP) has opened up opportunities for re-imagining user interactions in health & behavior change as conversational experiences. Prior work has indicated that a well-designed conversational approach can be more engaging, motivating, natural, personal, and understandable. It can also mimic the properties of some of the most successful human-led interventions, such as coaching and motivational interviewing. However, designing conversational interactions poses numerous challenges. Efficiently creating conversational content that is diverse, relevant for the context, and sounds natural is challenging. Furthermore, balancing the still limited AI capabilities with user expectations requires careful problem scoping and other design considerations. Finally, the mechanisms in which a successful conversational interaction can help improve user engagement are still not well explored.

In this dissertation I propose 4 different conversational systems that address some of the fundamental health & behavior change challenges. In Chapter 3 to address the intrinsic challenge of user boredom and engagement loss with repeated interactions - I propose a conversational system with value-based conversation topic personalization and diversification. In Chapter 4 to address the challenge of engaging users in mindful self-learning from their behavioral data - I propose conversational systems supporting structured reflection on

physical activity and on professional development at work. In Chapter 5 to support health data collection, especially to improve user comfort in sensitive topics and understandability among low-literacy populations - I propose a system for conversational survey administration. Finally in Chapter 6, to lower the effort involved in designing good quality conversational systems, I propose a tool for automated conversion of form-based surveys to a more engaging conversational format.

My work identifies and provides evidence for several benefits of the use of conversational interactions in health & behavior change. Among others, I demonstrate the benefits of increased engagement in interaction, improved motivation for performing activities, accessibility benefits related to familiarity, ease of use, comfort with sharing, and an ability to guide the users in the behavior change process via dialogue. I also identify several important challenges: perceptions of artificiality, managing high expectations of contextual knowledge, and social intelligence, as well as lower efficiency that could negatively affect the experience for some user groups. I further investigate the concrete links between conversational design elements and these benefits and challenges. My thesis demonstrates various design processes and automation techniques that can lower the effort of designing conversational experiences. As technology progresses conversational interactions can offer valuable support complimenting the existing automated tracking and the efforts of human health coaches. My work offers an important contribution to our understanding of how conversational interactions can play such a beneficial role.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

There are numerous people who greatly contributed to this dissertation and to my growth as a researcher in various ways. I will try to express my gratitude to all of them hoping I can be forgiven if I missed anybody's name.

I would first like to thank my advisor and mentor Gary Hsieh, for his guidance, understanding, and continued support through various challenges. I will always be grateful for numerous valuable skills I have learned from him and for his support in helping me understand the important first principles of research. I also want to thank him for allowing me to entertain my curiosity and ideas in other research areas, even if they sometimes expanded beyond his numerous areas of expertise. I am very thankful for his patience and tolerance of my, sometimes, slow progress and for never giving up on me.

I would also like to thank all my dissertation committee members for their support and sharing their expertise with me. I would like to thank Daniel Avrahami for helping me think about the implications of my research for practice beyond academia and for our collaborations on several exciting projects. I would like to thank Hannaneh Hajishirzi for sharing her expertise on modern AI which helped me understand the important trade-offs in applying such techniques in the HCI context. I would like to thank James Fogarty for helping me understand the technical HCI work and his guidance on communicating my research in a succinct, yet precise manner. A very valuable skill when having to pitch my research to new audiences. Finally, I would like to thank my GSR Dan Weld for asking insightful and thought provoking questions which helped me understand new facets of my work.

I would like to thank fellow department colleagues I had an opportunity to collaborate with on various research projects, also outside of my dissertation work: Nan-Chen Chen, Ray

Hong, Meg Drouhard, Jina Suh, Elena Agapie, Michael Brooks, Raina Langevin, and Amelia Wang. These collaborations allowed me to explore different areas, fueled my interests, and expanded my perspectives on broader research. I would like to especially thank Nan-Chen Chen and Ray Hong for our extended collaborations on several published projects.

I am also very grateful to all the members of the Prosocial Computing lab and my cohort: Ahmer Arif, John J Robinson, Mia Suh, Lucas Colusso, Kristin N. Dew, Arpita Bhattacharya, Christina Chung, Jenna Frens, Keri Mallari, Kerem Özcan, Kiley Sobel, Hyewon Suh, Amelia Wang, Spencer Williams, Himanshu Zade, and Andrew Neang. Thank you for numerous potlucks, happy hours, and sharing in my complaints about the hardships of graduate life.

I would like to thank faculty at HCDE who were always very helpful in supporting me and in providing resources I needed: Jennifer Turns, Sean Munson, Julie Kientz, Cecilia Aragon, David McDonald. Especially Jennifer Turns provided invaluable guidance and resources in helping me understand the important topic of refection. I would also like to thank my numerous external faculty and industry research collaborators who helped with various projects in and outside of my thesis: Andrea Hartzler, Jonathan Morgan, Dario Taraborelli, Dennis Hsieh, and Herbert Duber. I also appreciate all the rich perspectives and guidance from my various internship mentors: Andrés Monroy-Hernández, Justin Cranshaw, Daniel Avrahami, Saleema Amershi, Jonathan Bragg, and Doug Downey.

Outside of the research I had great experiences with learning how to guide students and support teaching during my long-term collaboration with Andy Davidson on teaching HCDE 539. I look fondly at all the fun physical computing project we were able to foster among our students and our numerous conversations on how to improve the course further. I am grateful for the invaluable knowledge about the teaching process I gained though my work with Andy. I am also really thankful for his understanding of my grading delays during the final quarter when I was busy wrapping up this dissertation. I also want to thank my other

teaching collaborators: Brock Craft, and Rafael Silva.

Last, but not least, I am extremely grateful to my family for their support. My Mom for always being there to listen to my challenges and support me with good advice. My grandparents for their perpetual concern about my eating habits and to my girlfriend Ada for always trying to align our busy research schedules to find some time to spend together.

DEDICATION

to my Mom for her continued support

Chapter 1
INTRODUCTION

1.1 The Promise of Conversational Agents

Conversational interaction, once a sci-fi dream, is now becoming increasingly common in everyday use of various computer systems. Voice Assistants (VA) such as Alexa, Siri, and Google Home provide voice-based interactions for accessing news [69], providing weather information [217], supporting scheduling [56], and controlling devices at home [137]. Similarly, text-based service chatbots support vacation planning, financial advice, and pizza ordering [112]. The principal value of conversational interaction in most of these scenarios is efficiency and claimed convenience for the user (e.g. hands-free interaction for voice [162], reusing a familiar messaging interface for chatbots [127]). Another, somewhat less commercially explored, benefit of conversational interaction lies in its potential for improving user engagement. User engagement is important in both task-oriented and non task-oriented applications [112]. Primary examples of such, arguably less task-oriented, applications are social chatbots with Xiaolce [249], Mitsuku [183], and Replika.ai [171] being the most successful. The engagement benefits of conversational interaction have been commonly attributed to human mimicking aspects of such interfaces [41].

1.2 Engagement Challenges of Current Technology for Behavior Change

In the domains of health & behavior change, keeping users engaged, particularly over a longer period of time and across multiple interactions has always been a challenge [7, 30, 182]. Technology has offered valuable support in automating many aspects of behavior change, but has been arguably less successful in improving user engagement [93, 107, 117]. Typical user journey in personal informatics (a technology based approach for supporting behavior

change) involves, among others, 1) execution of actions that would help the user change behavior, 2) reasoning based on past behavior and collected data, and 3) collection of data about one's behavior [73]. Each of these is associated with challenges. Wearable sensor-based technologies (e.g., Fitbit, Apple Watch) aid users in automated collection of behavior data and can be seen as some of the greatest benefits of the use of technology [117]. Yet, certain data can't be measured easily and needs to be collected through self reports [52]. Surveys are the primary method of collecting such data, but oftentimes users find them tedious and resort to various response satisficing behaviors [104, 199]. Particularly in health & medical domains where sensitive data and vulnerable populations may be involved this can lead to trust, engagement and understandability issues, which lower response quality and rates [96]. Even if the data can be collected automatically, a crucial purpose of this data is to support reasoning and reflection on it [23]. For this purpose, the technology currently supports users predominantly with dashboards and graphs. While useful, these often encourage only surface level habitual "glancing" of the data, without engaging users in deeper reflection on the meaning, interpretation, and mechanisms behind the information [22]. This is, in part, why human health coaches can be more effective in keeping users engaged [139]. Finally, sustaining behavior change actions via technology often relies on repeated reminding strategies [173] and planning support [4]. Yet commonly used automated reminders are repetitive, monotonous, and over a period of time tend to cause boredom and are often eventually ignored by the users [35, 101, 223].

1.3 Value of Conversational Approach in Behavior Change

Conversational approaches with their ability to mimic some of the human-human interaction aspects have the potential to improve on a number of these challenges. Human administered surveys, especially with vulnerable populations, have been reported to lead to higher quality responses, yet are more costly, less scalable, and not always possible [104]. Conversational interfaces have an ability to combine the best of both worlds by limiting costs through automation and providing some of the valuable human-human interaction aspects.

Similarly better learning from the automatically collected behavior data can be aided with well-crafted reflection dialogues around such data. Such dialogues could support user reflection and self-learning in a manner similar to what human health coaches do [198]. Finally, repetitiveness and artificiality of automated activity reminders can be reduced by mimicking some aspects present in human use of messaging platforms, such as rich diversification of topics and phrasing as well as tailoring based on understanding of conversational partner's interests and values. Such aspects come naturally in interaction with friends and professional human coaches [4].

1.4 Challenges of Conversational Agent Design for Behavior Change

However, designing conversational interactions for health & behavior change is challenging in a number of ways. Depending on the aspects being supported, the conversational agent might have a clearly defined task (e.g. collect specific health data) or be fairly open (e.g., help users reflect on their own data). Usually, however, a certain mix of the two is desired. Prior work indicates that users want to be efficient in their interaction with conversational interfaces, but at the same time expect some level of socialization even in more task-oriented applications [112]. Yet, previous work also warns against so-called 'mission creep' where the conversational interface introduces distracting interactions for the sake of being more social [97]. Combined with still profound limitations of AI technologies, this can easily lead to bloated expectations of intelligence that the agent can't sustain [156] and introduce the risk of interaction breakdowns [12]. On top of that, there are indications that not all users expect and enjoy the same level of socialization, even in the same context of use [147]. Furthermore obtaining rich, diverse, and personalized contents for conversational interaction is another design and technical challenge [76]. Existing datasets contain social media exchanges or tech support dialogues [5], which are not directly suitable for use in behavior change applications. Furthermore the design process involved in collecting or generating domain specific data from users in the form amenable for use in dialogue design is not yet well explored [49, 140, 192]. Diversification of contents, to create novel and engaging interactions each time is yet another

challenge [242]. Due to all these difficulties, designing good-quality conversational interfaces can be exceptionally challenging and existing tools fall short of supporting numerous nuances of the whole process well [97].

1.5 Thesis Statement

My thesis claim is summarized in the following statement:

> *Conversational interactions leveraging content diversification and tailoring can (T1) increase activity adherence, (T2) facilitate reflection, (T3) support collection of sensitive data. The effort of designing such interactions can be (T4) lowered with automation*

1.6 Research Overview

To demonstrate the fulfillment of this statement I explored the design of conversational interfaces for supporting health and behavior change practical applications related to some of the key aspects in this domain: 1) motivating & sustaining actions (Chapter 3), 2) supporting learning from behavior or reflection (Chapter 4) and 3) collecting personal data (Chapter 5). Through leveraging conversational design to support these applications I provide practical approaches for addressing the common challenges involved in conversational design in health & behavior change context. I also explore the mechanisms in which conversational interaction can provide benefits to user engagement and overall experience. Finally, I distill the common best-practice design principles I discovered across the settings to propose an automated support for lowering the conversational design effort (Chapter 6). Table 1.1 describes the research questions I examined aligned with the thesis claims. The structure of the rest of my dissertation follows below.

Chapter 2 discusses the challenges of current technology support for health & behavior change and how conversational approach could help address these challenges. I present key aspects of health & behavior change which require support according to prevalent behavior

Thesis Claim	Research Question	Addressed in
T1, increase activity adherence	**RQ1**: How can conversational approach improve repetitive activity triggers? **RQ2**: What is the impact of conversational content diversification on user boredom & adherence?	**Chapter 3**, through creation of two systematic content diversification strategies informed by cognitive space theory. Further through lab and field studies with *Fitness Challenges* system.
T2, facilitate reflection	**RQ3**: How to generate engaging personalized contents for reflection? **RQ4:** How to leverage dialogue structure to benefit engagement? **RQ5**: What is the impact of conversational approach & interaction modality on ease and depth of reflection?	**Chapter 4**, through proposing a workshop-based and structured reflection informed content generation. Further through development of *Reflection Companion* and *Robota* systems, which inject work tasks, fitness tracker data, and health goals into the dialogues. Finally through evaluation in the field studies.
T3, support collection of sensitive data	**RQ6:** How to design conversational data collection to improve engagement, comfort and understandabilty? **RQ7:** What is the impact of conversational approach to data collection on vulnerable populations?	**Chapter 5**, through development of *HarborBot* conversational social needs screening system with empathy and understandability features. Further through evaluation with high and low-health literacy patients in hospital emergency room setting.
T4, lower the conversational design effort	**RQ8:** How can designing conversational data collection be automated? **RQ9:** Which design aspects are easy and which are hard to automate?	**Chapter 6**, through proposing 4 tasks for conversational survey adaptation and creation of a repository of augmentations. Further though a 3-step evaluation of automation performance, remaining correction effort, and via a user study.

Table 1.1: Thesis claims, research questions and chapter organization.

change [191] and personal informatics [73, 142] models. I discuss the specific challenges identified in prior work and related to these aspects. I then discuss the characteristics and potential benefits of conversational interfaces based on theoretical indications and past lab studies. Finally, I discuss how these potential benefits of conversational design align with challenges of current technology support for health & behavior change.

Chapter 3 explores the potential for conversational design to motivate & sustain activity. In this work I redesign the activity triggers for physical activity promotion to follow conver-

sational style. I specifically introduce natural diversity of language & topics and personalize the interaction. For topic-based contents diversification I specifically propose two systematic strategies informed by cognitive space theory. I use a simplified format of conversational interaction focusing on single turn exchanges over a period of time on a mobile device. I test the proposed strategies in a lab experiment and a controlled field study.

Chapter 4 explores the potential of conversational approach to engage users in meaningful reflection. I design two conversational reflection agents for physical activity (Reflection Companion) & workspace productivity (Robota) settings. Both settings have been indicated in prior work as in need of support for reflection [139, 134]. In physical activity setting with the use of mini-dialogue structure informed by structured reflection theoretical model I address the challenges of lowering the effort and deepening the reflection by splitting the challenge into smaller manageable guided reflection turns. In workspace productivity setting I compare the use of voice and text-based chat for triggering reflection and explore the challenge of personal reflection design for semi-public space. In both settings I also personalize and diversify the interaction to make it more engaging. I test both approaches in field studies.

Chapter 5 investigates the opportunities conversational interaction offers for increasing comfort with sharing sensitive personal data & improving understandability among low health-literacy users. I design HarborBot, which is a mixed-modality (voice and text) chatbot for conversational administration of social needs screening survey in a hospital emergency department (ED) setting. To specifically address the challenge of comfort with sharing sensitive information I explore design for conversational empathy via social phrases, interface social cues, and empathetic reactions. Furthermore to address the challenge of understandability among low health literacy populations (common in this setting) I design conversational question rephrasing and the use of voice-based question readout. I evaluate HarborBot in a controlled study performed in ED setting with low and high health literacy patients comparing conversational and form-based social needs screening approaches.

Chapter 6 explores the opportunities for use of automation to support design of engaging conversational interactions. I design & develop an automated process for adapting survey-

based data collection to a conversational form. In this work I turn the manual design process I applied in Chapter 5 into a semi-automated process that lowers design effort and systematizes some of the engaging conversational design principles I developed in prior chapters. I perform a 3-step evaluation. I evaluate the data-driven machine-learning aspect of adaptation in a leave-one-out and cross-validation setups. I also evaluate and quantify the remaining survey administrators' manual correction effort (caused by automation imperfections) and finally evaluate the impact of the generated conversational surveys in a crowd-sourced study.

Finally, Chapter 7 discusses how the conversational application projects explored in my work provide evidence in support of my thesis claim. I also specifically discuss the benefits and challenges of conversational approach identified throughout my work. I conclude by briefly describing a few directions I plan to pursue in the future.

1.7 Summary of Key Contributions

Findings from my thesis provide a better understanding of how to design conversational experiences that address the key challenges of successful health & behavior change. Furthermore, my findings also indicate how the design process of engaging conversational experiences in these settings can be supported with automation. The contributions of my thesis include:

Conversational activity promotion (Chapter 3)

- *Design:* Proposed two systematic content diversification strategies informed by cognitive space theory: target-diverse and self-diverse.
- *Design Process:* Proposed a crowd-sourced process for generating motivational conversational prompts which are both tailored to individuals' values and diversified.
- *System Artifact:* Implemented a Fitness Challenges mobile conversational system for activity promotion.
- *Understanding:* Provided insights into user perception of tailored & diversified conversational prompts vs. repetitive non-conversational messages.
- *Evidence:* Demonstrated an ability of the self-diverse conversational prompts to signif-

icantly increase exercise completion.

Conversational reflection (Chapter 4)

- *Design Process:* Proposed a workshop-based process for generating diversified conversational reflection questions informed by a structured reflection theoretical model.
- *Design:* Proposed a 2-step mini dialogue informed by structured reflection model for lowering reflection effort and guiding the users towards deeper reflection.
- *Design:* Proposed a design for reflection at work which combines work-related benefit (i.e., journaling & reporting) via work-based channel with a personal benefit via a dedicated separate channel (e.g., organization, career goals).
- *System Artifacts:* Implemented Reflection Companion & Robota conversational systems for physical activity & workplace productivity respectively.
- *Evidence:* Demonstrated the ability of conversationally supported reflection to engage users in interaction as well as meaningful reflection.
- *Understanding:* Provided detailed understanding of how specific conversational elements (e.g., two-step dialogue, typing & sending response) affect user engagement and the quality of reflection.
- *Understanding:* Provided insights into user perception of reflection via different modalities (voice & text) and the related differences in interaction and engagement.

Conversational data collection (Chapter 5)

- *System Artifact:* A mixed-modality (voice and text) chatbot called HarborBot for administering a social needs screening survey in a conversational manner.
- *Evidence:* Demonstrated the benefits of conversational data collection (for social needs screening) with vulnerable populations.
- *Understanding:* Insights into how low & high literacy ED patients perceive different aspects of conversational social needs screening.

Automating conversational design for data collection (Chapter 6)

- *Design:* Proposed automated conversational adaptation of any survey in 4 steps: 1) addition of introduction & closing, 2) addition of contextual empathetic reactions, 3) addition of progress communication handling, 4) adaptation of question language to conversational style.
- *Implementation Artifact:* Implementation of the proposed 4-step conversion approach using ML techniques and a reusable repository of conversational augmentation phrases.
- *Evidence:* Demonstrated that the proposed automation approach can produce engaging conversational surveys (comparable to manual design) with only limited additional manual correction effort of grammar and misclassifications.
- *Understanding:* Provided insight into what it means to make survey-based data collection conversational & identified the trade-offs between survey administration requirements (e.g., dictated by validity) and an engaging conversational experience.

1.8 Summary of Key Findings

The summary of my findings from each chapter are presented in 1.2.

Chapter	Summary of Findings
Conversational Activity Promotion (Chapter 3)	• Self-diverse strategy significantly increase user activity performance in a 2-week long field study, making users 3.7 more likely to exercise. • Topic-based diversification of prompts can attract user attention (perceptually), provide informational value (more opportunities for new information), increase personal relevance (more opportunities for cognitive elaboration & human-like feel) • Non-diversified repetitive prompts are perceived more like reminders, in which users ignore the repetitive aspects (content blindness). • Conversational design of prompts can trigger higher expectations of intelligence & contextual meaningfulness and also invite higher scrutiny of the contents quality.
Conversational Reflection (Chapter 4) Both Settings	• Conversational approach can trigger different types of reflection (increased awareness, alternatives and future actions, and new insights) and offer tangible benefits (increased motivation, new behaviors, mindfulness, more realistic plans). • The use of personalized aspects in the dialogues (name, activity graphs, tasks) is useful for grounding responses in personal experiences, promotes engagement & motivation
Physical Activity Setting Specific	• Two-step mini dialogue structure for reflection can offer benefits: 1) extending thinking time for reflection, 2) encouraging deeper thinking and more meaningful answers, 3) lower reflection effort, but also runs the risk of disappointing users if the second step feels 'generic'. • Typing and sending responses in chat has the benefits of promoting deeper thinking, as well as seriousness & precision in making plans, but also incurred typing effort. • Once a day reflection supports reflection on a continual basis, enables devoting the whole day to deepen reflection on one aspect.
Workspace Productivity Setting Specific	• Too broad or out of context reflection questions at work can be perceived as a meaningless to reflect on and an unnecessary distraction. • Slack-based text modality for conversational reflection perceived as 1) easier to read questions and think about responses in their own time, 2) easier to reply in own time and describe the details, 3) easier to review and change responses, but at the same time to be: 4) more time consuming due to typing, and 5) less personal than voice. • Voice modality for conversational reflection considered: 1) valuable to have a separate channel just for reflection due to more personal feel and an ability to quickly capture 'quick' thoughts, 2) faster to answer questions with voice, as well as 3) more interactive, fun and engaging, but at the same time caused: 4) a pressure to respond immediately and 5) listening to own responses to be inconvenient and uncomfortable.

Chapter	Summary of Findings
Conversational Data Collection (Chapter 5)	• Conversational social needs screening applied in ED setting can be more engaging (due to feeling of talking to somebody), perceived as more caring (due to personality & empathy), and understandable (due to audio & question rephrasing) especially for low health literacy populations. • Efficiency of interaction is much more important for high health literacy, than low health literacy users • Perception of inefficiency can be triggered by 1) fixed & sequential pace of interaction, 2) a need to wait before question shows up (e.g., due to typing indicator 'ellipses'), 3) ability to read faster than the voice readout, 4) presence of additional conversational utterances, 5) inability to concentrate on reading with audio on. • Conversational social needs screening can feel 'pushy' due to: 1) direct questions, like from a teacher, 2) lack of lead in interaction between very sensitive questions, 3) perception of chat trying to repeatedly get information that was declined, 4) feeling rushed to respond due to short delays
Automating Conversational Design for Survey-based Data Collection (Chapter 6)	• A simple 4 tasks based conversion composed of 1) addition of introduction & closing, 2) addition of contextual empathetic reactions, 3) addition of progress communication & topic handling, 4) adaptation of question language to conversational style can offer engaging conversions while needing only relatively minor correction effort. • The same empathetic reactions to user answers in conversational surveys can be very polarizing depending on context: perceived as engaging, natural, pleasant an even 'cute' in one context and as judgmental and patronizing in another. • Proposed approach involving 3 types of empathetic reactions suffers from: 1) lack of appropriate reaction class for specific scenarios, 2) insufficient use of broader context, and 3) lack of specificity to survey contents. • Several conversion challenges are related to the trade-offs between survey requirements and conversational experience (e.g., rephrasing 2nd and 1st person survey items, addressing survey intrinsic question repetition), as well as availability of data matching socialization and empathy to the user.

Table 1.2: Summary of Chapters and Findings

Chapter 2
BACKGROUND

Many people aspire to seek to change their behaviors to better themselves in various aspects, such as eating healthier [72], exercising [93], being more productive [107], or better managing one's finances [119]. Increasingly changing lifestyles leading to widespread obesity in developed countries, aging populations, and disparity in access to health services, further emphasize the particular importance of behavior change in the domain of health [48].

Numerous existing behavior change frameworks (e.g., Theory of Planned Behavior [6], Transtheoretical Model of Behavior Change [191]) identify various factors important in influencing one's motivation to change behavior. These involve individual characteristics, such as intrinsic or extrinsic motivation [230] or self-efficacy (e.g., how much the person believes in successfully changing their behavior [17]). People may have different reasons for wanting to change behavior, such as specific one-time goals, maintenance of existing positive habits, or wanting to increase a particular behavior (e.g., increase physical activity levels) or even eliminate or decrease unwanted behaviors (e.g., smoking). Behavior change, to be effective, requires a fundamental ability to collect information about one's behavior, an ability to effectively use this information to introduce measurable changes to one's behavior, and an ability to sustain such improved behavior over time [73, 142]. Combination of all these factors make behavior change on individual and societal levels very challenging [120]. That's why, six months after making a New Year's resolution, only 46% of people were still on track with their behavior change goals [177]. Similarly in the health domain, only 22% of Americans follow the aerobic and muscle strengthening national guidelines [103], and less than 23% of the world population meets recommended guidelines [237].

Recently some of the more practical challenges in behavior change, such as tracking

activities and measuring effectiveness of interventions, have been increasingly supported by emerging technology-based tools. Such support was made possible by advances in wearable sensors, widespread internet connectivity, and prevalence of mobile and IoT devices which led to the creation of technology supported behavior change in the form of personal informatics [73, 142].

2.1 Current Technology Support for Behavior Change

Personal informatics relates to the use of technology for collecting and reflecting on personal information [143]. Li et al proposed a five-stage model of personal informatics that characterizes how technology can support people in behavior change by proposing phases of preparation, collection, integration, reflection, and action [142]. Epstein et al. further built up on this work with a lived informatics model, which adds, among others, aspects of lapsing and resuming [73]. Similarly, beyond individual (the main focus of personal informatics), a popular stages-of-change model [191] identifies several similar stages. The Precontemplation stage in which the user may be unaware of problematic behavior and may need evidence and support (also from others) in realizing that the change is needed. In the Contemplation and Preparation stages, the user wants to take action, but may need support in deciding what is the best action to take. Finally, the Action and Maintenance stages is when user takes actions, but may need support in maintaining the positive momentum over time. These models characterize how technology has and can be used for supporting behavior change. While different models propose different steps or stages, they all identify a number of common crucial behavior change aspects and further characterize how technology has struggled to support these aspects.

2.1.1 Challenges of Sustaining Actions over Time

The maintenance of positive activities is crucial in behavior change and a major challenge. Lived Informatics model includes lapses and difficulties in resuming activities as temporal aspects of the challenge [73]. Identifying important factors for predicting and affecting people's

intentions for actions is the main focus of many theoretical behavior change and persuasion models [57]. Successful change in behavior usually requires consistent and sustained execution of actions considered desirable for an extended period of time (e.g., running, going to sleep at proper times, eating healthy). Reminders and message-based triggers have been some of the most commonly used technical solutions for supporting such sustained user involvement in behavior change efforts [84, 167]. Yet despite their prominence, the design of effective message-based triggers is challenging [54, 84]. One of the main problems stem from the need for repeated user exposure to such triggers which can lead to annoyance, boredom, content blindness, or purposeful avoidance [35]. This "alert fatigue" has been linked to use of same or similar contents and lack of personal relevance of the message for users. Both issues call for designing motivational triggers that are diversely phrased, novel and personalized in their contents [65, 100, 169, 208]. Yet important challenges pertain: What aspects of the behavior change triggers need to be diversified to make them appear novel, engaging, and natural for users? How can we design diverse and personalized triggers in a way suitable for use at scale? Which and how mimicking aspects of human-human communication can help improve the efficacy of triggers?

2.1.2 Challenges of Reflection and Learning from Data

Helping users make sense and learn from behavioral data is yet another challenge. Technology has been used for combining user activity data collected from different sources (e.g., wearable activity trackers and self-reports) into a format intended to support exploration and self-learning (reflection) on past behavior patterns [22]. This has often been supported via dashboards [195], visual analytics tools [132] and glanceable displays [92]. Reflection is considered a crucial step that translates observations to actions [23] which can help users increase their self-knowledge [22], formulate realistic behavior change goals [139], and increase self-control while promoting positive behaviors [144]. Despite the importance of reflection, personal informatics models reveal little about how reflection can, or should be triggered via technology [22]. At the same time several personal counseling techniques (such as moti-

vational interviewing [198]), as well as commercial behavior change programs (e.g., Weight Watchers [113]) rely on engaging and insightful conversations with the goal of triggering reflection on one's own activity. Unfortunately, technology has struggled to successfully support reflection in practice at the same level as human counselors [218] and design for reflection is still in its infancy [23, 82]. As noted in [22] *"prior work carries an implicit assumption that by providing access to data that has been 'prepared, combined, and transformed' for the purpose of reflection, reflection will occur."* Current best practices rely on visualizations of self-tracking data [132, 55], or on journaling [149]. Both of these approaches assume that reflection will occur naturally when the data is presented. However, reflection is time consuming and not necessarily something that comes naturally to people [82]. In many cases people need a reason to reflect or at least an encouragement to do so [168]. Therefore important questions pertain: How can we design technology that could mimic some of the best practices of human counseling and coaching to better support reflection in behavior change?

2.1.3 Challenges of Data Collection

Technology has been particularly successful in supporting easy collection of measurable behavior and physiological data (e.g., steps, physical activity, heart rate). Unfortunately, not all data relevant for behavior can be easily collected or inferred with sensor-based tracking. Important aspects such as personality traits, individual motivations, as well as social determinants of health still need to be largely collected via self-reports [52]. Technology can still help, by means of electronic surveys which can easily and cost effectively scale, but numerous studies have identified challenges in how technology is currently used for survey administration. Electronic surveys can be easily ignored and can collect lower quality responses as compared to in-person interviews (in one study, 92.8% face-to-face response rate compared to 52.2% web-survey response rate) [104]. Furthermore, traditional surveys have been found to implicitly bias against non-whites [151], low income individuals, homeless or those that are disenfranchised with mental health and/or substance use [44]. Such disen-

franchised populations are likely to suffer disproportionately more from health, financial and legal issues and are in greater need of behavior change support [157]. These challenges have been partially attributed to the difficulties of understandability, trust issues, as well as to the rigid and impersonal way in which technology is employed to collect often sensitive and personal data [32]. Therefore important questions pertain: How can we support survey-based data collection that is understandable, flexible and empathetic for sensitive settings? How can we employ positive aspects of face-to-face interviewing to improve user engagement with automated survey administration?

In summary, the majority of technology support for behavior change relies on providing tools that can prove helpful for already motivated and engaged users, but technology support may fall short of effectively engaging less motivated ones. The often impersonal, rigid, non-empathetic, and less thought provoking use of current technical tools can limit the level of support the technology could provide for users in need. Therefore it is an interesting question whether it's possible to effectively redesign important behavior change interactions by mimicking some of the aspects of natural human-human interaction to make these interactions more engaging for the users? Furthermore, how to successfully design technology that mimics such human-human interaction aspects without running into the trap of overpromising intelligent behavior given, still profound, technological limitations? In the next section I look at how a conversational approach could potentially improve on some of the issues I have identified in the currently existing technology-based support for behavior change.

2.2 Potential value of Conversational Approach

Conversational agents (CA) with their ability to mimic aspects of human-human interaction have the capacity to improve on different behavior change challenges. Past work found that polite interruptions used by CAs [27] as well as use of social dialogue, empathy and expressions of friendliness [26] can improve users long-term engagement. Similarly, uses of persuasion [30, 187] and behavior change techniques [90] in conversational contexts have been shown effective for motivating users [206]. Furthermore, some of the intrinsic properties of

natural conversations, such as novelty of topics, natural phrasing, content diversification, and personal relevance also have the potential to help sustain long-term engagement and alleviate some of the alert fatigue experienced when supporting behavior change actions repeatedly.

To address the challenges of reflection and learning from behavior data conversational approach can offer a number of advantages. One of the main methods in which human-coaches engage people in learning is by repeatedly asking open questions that trigger deeper thinking [139]. Such questions can help people understand their own needs and motivations. Unfortunately, simple prompting approaches such as "tell me more" or restating what the user said in the form of a question (e.g., Eliza [235]) only have short-term value [28, 176]. A dialogue that can support structured progression of reflection and build up on user answers can help elicit contemplative [114] and metacognitive [81] thinking, encouraging people to think about the needs and wants beyond their first answers that come to mind when relating to singular prompts.

Conversational approach can also be helpful in supporting data collection. Past work demonstrated that conversational survey administration can increase user attention to survey questions leading to better quality responses and higher engagement [124]. These qualities have been specifically attributed to mimicking human-like interactions in chatbots [239]. Mimicking such aspects has also been shown to improve trustworthiness [194], which can be very valuable when personal or sensitive data is collected. Furthermore the use of voice in interaction can help mitigate understandability issues among low literacy participants [96]. This is particularly valuable as low literacy can be correlated with certain systematic health conditions that would particularly benefit from behavior change interventions.

In summary a number of aspects related to interaction and appearance of CAs can be valuable in improving user engagement and other crucial aspects related to various behavior change challenges. While prior work focused on embodiment aspects of CAs, the details of the language use (utterance phrasing, diversification, and novelty), the dialogues itself (topics, progression, and social dialogue), as well as the design process involved in creating these have been less explored. In this work I improve our understanding of these aspects.

Chapter 3

CONVERSATIONAL ACTIVITY PROMOTION: DESIGNING DIVERSIFIED AND TAILORED PROMPTS

In this chapter I focus on the challenge of motivating & supporting action in health behavior change. This goal aligns with Li et al's personal informatics [142] challenge of *promoting repeated actions* and with Prochaska et al's stages-of-change [191] *activity maintenance*. I investigate this challenge in the context of physical activity promotion via repeated message-based triggers, which are widely used by existing technology to sustain activities over time. I aim to demonstrate the value of conversational redesign of such triggers by means of conversational language style [124], diversification [78] & personalization [139] to address the challenges of repetitiveness [101] and content blindness [105]. I use a simplified format of conversational interaction focusing on single turn exchanges over a period of time on a mobile device, which is a predominant platform used by users wanting to change their physical activity related behaviors.

3.1 Background

Repeated triggers or reminders are one of the popular forms of motivating behavior change in various domains [167], including exercising [54, 210], sustainable living [3], or civic engagement [178]. Such message based triggers can serve to promote, remind, or even motivate action [84]. However designing effective triggers is challenging [84, 54]. The challenge largely lies in the need for frequent repetition of the triggers to sustain behavior over a period of time, which can lead to user annoyance and boredom [35, 223], purposeful avoidance [101], content blindness [105] and in extreme cases even lower motivation [208].

3.1.1 Conversational Inspirations

Social support has been shown to have a large positive impact on behavior change. Positive encouragements from friends and family in the form of social platforms or personal messages have been shown to have a big impact on user engagement [121]. Hence a proposed solution to adverse effects of repetition could be to make them feel more like coming from a person by making them sound more conversational by personalizing and diversifying their contents. Human health caches commonly personalize their communication with clients based on knowledge about an individual [202]. Content-based and linguistic diversification is a natural 'side effect' of how people communicate [185, 219] and has also been suggested in the domain of advertising [41]. Furthermore prior work has shown that short and similar messages coupled with high repetition accelerate the appearance of tedium (measured by annoyance and boredom) and some controlled experiments demonstrated the positive impact of diversification in constrained settings [100, 208].

3.1.2 Design of Motivational Triggers

While conversational feel, personalization and diversification of communication seem like good candidates for addressing adverse effects of behavior change trigger repetition, prior work has applied these strategies inconsistently and often on a study-by-study manual basis without a clear design process with predictable results [67]. Hence systematic reviews on SMS mobile messages as well as health related text messaging specifically pointed to a need for closer investigation of the relationship between design characteristics and user engagement and retention [79, 102]. In this chapter I therefore investigate: How might designers diversify, personalize, and make the behavior change triggers more conversational in a systematic manner? How can a systematic design process support such improvements?

3.2 Design Approach

To support systematic message diversification I employ a cognitive space modeling approach called Galileo Theory [1]. The theory operates with the notion of semantic similarity between different concepts. It is quite similar to the general notion of e.g., semantic relatedness of words based on Wikipedia links [71] or embedding-based word similarity in neural space [123], with the important difference that the Galileo space is personal for an individual or a group of individuals rather than universal. Using this conceptual framework allows me to diversify messages in a personalized manner [25]. The theory defines two distinct terms in this personal cognitive space: 1) Self referent term (e.g., "self") and 2) target term (e.g., "exercising" for physical activity). Concepts cognitively close to the "self" term are conceptually important for an individual, while concepts close to the target term are semantically similar to it. I used these two terms to propose two systematic content diversification strategies.

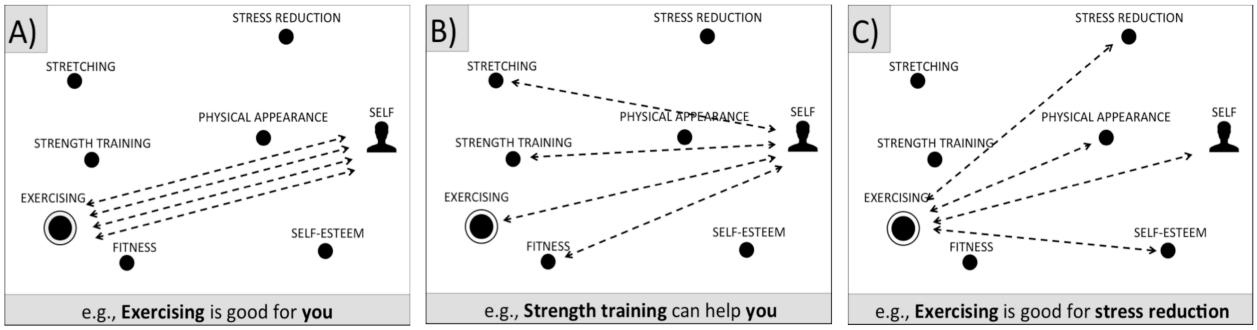

Figure 3.1: A non-diverse (baseline) and two diverse strategies as depicted in the cognitive space: A) Non-diverse - messages connect self and exercising, B) Target-diverse - messages connect concepts cognitively close to target (e.g. different types of exercising) and self, C) Self-diverse - messages connect concepts cognitively close to self (e.g. motivations) with exercising.

3.2.1 Target-diverse Strategy

The concepts close to the target concept of "exercising", being semantically close to it, but still somewhat different linguistically offer an opportunity to create diverse formulations of the messages while maintaining relation to the selected topic. While Galileo space is personal, the semantic relatedness of concepts not around the "self" are likely to be universal (e.g., "strength training" will be perceived as similar to "exercising" irrespective of the person). The related concepts can therefore be obtained using word use similarity. I used the "ensemble" semantic relatedness (SR) measure available in WikiBrain [210] to find an initial list of 8 concepts closest to "exercising" in terms of SR measure: "weight training", "jogging", "stretching", "strength training", "running", "walking", "aerobics", "body building". Based on the properties of cognitive space, by connecting concepts related to exercise (e.g., strength training) to "self" in a message, I can indirectly affect attitude towards "exercising" while providing surface level diversification of the message phrasing (Figure3.1.B)

3.2.2 Self-diverse Strategy

The second strategy takes a similar approach, but centers around the self-referent term of ("self") (Figure3.1.C). In this case, however, I can't rely on simple word similarity as the terms need to be "close" to the notion of "self", which in Galileo space means these are the concepts that the message recipients would care about (e.g., stress reduction). In that case the notion of similarity is likely not universal, but personal. Such personally relevant similar concepts can obtained in various ways, I used two approaches in my work: In study 1, I used prior literature around exercising motivations [135] which informed 8 key motivations people have for exercising: "stress reduction", "physical appearance", "increased vigor", "relaxation", "health", "fitness", "pleasure", "self-esteem", In study 2 I relied on Schwartz's values framework [209] to rate individual' closeness to 10 basic universal values of: "achievement", "benevolence", "conformity", "hedonism", "power", "security", "self-direction", "stimulation", "tradition", "universalism".

3.2.3 Generating Messages

While the above strategies provide a set of key concepts or terms that the message should revolve around (e.g., "fitness" and "aerobics" or "self-esteem" and "jogging") the actual text of the message (syntax) also needs to be designed. I incorporated two different approaches to do that. In study 1, I used a template based approach, where I keep the syntactic components intact and just swap different terms (e.g., "[Exercising] can help with improving [self-esteem]. Latest research has confirmed many of the anticipated benefits of improved [self-esteem]"). In study 2, I used a crowd-sourced generation where the crowd-workers were asked to write a motivational message that connects the given notions (e.g., instruction to connect "power" and "exercising" could result in a message: "People respect someone who makes a commitment to exercise!"). Examples are presented in Table 3.1.

3.3 User Study

I conducted two studies to investigate the effects of diversification. First study focused on comparing self-diverse, target-diverse and baseline strategies in a controlled lab study with 150 Amazon Mechanical Turk (AMT) workers using template-based message generation. The second study further tested the more personalized strategy ("self-diverse") in a 2 weeks long field deployment with 28 participants receiving messages along with exercise challenges on their mobile phones.

3.3.1 Study 1 - Controlled Lab Experiment

The AMT workers (45% male, median age group 25-34, 44% exercised regularly) participated in a between subjects study with 3 conditions: target-diverse, self-diverse and the baseline non-diverse. Participants were told that they will be asked to evaluate a draft of an informational website about health and nutrition. They were then shown a series of 4 web pages. Each web page contained health and nutrition information (100-150 words) and an associated image (e.g., presenting people running or eating healthy). I used the actual content from a

university's student health services' webpage. On each of the 4 pages the participants were shown 1 health tip consisting of an image (always the same for each message and condition) and a pop-up text message trigger (participants saw four different messages, one on every page). To protect against possible ordering effects, I counterbalanced the message order for the manipulation conditions. Participants received $2.20 for study participation. Based on the design goals I had 4 hypotheses:

- **H1:** Annoyance towards the message-based triggers will be lower when using diverse strategies.
- **H2:** Boredom with the message-based triggers will be reduced when using diverse strategies.
- **H3:** Informativeness of the message-based triggers will be rated higher using diverse strategies.
- **H4:** Helpfulness of the message-based triggers will be rated higher using diverse strategies.

To test these I measured annoyance (H1.1), boredom (H1.2), informativeness (H1.3), and helpfulness (H1.4) by asking the participants to estimate the experienced level of each towards the message contents on a 5-point likert scale. I measured behavior intention and attitude using TPB [6]. The reliability of both TPB measures were high, attitude: α=0.88, intention: α=0.77. I also measured reactance (negative impact of persuasion) following [66].

3.3.2 Study 2 - Controlled Field Deployment

I also conducted a second study that involved a field deployment. This was meant to address the two main limitations of the first study. First is the lack of realism. Study 1 involved only four messages and participants were not sent these messages in the context of behavior to perform. The second major limitation is that I collected only perceived measures. I aimed to

address these in the field deployment where users were asked to perform actual exercises and report their performance with a reply text message on their mobile (3.2). I used a between subject design, comparing the stronger of the two strategies (self-diverse) to the baseline, non-diverse. A stratified randomization based on the level of physical activity was used to assign participants into two experimental groups (self-diverse and non-diverse).

I recruited 28 participants online and through fliers distributed at a university campus (18% male, median age of 31, 32% claimed to exercise regularly). Participants were invited to use a daily-challenges application that I built for the study. The application presents 4 daily exercises that the participants were asked to complete: 2-4 push-ups, 12-15 crunches, 12-15 lunges, and 12-15 jumping-jacks (these numbers were chosen using feedback from pilot studies). Participants were notified via a message trigger to perform these activities four times a day (the order of activities was randomized daily). Participants were asked to perform the activities at 9:30 am, 11:30 am, 1:30 pm and 3:30 pm. They also received a daily summary message at 6:00 pm, which linked them to a webpage showing completion status dashboard (Figure 3.2). Participants could mark their activity completion either directly through the communication channel (e.g., texting "done" back), or if they forgot, they could manually enter in their completion on the website.

In addition to the 4 hypotheses from study 1 I also tested the messages' effect on adherence:

- **H5:** Diversification will increase exercise completion

Each participant was awarded $35 for participation and an additional $20 if she agreed to participate in a follow up interview. On top of the measures from study 1 I collected the self-reported exercise completion rating and conducted one-hour semi-structured interviews with 14 participants that expressed interest in being interviewed.

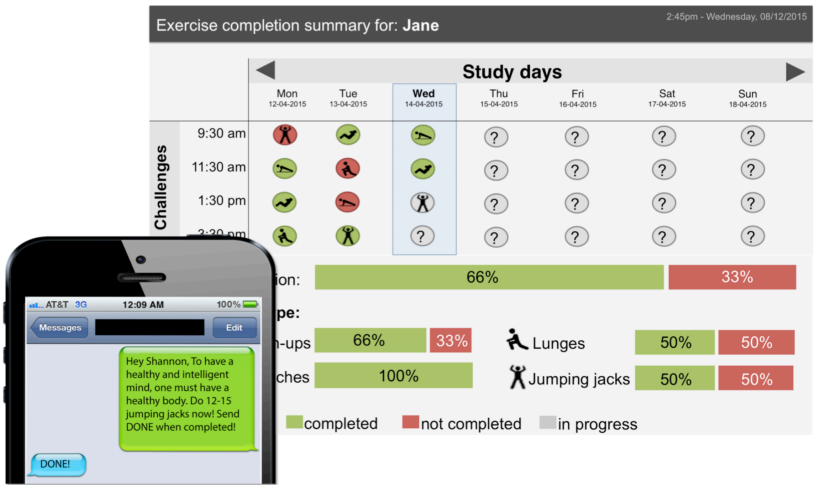

Figure 3.2: The exercise completion webpage used in the study and an example conversational SMS prompt delivered on a mobile phone.

Value	Message text
Achievement	"To have a healthy and intelligent mind, one must have a healthy body."
Benevolence	"Physical activity promotes general happiness and self love."
Conformity	"You can take care of ageing parents better by being healthy yourself."
Hedonism	"Exercising will make you feel and look better!"
Power	"People respect someone who makes a commitment to exercise!"
Security	"Exercising is extremely helpful for your body and mind."
Self-direction	"I exercise so that I have the freedom to do whatever I want to."
Stimulation	"Smell the air, touch the world, experience life. Exercise!"
Tradition	"Even modest effort can have measurable results."
Universalism	"Moving the body is activation of the brain & being one with nature."

Table 3.1: Examples changing part of the conversational prompts used in the field.

3.4 Results

3.4.1 Quantitative Results

In study 1 I found that both strategies were considered significantly more informative and helpful as compared to the baseline (Table 3.2), however, only the self-diverse strategy offered significant reductions in annoyance and boredom. Hypotheses H1 and H2 were only partially supported (not supported for target-diverse), while H3 and H4 fully were supported. The results indicate that the self-diverse strategy performed better compared to the target-diverse one. It could be that the self-diverse strategy addresses more personally relevant issues and such personal relevance may render it less annoying and boring.

In study 2, on top of the measures from study 1, I collected the self-reported exercise completion rating, which I focus on first as the most direct behavioral measure of message effectiveness. The main hypothesis for the field deployment was that self-diverse messages

will increase exercise completion (Table 3.3).

	Non-diverse (baseline)	Self-diverse	Target-diverse
Annoyance	3.85	3.07*	3.33
Boredom	3.79	3.17*	3.24
Informativeness	2.53	3.57***	3.26**
Helpfulness	2.44	3.57***	3.07*
Reactance	1.98	2.05	1.89
Attitude (TPB)	5.26	5.22	5.03
Intention (TPB)	5.15	5.26	4.88

Sig. compared to non-diverse: *** p<0.001, ** p≤0.01, * p≤0.05

Table 3.2: Summary of the results from study 1 - differences between the conditions based on the post-study measures

	Non-diverse (baseline)	Self-diverse
Exercise completion	2.45	2.90†
Annoyance	2.79	3.08
Boredom	3.21	3.08
Informativeness	2.93	2.62
Helpfulness	3.93	2.69***
Reactance	2.43	2.21
Perceived as different	2.00	3.00*

Significance: *** p<0.001, ** p≤0.01, * p≤0.05, † p≤0.1

Table 3.3: Summary of the results from study 2 - differences between the conditions based on the post-study measures

I found that this hypothesis was supported. First, through visual inspection of the data, those in the self-diverse condition completed more exercises in 12 out of the 14 days and on the other 2 days the differences were negligible (Figure3.3); Second, I constructed a mixed-effects logistic regression model for predicting the completion of each prompted exercise (Model 1 in Table 3.4). I found that those in the self-diverse condition are 3.7 times more likely to exercise, but this result is only weakly significant (p=0.09). Using the second model I analyzed exercise completion per day. I binned the number of daily exercises completed into two levels using median split (0-2 completed as fewer and 3-4 completed as more) and used a similar mixed effects logistic regression to predict likelihood to complete more exercises (Model 2 in Table 3.4). This model also shows that participants in the self-diverse condition exercised more (about 6.5 times more likely to complete 3-4 exercises, daily; p=0.04).

Finally model 3 leveraged the fact that there were message repetitions in our self-diverse condition (I did not have the 4*14 messages needed for the full duration of the study), which allowed me to more specifically test whether message repetition affects exercise completion. Using the self-diverse only dataset, I coded up how many times a specific message has been

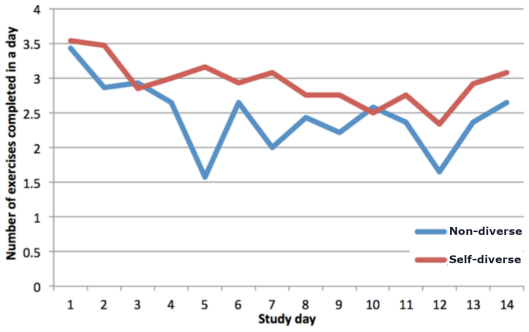

Figure 3.3: Average self-reported exercise completion per study day. Big drops represent weekends

	Exercise completion		
	Model 1 (per exercise)	**Model 2 (per day, 2 tiers)**	**Model 3 (per exercise, self-diverse only)**
	Exp(B)	**Exp(B)**	**Exp(B)**
Condition	3.68†	6.46*	
Intention (TPB) - pre-test	0.63	0.57	1.76
Age	1.10	1.13	1.17
Gender	0.84	0.84	0.88
Exercise day	0.91***	0.89**	1.06
Repetition count			0.52***

Significance: *** p<0.001, ** p≤0.01, * p≤0.05, † p≤0.1

Table 3.4: Mixed-effects regression models for predicting exercise completion

sent to an individual participant and used that as a predictor variable in the model (Model 3, repetition count in Table 3.4). I found that, controlling for the time progression of the study (exercise day), the number of repetitions was indeed a significant factor influencing exercise completion (p¡0.001) with participants being 0.5 times less likely to complete the exercise with each message repetition. Interestingly, in this model, the effect of exercise day was not significant, suggesting that the decline in the self-diverse condition is mostly due to the message repetition rather than potential novelty effects associated with study participation.

On the other hand, post-study self-report measures (Table3.3) indicated that none of the H1-H4 hypotheses were supported. To help explain these potentially conflicting results, and to explore whether and how these triggers helped, I turn to our qualitative results from study 2 interviews.

3.4.2 Qualitative Results

Based on interview feedback I found a number of themes related to effectiveness and perception of conversational triggers. First of all, across both conditions, all the participants appreciated that the triggers reminded them to regularly perform exercises during the day. Some even pointed out that they would not have done any exercises during the day if they

were not reminded of them. Even if the participants felt that they triggers were pushing them a little, they still considered it to be a positive push: *"I liked that it was annoying me to do exercise. Kept nagging at me to do the few workouts needed (. . .)"* (P5, non-diverse). Many participants appreciated that the messages felt positive, encouraging, but also to the point. This shows that regardless of trigger type, having triggers was helpful in general.

Diversification helps

Diversification used in conversational messages in the experimental condition was generally appreciated. In-depth analysis of user feedback allowed me to identify three most common ways in which diversification was considered helpful: 1) attracting attention, 2) providing information, and 3) personal relevance.

Attracting attention: All the participants noticed that the contents of the messages were changing. These constant changes built an expectation of novelty each time a message arrived, which in turn increased attention, sustained engagement and interest in reading the messages: *"I would definitely skip over them if they were all the same message."* (P26, self-diverse) Furthermore, the diversity of the messages led to increased curiosity and introduced a certain element of "fun": *"With all different messages it's fun to read what they say. (...) That's a fun element to it I guess. I remember a few of them"* (P22, self-diverse).

Providing information: About half of our participants in the self-diverse condition liked the fact that the messages delivered small informational pieces about the different benefits of exercising: *"I liked that they talked about all the different benefits of exercises. One of them was learn about yourself, which I thought was cool. Something about endorphins."* (P2, self-diverse).

Also some of the information in the messages was not necessarily obvious to the participants and was therefore somewhat revealing. On top of that, the participants perceived such informational pieces as inspirational: *11I like that it seemed informative. Some of the things*

weren't really as obvious. It's kind of inspirational (...). Yeah, so like, there were some that were kind of more factual." (P15, self-diverse).

Personal relevance: Most of the participants also resonated with specific messages or specific keywords in them. These participants specifically remembered selected messages that had some sort of personal value to them either by relating to their past or current experiences: *"I remember one, 'Being healthy yourself helps your aging parents better, as they get older.' (...) I think that one was the one that I picked up most, because my parents are getting old."* (P24, self-diverse).

In that sense the message diversity was very valuable for helping the participants cognitively elaborate on the personal value of exercising. It is worth noting that we have also proposed the ease of making the diverse messages personally relevant as an explanation of the higher informativeness and helpfulness ratings in our lab study.

Additionally, what also seems to have contributed to the sense of personal relevance was the fact that the messages felt as if a person wrote them: *"I did think that they were written by a person. Certainly, there's a sense of someone writing them for you at some point"* (P3, self-diverse).

Diverse conversational triggers - reminders & motivators; Non-diverse - reminders only

Further analysis of participants' comments revealed that the triggers were perceived and evaluated differently between conditions. In the non-diverse condition, the participants almost immediately noticed that the motivational part of the message is fixed and would subsequently focus on just the part that changes - the exercise to complete: *"My first reaction was 'Yes, exercise is good for me'. Then because they never changed my brain just looped past entirely. (...) I stopped really paying attention to it so much and was going straight to what is the exercise."* (P4, non-diverse).

Consequently, they perceived the messages as simple reminders about exercising, which they considered helpful: *"The message is very straightforward. It's short. 'Exercising is good*

for you and then do something now'." (P25, non-diverse).

On the other hand, those in the diverse condition employed a more critical evaluation of the contents. The changes from message to message solicited more attention from the participants. They scrutinized the messages more than those in the non-diverse condition to see whether the particular contents they received this time is actually appropriate and helpful to them given their context. They expected the messages to be somewhat intelligent and meaningful motivators rather than just simple automated reminders: *"It would be like 'Oh, nature is the same as exercise.' It's like what does that mean? (...) I love the self-help stuff. I love motivation, but it needs to actually make sense to me or seem like somewhat logical I guess"* (P22, self-diverse).

This effect, coupled with the fact that people tend to remember the negative more than the positive [21], resulted in the participants exposed to the diversification strategy recalling and focusing on incidents in which the messages were less helpful. They evaluated the messages not just as a reminder for which exercise they need to complete, but also for their motivational component.

Challenges in designing diversity

Despite many benefits, I also identified a number of challenges that designers need to consider to improve on the use of diverse conversational messages. These are: 1) quality of diversification, 2) challenge of the need for perpetual novelty, and 3) contextual relevance.

Opportunities to increase diversity: Despite the diversification, a number of participants still felt that the messages were not that different. These participants commented that they indeed noticed that the messages were technically different, but felt that they were also very similar in terms of tone and framing: *"(...) They seemed pretty similar in terms of being encouraging of exercise and talking about the different benefits. It seemed like they were the same in tone but certainly, each one was different."* (P7, self-diverse).

Many participants also commented about the practical value of information in the mes-

sages. Although the messages were generally perceived as presenting diverse information about the benefits of exercising, which was appreciated, a number of participants felt that the provided information was not necessarily very revealing for them. They generally felt that they already knew most of the information: *"I think very few of the things were really novel to me, once they're saying, 'Exercise to look and feel better,' you kind of know that."* (P16, self-diverse).

Repetition is a problem: Making sure that the messages stay novel is important. For this 2-week study we did not have enough different messages to ensure that the participants will not receive the same message twice. Unfortunately, almost all the participants noticed this fact and it lead to disappointment in each single case. The fact the messages were repeating, led the participants to lose interest in reading them. Majority reported that they started skipping the motivational part at and focused on the exercise they have to complete once they noticed that the messages are not novel: *"At first, you know ... I think I would have liked it a lot more if you guys had a whole new set of different messages. Make the reading more enjoyable (...) Later when the messages were ... Seems to be repeating I stopped reading them."* – (P11, self-diverse).

It is worth noticing that this qualitative feedback is consistent with my qualitative analysis showing that the message repetition had a significant negative impact on exercise completion.

3.5 Discussion

In this research, I sought to evaluate and understand the feasibility of the two proposed strategies for diversification of behavior change messages. Through both studies, I found several benefits of using the proposed approaches. In a controlled lab study setting, the strategies resulted in messages that were perceived to be more informative and helpful and the self-diverse strategy also reduced annoyance and boredom from repeated exposure. Applied in the field and in a more conversational context of SMS exchanges, the self-diverse strategy also led to an increase in behavior change adherence. This application represented the

Generation step	Goals	Specific approaches explored	
		Self-diverse	**Target-diverse**
Concept generation	• Diverse, but still on topic	• Past literature • Universal values	• Semantic-relatedness (WikiBrain)
Concept selection[1]	• Lowering cost (narrowing down the cognitive space)	• Distances in the cognitive space (crowd sourced)	
Message generation	• Balancing cost and quality	• Fixed message template • Manual writing (crowd based)	
Creation[2]	• Creativity, formulation diversity and natural feel	• Crowd prompts with concepts	
Evaluation[2]	• Ensuring quality	• Evaluation criteria: "on topic" and "motivational"	
Message selection[3]	• Personalization • Context matching	• Individual values scoring	

1 – this step is optional and mostly applicable if the number of concepts is too large to generate messages in a cost-effective manner.
2 – these steps are specific to the crowd-sourced message generation approach and may not be applicable to other methods.
3 – this step is optional and relates to selecting personalized messages or matching messages to the delivery context.

Table 3.5: Summary of the proposed diverse message generation process based on the approaches explored in both studies.

basic use of conversational design (single turn, agent initiative only). At the same time, the exchanges were sustained and repeated over a longer period of time making the setting closer to long-term relational agents employed in [28]. Even this simplified application showed the benefits of a diversified, personalized and more conversational approach to motivating behavior change on an everyday basis that can be further improved with a more elaborate conversational approach.

3.5.1 Conversational Message Triggers Design

Through this research, we proposed two strategies and a general process to generate diverse trigger messages. These strategies provide different benefits making them useful under different circumstances. The self-diverse strategy seemed more effective in mitigating negative effect of repetition. But the target-diverse strategy may also be useful in settings when the target concepts should not be addressed directly, e.g., in anti-smoking campaigns talking

about smoking directly may actually induce more smoking from smokers [95].

Tone & Framing Diversification: In relation to the messages contents itself, the qualitative feedback from the field deployment helped me identify that despite the measurable effectiveness of the diversification strategy, there are opportunities for improvement. One such direction relates to the possible use of different framing or tones for the messages. Currently all the messages are generally positively motivational. I could imagine, the prompts in my crowd sourced message generation process that are more focused on challenging the recipient, pointing out negative consequences of inaction or that employ social comparison for the purpose of triggering competitiveness or cooperation. This would increase the syntactic diversity of the messages within the framing of the prompted concepts.

Addressing Long-Term Repetitiveness: Another aspects of the diversification that could further be improved relates to the repetition of the messages. Both quantitative and qualitative data indicated that when the messages started repeating at some point it had a measurable negative impact on exercise completion and participants' overall experience. Unfortunately, it may be impossible (and too costly) to generate an infinite number of diverse messages. There are, however, a number of options I could explore. One, we could increase the perceived novelty of the messages following some of the techniques discussed in the previous paragraph. Another possibility is that given a sufficiently large set of messages, people might start forgetting past exchanges. There might be an optimal threshold for total message count dependent on use frequency. This could require a more costly generation process, but effectiveness of varying exposure has already been indicated in [102]. Yet another strategy could be to expand the current single-turn, agent initiative only interaction to more of a mini-dialogue with different interaction paths dependant on user answers. Such mini-dialogues could render slightly different exchanges each time contributing to perception of diversity. On the other hand more elaborate interaction could involve more user effort.

Context matching: Finally, the current diversification and message delivery did not take into account the context in which the interaction takes place. Many participants pointed out that they expected the messages to "match" the activity they are expected to perform or change in respect to the time of day or social setting they are in. Lack of such matching negatively affected the perception of personal relevance and introduced a sense of artificiality. I could address this mismatch, by prompting message generation for specific contexts in advance and then try to match these messages to the appropriate context; this would unfortunately increase generation costs. Another approach would be to automatically modify the already existing messages (e.g., via providing templates with slots) to make them more appropriate for the specific context [220].

3.5.2 Diverse Message Generation Process

Aside from demonstrating effectiveness, the practical execution of the designs also provides design insights into the processes and workflows that can be used for generating diverse conversational messaging content. Based on my experiences through study 1 and 2, I propose a four-stage process of systematic conversational message trigger generation (Table 3.5): 1) Concept generation, 2) Concept selection, 3) Messages generation, and 4) Message selection. I summarize the value and importance of each generation step.

Concept Generation: The first step of the process with a goal to generate a diverse set of concepts related to the target concept I proposed two ways of generating diverse concepts: target- and self-diverse. For target-diverse approach, I used semantic relatedness measure (based on WikiBrain [71]) to help assess concepts that are related to the target concept (in my studies - "exercising"). Other approaches that assess relatedness can also be employed (e.g., embeddings [141]). For self-diverse approach I used past-literature (study 1) and values framework (study 2) for generating personally relevant concepts. For settings where the motivations are broad or unclear, the values framework offers an alternative strategy. It provides a manageable set of universal values that people across cultures care about, just at

varying degrees [49]. Generating the contents using these values can then result in a number of personally relevant conversational triggers.

Concept Selection: The goal of this step is to narrow down the size of the concept space. This is optional and mostly important for reducing the costs involved in executing the next stages. The set of concepts can be focused around those most closely related to the target concept or self. This was done in study 1, where I selected the 3 most relevant concepts out of 8 initial ones based on the crowd-generated cognitive space. However, generating cognitive space required laborious comparisons of pairs of concepts. In the near future, this process may be done algorithmically.

Message Generation: This step is where the concepts are turned into concrete text. One appropriate approach is to use fixed sentence templates as I did in study 1. This, however, can produce messages that feel artificial (due to limited lexical diversity). Existing fully automated methods of natural generation could be used, but the exact outcome can be hard to control [33]. In study 2 I explored a crowd based generation, which results in messages that seem much more natural. The messages are also likely to be more creative as well as lexically and semantically diverse, as crowd-workers have the freedom to weave in other concepts, aside from the ones prompted. This benefit has already been observed in previous work, where crowd-workers introduced topics from personal experience [51].

Message Selection: This step focuses on selection of messages, from the generated message corpus, that are the most relevant for a particular participant or particular context of delivery. The goal of this step is to further increase the "natural feel" and personal relevance of the messages. In this work, I focused on personalization based on individuals' values. I asked participants to fill out a short survey to assess their value orientations. Then I sent them a set of the messages that are more personally relevant. To reduce cost, future versions may be able to utilize recent NLP advancements in social media based personality profiling

[42]. Although careful consideration of user privacy and permission for data use needs to be considered.

Context matching can also be important as I have learned in the field deployment, where the pre-generated text did not always go well with some activities or specific social or time-based context. Such mismatch has been picked up by our participants and affected the perceived quality of the conversational triggers. It might also be valuable to include context information already in the "message generation" step to generate text for the set of expected contexts. Another approach would be to use automated natural language generation techniques, to slightly alter the messages on the fly to make them fit the context better.

3.6 Summary of Contribution

In this chapter I examined a more conversational approach to design of motivational triggers, which are commonly used in behavior change for activity promotion. I identified as the major challenges to current technology support for activity triggering: repetition and limited tailoring to an individual. I consequently redesigned these triggers using aspects of a natural conversation: lexical & topic diversity [78], personalization & tailoring used by human coaches [139], as well as conversational language style [124]. I specifically proposed two systematic content diversification strategies informed by cognitive space theory: target-diverse and self-diverse. Target-diverse strategy uses concepts related to the target concept (e.g., "exercising"), while the self-diverse strategy uses concepts related to an individual (e.g., motivations for exercising) to inform diversification. I paired these strategies with topic-based tailoring informed by an individual's values profile [209]. I evaluated both strategies in a lab study as well as in a 2-week long field deployment. I demonstrated that the conversational design of triggers based on these diversification strategies results in higher perceptions of informativess, helpfulness, as well as reduced annoyance and boredom (Study 1) and most importantly can lead to higher real-world exercise completion (Study 2). Designers and practitioners in health & behavior change could use the proposed strategies to improve effectiveness of their motivational approaches. Also designers of conversational systems can

leverage these strategies to inform diversification in the conversational design. Finally this work, aside from the designs themselves, proposed a systematic process employing crowdsourcing and computational semantic-relatedness for effectively reproducing the designs for use in different settings.

Chapter 4

CONVERSATIONAL REFLECTION: DESIGNING FOR PHYSICAL ACTIVITY & WORKPLACE PRODUCTIVITY

This chapter aims to address the challenge of helping the users reflect and learn from their activities, which relates to stages of 'integration' and 'reflection' from Li's five-stage personal informatics model [142]. In terms of stages-of-change [191], this goal aligns with challenges users experience predominantly in 'Contemplation' and 'Preparation' stages. Existing technology for supporting this goal in behavior change uses visual analytics dashboards or journaling, which oftentimes rely on substantial prior user motivation, effort and graph literacy [87] to be effective. The goal of work presented in this chapter is to engage users in 'reflection' on their activities and their data using conversational approach inspired, among others, by approaches employed by human-coaches. I design and evaluate the use of conversational approach for reflection in two settings: physical activity and work productivity. Both settings represent a unique set of challenges and have been indicated in prior work as in need of support for reflection [139, 134]. Physical activity relies on user self-defined actions and goals, is largely personal, and likely performed in a private setting. In productivity settings work tasks are to some extent assigned by others and need to be reported. This setting also involves semi-public space for interaction (i.e., office). Through the course of these works, I explore two approaches to domain specific dialogue content generation: 1) workshop-based, and 2) literature-based, and also investigate the differences between the two modalities of conversational interaction: 1) voice-based and 2) text-based.

4.1 Physical Activity Setting

4.1.1 Background

In physical activity tracking context, mobile, wearable consumer devices allow people to collect and examine large amounts of data about their activities, behavior, and wellbeing. However, a gap remains between our ability to collect and visualize data, and our ability to learn from-, and act upon this data in meaningful ways [143]. A key component for bridging this gap is to facilitate reflection [23, 45, 144]. The value of engaging users in reflection has been identified as a key element of successful health behavior change [144, 158]. Through the process of reflection, users can increase their self-knowledge [22], formulate realistic behavior change goals [139], and increase self-control while promoting positive behaviors [144]. Reflection has been considered an impetus that moves the individual from examinations of his or her data to action [23].

Technology Support for Reflection

Despite the importance of reflection, behavior change models reveal little about how reflection can, or should be triggered [22]. Consequently technology has struggled to successfully support reflection in practice [82, 196]. As noted in [22] "prior work carries an implicit assumption that by providing access to data that has been 'prepared, combined, and transformed' for the purpose of reflection, reflection will occur." Indeed, one of the main means of facilitating reflection using technology relies on visualizations of self-tracking data, such as Fish'n'Steps [148], UbiFitGarden [53] for physical activity; Affect Aura [161] for affective states and LifelogExplorer [132] for stress. The other approach relies on journaling [188], such as SleepTight [45] for sleep and Affective Diary [149] for manual journaling of emotions. Both of these approaches assume that reflection will occur naturally when data is presented. However, reflection is time consuming and not necessarily something that comes naturally to people [82]. In many cases people need a reason to reflect or at least an encouragement to do so [98, 168].

Human Health-Coaches

Taking inspiration from personal counseling, supporting reflection through conversation seems like a promising approach. Several personal counseling techniques, such as motivational interviewing [198] and commercial behavior change programs (e.g., Weight Watchers [113]) rely on engaging and insightful conversations with the goal of triggering reflection on one's own activity. Personal coaches "repeatedly ask questions to get at hidden motivations" and that asking reflection questions can help people understand and articulate their underlying needs and goals [139]. Such conversations can elicit contemplative [114] and metacognitive [81] thinking, encouraging people to think about the needs and wants beyond their first answers that come to mind. In this chapter I therefore investigate: How should a conversational system facilitate reflection on physical activity? Further, can a conversational system support reflection that is engaging rather than burdensome?

4.1.2 Design Approach

The design process for the Reflection Companion conversational agent involved two parts: 1) a workshop with activity tracker users to generate reflection questions, 2) modification of the questions to fit dialogue context and formulation of two-step reflection dialogues.

Workshop-based Content Creation

I organized workshops with activity tracker users to prompt them to write questions about physical activity structured by provided reflection framing. Structured reflection models provide insights for designing reflection-centered interactions and offer support for how reflection can be supported to evolve with time [82, 109]. Such models see reflection as a process with stages or levels. Atkins and Murphy [14] in their review of literature on reflection, identified three commonly-shared stages: 1) awareness of uncomfortable feeling and thought, 2) critical analysis, and 3) development of new perspectives. My approach for structuring the reflection dialogue aligns with these three stages, which for simplicity I refer to as stages of: Noticing,

Understanding, and Future actions (Figure 4.1)

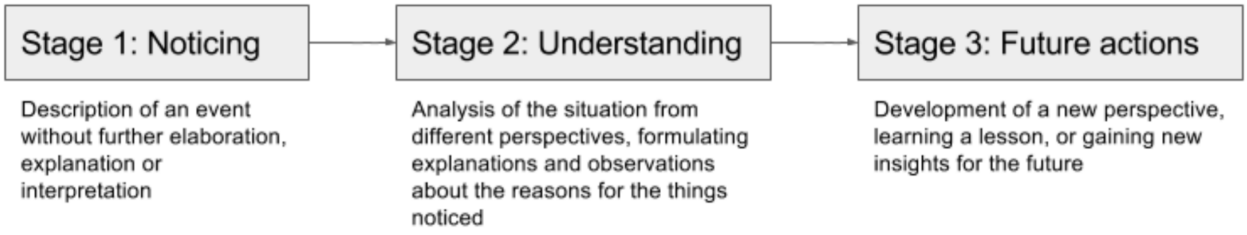

Figure 4.1: Reflection depicted as a process with stages of levels synthesized based on multiple structured reflection models.

Another critical component of a conversational system for reflection are the questions that trigger users to reflect. In this work, I employed a workshop-based approach to generate a set of reflective questions that could be used in the dialogue to trigger reflection. Working with 12 existing users of activity trackers, (8 female, 4 male) with an average age of 27.3 (SD=2.9), the workshop approach helped me generate a diverse set of reflection prompts (Table 4.1).

Workshop participants generated a total of 275 questions in 3 categories I prompted for: Noticing (n=76), Understanding (n=116) and Future actions (n=83). Following analysis of the generated questions, I found the questions within one reflection stage were not all the same and, in fact, could be further sub-categorized by topical aspects of interest. I decided to perform this categorization to be able to later select the most diverse representatives for each discovered category. To do that, I performed affinity diagramming among 3 researchers. The most frequent categories are presented in Figure 4.1. These categories represented different specific aspects of behavior change the participant wanted to reflect on.

Conversational Agent Design

Based on the outcomes of the workshops, I set out to design a system with the following three goals: 1) To guide users towards deeper reflection on physical activity through dialogue

	General/ Context (n=27)	Goals (n=50)	Tracking (n=29)	Observations/ Patterns (n=69)	Motivations (n=20)	Plans/ Scheduling (n=24)
Noticing (n=76)	What are you doing to be more active?	How many days did you meet your goals?	How many days did you wear your tracker this week?	Which days do you walk more?	Did you notice any especially motivating moments this week?	What were the discrepancies between your plans and your actual activities?
Understanding (n=116)	What are the top 3 reasons you're stationary?	Why do you only hit your goal on certain days?	What actions lead you to logging your food?	What happened during peaks/low points during your week?	Why were you sometimes unmotivated this week?	Why did you skip some part of your plan this week?
Future Actions (n=83)	What events could affect your activity next week?	Should you reevaluate your future goals?	What other metrics do you want to track?	How can you avoid low activity days next week?	How can you encourage yourself to exercise regularly?	How can you set yourself up to have a day similar to successful days before?

Table 4.1: Examples of reflective questions generated during the workshop sessions. Questions are grouped by the main prompted categories (rows) and categories identified in through affinity diagramming (columns). Only the 6 most frequent categories are shown. The five white cells represent intersections for which the workshop participants generated no questions. For creating diverse and novel questions, I suggested questions for these intersections.

progression, 2) To provide engaging, novel and diverse conversations around reflection, and 3) To enable interaction on personal mobile devices (predominant platform used among workshop participants).

I designed a conversational system - Reflection Companion - that engages users in reflection on aspects of physical activity through reflection prompts. Reflection Companion uses SMS/MMS for the conversational exchanges. It initiates a short conversational exchange with an opening question sent once a day at random time within a time range specified by the user. I implemented this system as a PHP server using a Twilio API[1] for managing the SMS/MMS exchanges (Figure 4.2). To generate graphs of users' physical activity, I used FitBit API[2] to download the latest synchronized user activity data periodically throughout

[1]https://www.twilio.com/docs/usage/api

[2]https://dev.fitbit.com/build/reference/web-api/

Figure 4.2: Example of an actual user exchanges with our system's mini-dialogues on the left. On the right a block diagram of an example dynamic mini dialogue with: actual user replies, user intents recognized based on free-text replies, and the system tailored follow-ups. The red boxes represent a path where user reply was not recognized and has been handled by a "generic" (non-tailored) follow-up.

the day. To make the reflection conversation engaging and to encourage a deeper level of reflection, I employed three strategies: the use of a two-step minidialogue structure, everyday short reflection sessions, and personalization. Here I further specify the details of these strategies.

Guiding Towards Deeper Reflection through Mini-Dialogues: To support deeper reflection, I used a question-follow-up question design, or what I will refer to as a mini-dialogues design. The mini-dialogues have an opportunity to direct user reflection towards deeper levels by bringing users' attention to different aspects of the reflection process based on user response to the initial question. To build such mini-dialogues, I created follow-up

questions to most of the initial reflection prompts. I followed the progression of the reflection process: questions about awareness would be followed by questions about understanding, whereas questions about understanding would be followed by questions about future actions. The follow-up is asked only after the user provides a response to the initial question. I designed 25 different mini-dialogues, 10 of them have the same follow-up question regardless of what the user writes in their initial response. However, the remaining 13 mini-dialogues feature a dynamically tailored follow-up question. In such dialogues, a different follow-up question is delivered depending on the user's initial response. The tailored follow-ups are designed in such a way as to build upon the user initial response and encourage a deeper level of reflection on the shared information, e.g. if the initial question asked: "What are some of the ways that your work has impacted your physical activity this week?" and the user replied with "Work impacted my exercise because I sit at a desk most of the day", then the follow-up question would be "What could you do to prevent your work from impacting your physical activity?" (Figure 5). On the other hand, if a user replied to the same question with: "I walked a lot this week at work because we were changing offices" then the follow-up would be: "How could you set up your work to help you be more active in the future?" In this example the mini-dialogue is trying to guide the user from understanding how the work impacted her activities, to future actions that can help with being more active. This is following the progression suggested by structured reflection models depicted in Figure 4.

Everyday Reflection Session: An important aspect in the design of our system was the frequency of prompting users to engage in reflective conversations. Too frequent requests for reflection can potentially make the topics to reflect on repetitive and can lead to boredom or frustration given similar activity data and finite diversity of our mini-dialogues. On the other hand, too infrequent reflection can cause people to forget previous revelations, preventing them from building up on past observations and disrupting support for reflection as a process [218]. Human-provided counseling sessions happen infrequently, no more than once or twice a week, similar to the frequency of meetings observed in programs such as Weight Watchers

[113]. These sessions are, however, much deeper and more extensive than what the Reflection Companion can currently support. The mini-dialogues are designed to provide brief moments of reflection, rather than support full motivational interviewing sessions. Given indications from past work that users of mobile activity trackers frequently engage in short awareness interaction sessions with their data within one day [93], along further feedback from the workshops, where active tracker users indicated checking their data on their mobile phone at least once a day, I decided to prompt users daily.

Providing Personally Relevant and Diverse Conversations around Reflection: To make the reflection dialogues engaging, I personalized the experience by introducing questions that referenced users' own behavior change goals using an introductory phrase such as: 11Hi Jake, you listed as one of your goals: 1taking regular breaks daily'... " after which a reflection question would be presented. The introductory phrases changed each time to provide for a more natural experience. Five mini-dialogues referenced users' behavior change goals. These mini-dialogues were template based and automatically used the user reported daily, weekly or long-term goal. Each dialogue also addressed users by name and employed a friendly conversational tone following indications from [131, 239]. Furthermore, in order to make the reflection focused and personally relevant, 17 mini-dialogues were delivered with a graph showing the user's physical activity metrics (15 plotting steps, one calories burned and one sleep). 14 of these graphs showed a week worth of data, 3 showed a comparison of two weeks of steps (see Figure 5). To provide an explicit link between the data shown in the graph and the reflection questions, these mini-dialogues would open with phrases such as: "Hi Kate, please take a look at your graph...". Such introductory phrases again varied each time to provide a more natural experience. Finally, to diversify the dialogues and to keep users engaged for longer and avoid boredom following indications from [30], I made the dialogues different in terms of the behavior change aspect (reflection topics) they addressed. Following the categorization from the workshop presented in Figure 4.1, 8 dialogues were related to observations/patterns, 6 to goals, 4 to plans/schedule, 3 to tracking and general/context,

and 1 to motivations. I also diversified them in terms of the starting reflection level - 11 started with noticing, 8 with understanding, and 6 with future actions - and question format - 15 were closed questions and 10 were open questions. This is on top of delivering some of the mini-dialogues with associated activity graphs

4.1.3 User Study

To evaluate Reflection Companion's performance, conversational design choices, and the ability to trigger reflection and encourage participation, I conducted a 2-week field study approved by the university's Institutional Review Board.

Participants: A total of 33 active Fitbit users (29 female, 4 male) between ages of 21 and 60 (M=36.5, SD=11.2) were recruited through social media. They used Fitbit for at least 2 weeks, were willing to provide access to their Fitbit data, and were willing to receive up to 4 SMS/MMS messages per day on their mobile phone for a period of 2 weeks. Participants logged 10,133 steps per day on average (SD=$6,521$, range: $1,768 - 36,757$) during the week before the study. Five participants logged fewer than 5k and 13 more than 10k steps per day. 19 of the 33 participants were interviewed after the study.

Procedure: At the start of the study, participants provided access to their Fitbit data. Then they completed a survey, in which they shared their daily, weekly, and long-term behavior change goals and indicated the time frame during which they would like to receive the reflection mini-dialogues. During the study, participants received one mini-dialogue per day over the course of 2 weeks, delivered to their mobile phones via SMS/MMS. At the end of the 2 weeks, participants completed a post-study survey. Finally, they were able to choose to use the system for 2 more weeks without additional compensation (I clarified their decision would not affect payment).

Measures: To assess the impact and success of Reflection Companion, I looked at measures of engagement. I looked especially at participants' willingness to use the system for an additional 2 weeks without compensation. Prior work indicates that continuous engagement intention is strongly related to perceived value and satisfaction with the system [125]. Participant interactions with the system were logged and analyzed. This includes the number of dialogues responded to, the time until a response was made (and whether a reminder was used), as well as the length and content of responses. These measures along with continued participation were used to assess engagement with the system. I further collected self-reported health awareness (9-item questionnaire adapted from [106]), level of reflection around self-tracking (Kember's 12-items [122]) and general mindfulness (13-items [229]). Changes in pre- and post- scale ratings were analyzed using paired t-tests. Further user replies to mini-dialogues over two weeks were analyzed and categorized. Semi-structured interviews (40 minutes on average) were conducted and audio-recorded following the study. Interviews were first transcribed and quotes related to each of the categories covered in the interview were extracted using a closed, selective coding approach following a general procedure for analysis of qualitative data described in [138].

4.1.4 Results

I present the results of the field deployment in the physical activity setting by looking at engagement measured by the system use behavior, impact on user self-reported reflection, and feedback from the interviews. Given that the system relied on limited NLU for user intent recognition, I also report system performance.

Engagement: During the 2 week main study deployment reflection companion sent a total of 462 prompts and 429 follow-ups, receiving 829 responses from participants. Participants responded to 96% of all initial questions and to 90% of the follow-up questions. While 11 participants responded to all questions, the lowest rate for participant responses to initial and follow-up were 23% and 64%, respectively. Overall response rate stayed fairly consistent,

indicating generally high engagement throughout the study. However, Figure 4.3 shows a decline in the length of response as the study progressed, decreasing from an average of 170.1 characters in the first week (SD=31.8) to 138.1 characters in the second week (SD=17.0). Participants took 50 minutes on average to respond to the first question and 13 minutes to respond to the follow-up. Reminders were sent in 39% of cases. Encouragingly, 16 out of the 33 participants elected to continue using the system for 2 additional weeks without reward. Furthermore, these participants continued to engage with the system at a high rate, responding to 83% of the initial questions and 76% of the follow-up questions during the additional 2 weeks. Average response length during the additional 2 weeks was 98.4 characters (SD=74.9).

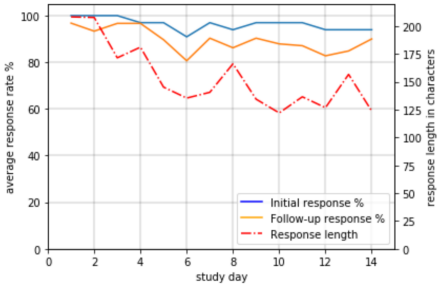

Figure 4.3: Response rates to initial, follow-up questions, and average response length in characters for 14 days of core study.

Mapping to the stages of a structured reflection process	Measures adapted from Kember [42]	Pre study	Post study
Stage 1: Noticing	Level 1: Habitual action (HA)*	3.16	3.53
Stage 2: Understanding	Level 2: Understanding (U)†	3.60	3.92
Stage 3: Future actions	Level 3: Reflection (R)	3.54	3.64
	Level 4: Critical reflection (CR)	3.60	3.85
Other measures		Pre study	Post study
	Mindfulness	2.52	2.63
	Physical activity awareness	5.63	5.73
	Step count (weekly mean)[1]	10,133	11,165

*Significance against the pre-study measure: * p<0.05, † p<0.1,*
1 -steps for a week before and after the study as our participants allowed us to stay connected to their Fitbit for an additional time.

Table 4.2: Summary of pre- and post study measures. The levels from Kember's survey are mapped to the stages of reflection in the structured reflection process.

System Performance: Reflection Companion relied on the NLU classification to categorize free-text user responses to select appropriate follow-up. For the 224 replies logged for these dialogues in the core two weeks of the study, more than 72% have been automatically matched with a known intent and resulted in presentation of a tailored follow-up. We coded the quality of the follow-up question into: Good match (follow-up question provided a good continuation of the dialogue), Acceptable match (follow-up question only partially build-up

on user question or required users to repeat some of the initial response), and Poor match (follow-up question made no sense in the context of user reply). Out of the automatically recognized intents, 95% of the presented follow-up questions would be a good (69%) or acceptable match (23%). This means that the system made very few "hard" mistakes, such as recognizing that the user expressed a negative impact of work on physical activity, where in fact the user described a positive impact. For the 62 (22.68%) cases where the system was not able to recognize any intent from user response and for which a non-tailored follow-up was presented, 92% of the presented follow-ups offered a good (58%) or acceptable match (34%), with only 8% of "hard" mistakes.

Impact on Reflection: Analysis of user responses to the reflective mini-dialogues provides numerous examples where dialogues were successful in supporting discussions around awareness related to goal accomplishment, self-tracking data, and trends in behavior: *"I like to be active on the weekend and it catches up to me on Mondays so I take it easy, then it's back to working out on Tuesdays and Wed."* Mini-dialogues also appear to have helped participants to better understand their behaviors and helped users draw connections between the step count and their context. Additionally, participants reflected on multiple higher-level aspects such as the value of physical activity, the meaning of a healthy lifestyle, the value of comparing oneself to others: *"My best friend is a doctor and has 3 kids and exercises way more than I do. (...) So sometimes I feel lazy when I compare myself to a friend, but most of the time I realize this is my life and comparing myself to someone else is not a mentally healthy practice, so I give myself grace."*. They also often reflected upon things that worked for them: *"Jogging helps me towards the goal of jogging a half marathon. Writing out my training plan on a calendar has been helpful."* as well as the things they could possibly change: *"Short runs before or after work. I enjoy running but I don't often make the time anymore. Standing at my desk more. Taking breaks not just at lunch. Getting a dog."* Aside from reflection, the dialogues provided additional benefits. For example, the prompts enabled users to vent: *"Annoyed that some of them are thin without even putting*

in that much effort. Sometimes annoyed that I can try so hard for less rewards" and also often served as additional reminders: *"Today is my first day back at work so I have not done it yet - will do it if I go to a diff floor".*

At the same time self-reported ratings (Table 4.2) indicate a significant difference in Habitual Action (HA) for pre (M=3.16, SD=1.06) to post (M=3.53, SD=0.89) study measurements; $t(32)=-2.0386$, $p< 0.05$ and a weakly significant increase in Understanding (U) from pre (M=3.60, SD=0.98) to post (M=3.92, SD=0.84); $t(32)=-1.8994$, $p=0.07$. The increase in Understanding level indicates an increase in users' analysis of the situation from different perspectives, formulating explanations and observations about the reasons for the things noticed. On the other hand the increase in HA is somewhat surprising because it reletes to activities performed habitually. One likely explanation for the increase in HA is that our system enabled a decoupling of the activity (here, physical activity) from reflecting on the activity (here, taking place when engaging with the system).

Interview Feedback

Types of Reflection Triggered: The 19 interviews confirmed and expanded on the results of the analysis of user responses to the mini-dialogues showing that the system was successful in triggering reflection on past activity patterns, on possible future actions and on new, previously not considered aspects.

Increased awareness: Ten participants reported that system increased their awareness of past physical activity. It specifically helped them realize how much they were recently doing and notice repeatable patterns in their own physical activity: *"It made me more aware that I am doing more steps when I'm at home and on the weekends. It just made me much more aware of how little and how much I'm doing on certain days."* (P8). Four claimed that the system helped them think about how they currently plan and allocate time to their activities: *"Got me to go back through my data and my calendar, and really stop and spend time thinking about, 'Okay, am I really prioritizing this or not?"'* (P14). Another four

reported that it led to them thinking about the relationship between activities, data, and the health outcomes: *"It opened my eyes to a few things... how my steps were affected by what sleep I had... and tracking my patterns on what days I did what."* (P10).

Alternatives and Future Actions: Eight interviewees reported that interacting with the system led to reflection on the actions they were currently taking to achieve their goals and made them critically re-evaluate these actions to think about possible alternatives: *"I definitely thought about whether I was doing as much as I could to be able to reach those goals. More about what were the barriers that were making it where I wasn't reaching those goals."* (P13). The prompts also triggered thinking about planning possible strategies to achieve enough physical activity based on what they have learned from the past: *"Partially, it's about reflection, but it's more of planning ahead, like what I should do and what I will do... by reflecting on the past behavior."* (P20). Such reflection was for many participants a prerequisite for trying out new behaviors.

New Insights: Four participants indicated that interacting with the system led them to reflection on aspects they had not thought of before, such as considering possible alternative metrics: *"It got me thinking about what other interesting metrics are there? I had never really thought about what I track or pay attention to that carefully. I just kind of use whatever the given dashboard is."* (P14). In other cases, it triggered critical thinking about how they currently use the tracked metrics, and what they can learn from them. The system also introduced new ways to evaluate data by presenting them in a different timeframe (e.g., two weeks): *"It was my first time to see an overview of my weekly activity... I had never done it before. Thinking in a way of a week cycle was interesting... Thinking of two weeks in parallel, is there any seasonality or any cycle."* (P20)

Benefits of Reflection: Reflection was beneficial in many ways: it increased motivation towards physical activity, introduced changes to participants' actual behavior, increased

mindfulness, and encouraged formulation of more realistic strategies for increasing activity.

Increased Motivation: Participants found the reflective dialogues to be motivating. Five reported that the mere presence of the prompting mechanism provided focus, kept them in check, and consequently led to increased motivation. In some cases, the daily presence of the dialogues created a sense of accountability, which provided additional motivation: *"They were a form of encouragement to me, because it's like I knew that there was accountability on my part, that if I had a poor day that I had to explain why, reflect on that on, what would I do the next day."* (P22). Eight further reported that the dialogues helped them realize their barriers, formulate clear action plans and define small, concrete and attainable steps for achieving their goals. Interviewees considered these aspects to be motivating: *"It was like 'What little changes could I do?' And that was helpful 'cause like making the time for an hour workout every day seems daunting, but going for a walk on my lunch is doable. Going for a walk after work is doable."* (P25).

Leading to New Behaviors: For many participants, engaging in reflection resulted in the adoption of new behaviors. These behaviors were usually small changes to daily routines, such as parking further away from office or parking meter to walk more, walking to a grocery store instead of taking a car, or using stairs instead of an elevator: *"I actually did little things to make myself more active during the day. The prompts got me like, one day I'm talking about walking more during break, and so since then I've made a point to get out of the office and walk during my lunch. Just doing little things."* (P25). In some cases, the dialogues served as an additional push on top of a request from a family member, e.g. a request from participant's daughter to go for a walk or an evening walk with wife in case of another participants. In some cases, the prompts also triggered a return to past behaviors that have been abandoned: *"It actually got me to get back into running, which is what I had gotten out of for a little while so that was kind of nice."* (P24). In a number of cases, the mini-dialogues led to behaviors that facilitate physical activity, such as wearing Fitbit more

often, downloading an additional app for tracking running progress or scheduling a class at the gym: *"After I would get the message, if I hadn't already scheduled class at the gym for that day, it would usually be a good reminder."* (P14).

Increased Mindfulness and More Realistic Plans: Six of the interviewees said that the mini-dialogues helped them better assess their progress and become more mindful of their own tendencies and inclinations: *"I realized something about myself that I like to work out...[by doing] another activity. For example, going to the museum."* (P14). In many cases, this led to an increased understanding of factors that help participants meet their goals, or barriers that prevent them from doing so: *"I guess just becoming more aware of the barriers to some of the stuff keeping me from my goals."* (P26). This helped interviewees realize the need for specific and realistic actions to achieve their goals: *"I think it helped me be more realistic. A lot of times where you're like 'Oh I can do this in a month or something like that.' But in reality, it's a lot tougher so it's nice to have that reflection"* (P24).

Impact of System Features: Additionally I explored the impact of key elements of the system on user experience: the two-step mini-dialogue structure, continuous reflection through daily conversations, the need for typing & sending a response, and personalization using the activity graph of personal Fitbit data as summarized in Table 4.3.

4.1.5 Discussion

In this work, I argue that a conversational approach, using what I refer to as "mini-dialogues" design, can be effective in eliciting reflection. Indeed, in the deployment, *Reflection Companion* conversational agent successfully led to reflection at three levels: awareness, understanding, and new insights for the future. I show that such reflection can help users become more motivated and can lead to defining action plans better aligned with users' long-term goals and actual abilities. Here I further discuss some aspects of the approach.

	Aspects of mini-dialogue based reflection guidance		Reflection frequency: One dialogue a day	Aspects of Personalization & diversification
	Two-step mini-dialogue structure	**Typing and sending responses**		
Positive aspects	Extends thinking time for reflection	Promotes deeper thinking, seriousness, and precision	One dialogue a day just right: allows for reflecting on continual progress	Graph useful for supporting response closely tied to the personal data
	Encourages deeper thinking and more meaningful answers	Creates a sense of commitment and accountability	Enables devoting the whole day for reflecting on one aspect, which was appreciated	Prompts useful for bringing attention to the data aspects not considered before
	Having two smaller questions lowered the reflection effort	Helps remembering, serves as a mental note	Useful as a momentary trigger and check-in	Personal data promoted engagement and motivation
Observations /Challenges	Some follow-up questions felt generic and computerized	Required additional effort from the user	Aspects discussed between dialogues are sometimes repeated	Despite diversification graphs and questions felt similar

Table 4.3: Summary of the positive/negative aspects of the system design choices based on feedback from participants.

Benefits and drawbacks of reflection on physical activity: I have shown that reflection helps increase awareness, mindfulness, and triggers consideration of new aspects. This is supported both through the interview data, as well as the pre-post study increase in "understanding" rating. I also found that reflection activities can serve as a prerequisite to better goal setting and more feasible future actions. While most of the participants did not revise their physical-activity goals during the study, many reported that the 2-week period was too short to compel such a revision. I found, however, that reflection serves as a preparation for considering new goals and feasible future actions. Further, I found that reflection provides a non-judgmental, neutral interaction that was appreciated by many participants. The reflection activities offer participants a break from the often judgmental and persuasive nudges built into current behavior change systems. Nevertheless, for some, a concern was

that reflection activities are not necessarily actionable. Finally, I should note that reflection might potentially lead to discouraging revelations (e.g., less activity than expected) as noticed in the exploratory workshop. Encouragingly, I did not notice any indications of the mini-dialogues having such negative effects during the field deployment, but this still remains a remote possibility.

Insights About Designing a Conversational Agent for Reflection: Through the study, I uncovered three key benefits of the conversational approach to reflection. One is that it has an ability to actively shape the direction of user thinking. I found that the mini-dialogues, through building-up on user responses, have an ability to guide user thinking in a specific direction. I also found that having multi-step conversational exchanges extends the time a user spends reflecting and that everyday conversations can help users learn over time. Last but not least, the conversational approach provides an engagement boost through perceived accountability and commitment (even if the user is aware of talking to a computer system). The act of typing and committing to an answer brings benefits of precision in planning, deeper thinking, and accountability.

However, there are also drawbacks of using a conversational approach. First is that doing so runs the risk of building-up and disappointing user expectations. Second, conversational interfaces are at least currently harder to design for; more effort and resources are required. One key challenge with building a conversational system for reflection (or a conversational system in general) is to generate a set of sufficiently diverse and topic-appropriate dialogues. This is especially important for the purpose of continuous, everyday coach-like interactions.

Extending the Long-Term use of Reflection Companion: In order to make the dialogues even more engaging, especially for longer-term use, a number of potential approaches such as diversification, tailoring, memory & adaptation can be explored. Diversification focuses on making the dialogues novel each time. It can be applied on syntactic (sentence composition), semantic (topics), and dialogue structure level. Diversification, however, does

not build up on past exchanges or increasing knowledge collected about the user to make the conversation more engaging. Another approach involves improved personalization & tailoring. Reflection Companion used personalization by addressing the user by name, presenting a plot of personal data, and weaving in user goals into selected mini-dialogues. The topics introduced by the dialogues were, however, not tailored to the user's interests in any way. Future work could explore tailoring on the level of topics of interest using e.g. Schwartz's 10 basic values, representing universal motivational constructs [209] which I have used in my work described in Chapter 3. Yet another option could be tailoring the dialogue structure itself, which has been explored for cultures [62]. Arguably most valuable for long-term, but also most technically challenging, would be to remember aspects users shared and adapt the mini-dialogues to include those aspects. Currently user response to the initial prompt is classified and "remembered" only to decide on the follow-up to present. Unfortunately, no long-term memory or common ground is retained. This requires asking each time e.g., what is the user barrier for a specific goal or activity, or having to switch to a new topic to avoid repetition. Remembering information from user past responses has obvious long-term benefits: it allows to deepen the reflection on relevant topics over time, it communicates to the user that the shared information is appreciated, and it partially addresses the issue of topics exhaustion as dialogues can also go in depth on one topic over time.

4.1.6 Conclusion

In this work I introduced a mobile phone based conversational system for supporting reflection on everyday physical activity - Reflection Companion. The system prompted users daily to engage them in reflection on various physical activity related topics. Interaction was in the form of mini-dialogues incorporating user's personal goals, and activity graphs (from Fitbit tracker). The conversation questions were generated via workshops with activity tracker users and informed by structured reflection model to help fuel content diversity. In a 2-week deployment with 33 users I have found that Reflection Companion offered an engaging interaction, with half of the users electing to actively use the system for an additional

2-weeks outside of the study without any compensation. The system was also successful in increasing user awareness, supporting reflection on activity alternatives and future actions, and prompting new physical activity insights. Users reported feeling more motivated, mindfull, and encouraged to try new behaviors. On top of that, the users linked the majority of the reported benefits to specific conversational design elements. They attributed deeper and longer reflection and more truthful sharing to *two-step mini-dialogue* design. They linked the ability to build-up on prior reflection and the lower cost of reflecting to *continual daily interactions*. Further, they connected the need for deeper thinning, precision, as well as the sense of commitment and accountability to having to *type & send responses* knowing that "someone" is reading them. Finally they attributed the sense of progress, focus, attention, and personal relevance to the weaving in of *personal activity graphs and individual goals* into the conversation. The system offers empirical evidence for the value of conversational reflection, proposes a design process for feasibly realizing such design, and further offers insights into promising future directions for most impactfull improvements.

4.2 Workplace Productivity Setting

4.2.1 Background

For knowledge workers in companies, keeping track of work activities and accomplishments can be a useful practice but one that can be hard to sustain. Awareness of one's own activities, and reflection on aspects of learning at work are important for professional development [205] and can lead to tangible performance improvements [64]. It builds worker confidence in the ability to achieve goals [64], improves the depth and relevance of individual learning [168], supports emergence of self-insight and growth [166], and consequently leads to performance increases [115, 250]. Performance increases are said to come from understanding of the causal mechanisms behind actions and outcomes [250] and by learning from accumulated past experience [64].

Challenge to Reflection at Work

Yet increasing time pressures in the modern workplace make taking time to step back and engage in efforts to learn from one's prior experience seem like a luxurious pursuit [63]. Employees would rather decide to gain additional experience doing the task than take time to articulate and codify what they learned from prior experiences. In fact this kind of 'doing more' behavior is still encouraged in many workplaces [64]. Finally, reflection itself is time consuming and not necessarily something that comes naturally to people, they usually need a reason to reflect or at least an encouragement to do so [98, 168]. Supporting reflection through computerized systems has been identified as a vital field of research [22, 153] with computer-supported reflective learning specifically in work settings being identified as crucial [134]. Still, few systems exist for supporting reflection in the workplace.

Potential for Conversational Support

To help with professional development and learning from work activities, institutions of career counseling and development exist in bigger companies [19] as well as outside of company structures [189]. Conversational agents, whose use is growing in popularity, stand to play an important role in supporting behavior change and well-being. While chat bots and other "virtual assistants" have been motivated by, developed, and tested in a variety of contexts from customer service [62, 181, 240] to health-related behavior change [27, 186], to simulated job interviewing [145], our focus is on the role of conversational agents for organization, productivity, and self-learning in the workplace. In such settings, user needs may be different and avoiding disrupting work and improving efficiency are important. Prior work also identified potential benefits of talking to an agent instead of a human in contexts where people are less afraid of being judged and more willing to disclose [155]. In this chapter I therefore explore: How can we design conversational experience to support reflection in the work context?

Interaction Modalities of Conversational Agents

Furthermore there are indications about the differences in the impact of modalities on user behavior and perceptions. In movie recommendation context spoken queries were longer and more conversational, with more subjective features than typed queries [116]. A study on using voice for providing edits and comments in writing tasks, showed that voice-based comments may be easier and more natural to leave (as opposed to text) from the point of view of an editor, and also that people leave different types of comments using the two modalities [174]. This combined with the fact that a recent poll [94] showed an increased adoption of voice interfaces, with 63% of Americans surveyed using voice assistants such as Apple Siri, Google Assistant, or Amazon Alexa makes it interesting to explore the use of voice modality for reflection as well. I therefore also explore: What are the differences between voice and text-based conversational reflection?

4.2.2 Design Approach

The design process for the Robota conversational agent involved three parts: 1) generation of meaningful reflection questions for the workplace context based on existing knowledge on reflection in learning [168], education settings [8], behavioral questions from job interviews [222], and career development sources [226], 2) designing dialogues combining work activity reporting and reflection for voice and chat modalities, 3) designing supporting personal informatics elements: dashboard and reminders

Literature-based Content Creation

For the conversational reflection in the productivity settings I generated a collection of work-related reflection questions inspired by structured reflection theoretical frameworks such as Moon's reflection in learning [168], Gibb's reflective cycle [89] and Bain's 5Rs framework [16]. I also drew from concrete examples of reflection questions in educational settings [8], behavioral questions from job interviews [222] and career development sources [163].

I attempted to cover the following categories with my questions, aiming at encouraging workplace reflection:

Task-related questions: These questions ask about tasks and activities and how aspects of these tasks and activities may contribute to learning; for example: *"How can you make the activities you planned for today more enjoyable for yourself?"*

Planning and organization: These questions focus on understanding factors affecting performance and learning points from organization of work in scope of a day as well as the week; for example: *"How satisfied are you with how you organized your work today? Is there anything you have learned?"*

Short-term and long-term activities and goals: These questions focus on realizing relations between activities and goals, barriers to goals accomplishments, as well as on exploring the value of having a longer-term goal; for example: *"Do you feel the activities you did today contributed to your goals? Why or why not?"*

Motivation and satisfaction at work: Questions in this category triggered exploration of sources of positive and negative emotions at work as well as moments of satisfaction; for example: *"What were some of the most satisfying moments at work for you this week and why?"*

Personalized questions: Questions in this category include dynamic elements extracted from user's work journal entries; for example: *"Did ¡task¿ help you learn anything new that could be valuable for the future? What did you learn?"* Past work identified the use of record of events as one successful way to enhance reflection [111]. Such a record can be looked at again to provide time and focus attention on different aspects of the experience on each return, especially if some guidance as to what to focus on is provided [216]. These ques-

tions further highlight the link between the journaling activity over Slack[3] and a continued engagement through the reflection questions.

Conversational Agent Design

I designed and implemented a custom conversational agent called Robota (which stands for "work" in Polish) to support workplace journaling and reflection. Workers interact with Robota through chat and voice, and can explore past interactions through a web dashboard. Figure 4.4 illustrates the overall architecture of the system: The core Robota logic is implemented in the cloud as a timed state-machine using Python's Flask[4] and SQLAlchemy[5] frameworks on top of MySQL database[6]. This common backend supports the chat and voice modules as well as the web dashboard, described later.

Chat Modality (Slack-bot): I implemented Robota's chat module as a "Slack bot" via the Slack API. The bot has the ability to send and respond to direct messages on Slack (a Slack bot appears just like a person on Slack, appearing in the user's contact list). A journaling prompt, illustrated in Figure 4.5, consists of an introductory message followed by a request for accomplished activities. Robota then asks the user to record her plans. The user responds in open, unconstrained text. In addition to journaling, the chat module is responsible for delivering chat-based reflection questions, and for prompting the user to perform voice-based reflection (described next).

Voice Modality (Amazon Alexa Skill): I used Alexa Dash Wand[7] - a handheld cloud-connected device with a built-in speaker and microphone that allows the user to take it to a quiet room and speak to it discreetly. The Dash Wand supports Alexa Voice Service

[3]http://slack.com and https://api.slack.com/bot-users

[4]https://www.fullstackpython.com/flask.html

[5]https://www.sqlalchemy.org/

[6]https://www.mysql.com/

[7]Dash Wand - https://www.youtube.com/watch?v=s7IExS483wE

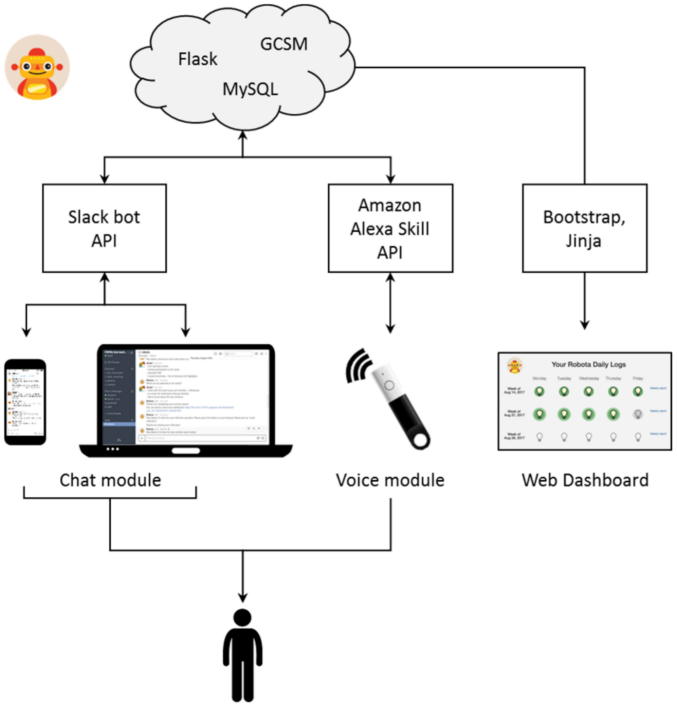

Figure 4.4: System architecture of the Robota conversational agent. A common backend supports chat interaction as a Slack bot and voice interaction as a custom Amazon Alexa Skill using an Amazon Dash Wand.

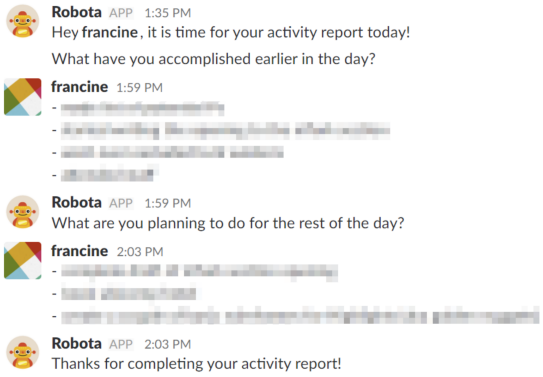

Figure 4.5: An example of interaction with Robota using the chat module, in this case, a mid-day journaling prompt.

Journaling time	**Questions** (through the chat module)
Morning (10am)	- *What have you accomplished yesterday?* - *What are you planning to do today?*
Mid-day (1:30pm)	- *What have you accomplished earlier in the day?* - *What are you planning to do for the rest of the day?*
End-day (4pm)	- *What have you accomplished today?* - *What are you planning to do tomorrow?*

Table 4.4: Work activity journaling prompts for different journaling schedules (schedule selected by the user).

(AVS) and custom-built apps (called "Skills"). I implemented a custom skill using Amazon Alexa Skill API [8]. To prompt the user for voice reflection, Robota sends a Slack message asking the user to initiate reflection. This is to link the slack journaling and voice reflection as parts of the same system, but also to give the user freedom to initiate voice interaction at a convenient time. The user then holds down the Dash's button and says "Start Work Reflection." Robota speaks one of the reflection questions (described later) and listens for the user's response. The user may ask Robota to repeat the question. Robota speaks a 'thank you' message in voice, and also sends a 'thank you' message on Slack. Finally, both chat and voice interaction collected information are collected in the user's dashboard, described next.

Supportive Personal Informatics Elements

Web Dashboard: To allow users to review their work journal entries and their responses to reflection questions, I implemented a web based dashboard. The dashboard uses badges to represent each day to encourage continued participation. Reviewing journal entries and reflection for a specific day is done by clicking on a badge. Due to low performance of speech-to-text services, for user responses through the voice module, I provide links to the original voice recording instead of a (likely faulty) transcription. Finally, to support sharing work reports with others, the dashboard includes a link to a weekly compilation of all journal entries.

Chat-based Reminders: An important aspect in designing successful conversational agents for the workplace, is to balance engagement and interruptions. Since reflection questions were designed to follow and, in some cases, rely on journal entries, I implemented a reminder strategy that used long and growing time spans for subsequent reminders.

[8]Alexa Skills Kit - https://developer.amazon.com/alexa-skills-kit

4.2.3 User Study

I conducted a 3-week, within-subjects controlled deployment with 10 participants from a company lab (3 female; 7 male). Five participants were between ages 25 and 43, three 35 to 44 and one for each age group of 18-24 and 45-54 years old. None of the participants were involved in this research project. Participants included three research staff, four interns, and three developers/support and represented a diverse set of accents: English native for only 2 out of the 10 participants, the rest included Japanese, Chinese, and French. This is particularly challenging given "Robota" used voice in one of the conditions.

Procedure: At the beginning the participants completed a short survey and each participant also chose when they wanted Robota to prompt them to journal their activities and plans, between morning, mid-day, and end-of-day journaling (as described above). During the first week of the study, participants used Robota for daily journaling only, through Slack (Journaling-only condition). In the following 2 weeks, participants would respond to reflection questions (10 questions total: one question a day, for two 5-day workweeks) through chat (Chat-Reflection condition) for one week and using voice (Voice-Reflection condition) in the other week in a counterbalanced order. At the end of each week, participants were asked to compose a weekly report and respond to a survey. Finally, participants completed an end-of-study survey and took part in a short interview.

Measures: On the Friday before the beginning of the study, participants were asked to write a weekly report summarizing their work activities, and evaluated the difficulty of writing the report, the report's clarity and level of detail. Every Friday afternoon throughout the study participants similarly wrote a weekly report of their work activities and provided ratings. In addition to weekly reports, participants responded to questions regarding their interaction with Robota during the week. At the ends of weeks 1, 2 and 3, these included questions about the journaling activity. For example, *"Did logging your daily activities influence your work? If so, how?"*, *"Did logging your daily activities influence writing the*

weekly reports? If so, how?". And 7-point Likert scale: *"How easy or difficult was it to log daily activities?"*. At the ends of weeks 2 and 3, these included questions about the modality they used. For example, open-ended: *"What are the main things you liked about using the chat bot to reflect on your work?"*. 7-point Likert scale: *"How easy or difficult was it to respond to the reflection questions?"*. For the final survey, at the end of week 3, participants were asked about the value of reflection: *"What benefits, if any, did you get from reflecting on your work (using either the chat bot or Alexa)?"* and to directly compare their interaction with the voice and slack channels: *"Considering the two methods for reflecting on work (the chat bot and Alexa), please compare your experience of the two."*

4.2.4 Results

I present the results of the field deployment in work productivity setting by looking at system use, user reported value of work journaling support, impact on work reflection, and differences in the impact of voice and chat modalities.

System Use

Participants used the system consistently throughout the study, responding to 99% of the activity journaling and reflection requests. Responses arrived within a median of 31 minutes. Robota sent a total of 174 reminders for journaling. Robota also sent 98 requests for reflection followed by 59 reminders (34 in the Chat-Reflection condition and 25 in the Voice-Reflection condition). The average length of a daily activity log was 292 characters (SD=239.62). The average length of a response to reflection questions using chat modality was 131 characters, compared to 98 using voice modality.

Value of Work Journaling Support via Chat:

Through analysis of the end of week surveys and interviews, I found that all participants rated journaling as useful for composing weekly reports. Participants reported that the daily

activity journaling helped them directly with work tasks by: 1) increasing their awareness and productivity and 2) helping with composing reports. Still, a number of challenges to journaling surfaced, mainly: no tasks worth recording for a day, lack of progress perception, and duplicate entries for long-running tasks.

Increased Awareness & Productivity: Three participants reported that journaling increased their thinking about their daily activities and work organization as well as lead to increased awareness of progress: *"Sometimes it made me realize that there was little progress on some days"* (P9). Two others felt that journaling positively impacted their productivity, this was mainly through the aforementioned awareness of limited progress: *"If I found I didn't make much progress on a day, I would try to do more on the next day."* (P4) or through concern that they will have nothing to report at the end of the day: *"Maybe more productive. I don't want to have nothing to be logged at the end of a work day."* (P10). Five other participants, when asked directly in a post-study interview, reported that journaling had no specific impact on their work awareness and productivity. In one case it was because the participant already regularly journaled her activities (P6). In the other four cases, participants did not feel a direct impact on their work, as journaling itself didn't suggest concrete changes. They, however, still reported an indirect impact, such as help with keeping track of time and tasks (P3, P8), assistance with work organization (P5) and help with deciding on the relevant tasks to pursue (P2).

Helped with Composing Reports: All the participants considered daily activity journaling useful for composing weekly reports. For eight individuals, activity journaling helped by making it easy to recall things done throughout the week: *"I didn't need much effort to remember this week's activity because I logged it on Robota every day."* (P7). Some also felt it helped them make sure they did not miss any important points from their reports: *"I can refer to these logs to have a better summarization without missing important points."* (P4). For four participants, daily logs served directly as a source material for copy-pasting

relevant items into their weekly reports: *"I simply picked the important points from the daily reports and used them."* (P2). For two people, daily logs helped with organization of their reports: *"Yes, I think it helped me to remember and organize what I have done."* (P9). Finally, for two more participants, having all the relevant information about their activities in one place helped them avoid collecting information from various sources: *"It was easier to compose from Robota logs because I didn't need to go back and forth within different sources for collecting my activities."* (P7).

Work Reflection with Robota

Eight participants rated the act of answering reflection questions as useful, somewhat useful or neutral (eight in chat, six in voice, and six in both). Comments from the interviews suggest that reflection aspects of the system helped participants: 1) improve work organization, 2) look at their work from different perspectives and even 3) consider higher level goals of their careers.

Improved Work Organization: Three participants mentioned that the reflection prompts made them think about how they organize their daily activity: *"It makes me think about the efficiency, the organization, and other things. This will further help me increase my efficiency."* (P4). In some cases, it also helped with planning activities and making sure that important things are not forgotten: *"Remind me that some things are needed to do."* (P5).

Helped Gain New Perspective: Six participants indicated that reflection with Robota gave them opportunities to think about the value of activities they perform: *"It made me keep track of what I have learned from my work, which was different from what I usually write on daily reports"* (P9), or encourage new ways of thinking about work: *"Robota pointed out what I haven't thought ever and it was a good chance to think about it."*(P7). Finally, they also reported that it was valuable to find some time to think more deeply about their activity:

"Helps me take a moment to be reflective, almost meditative, during the day about the process of how I work instead of just thinking about the content of the work." (P6).

Helped Consider Higher-level Goals: Three participants also discussed how Robota helped them think about the meaning behind their work: *"Force me to think about the impact of things I did."* (P5). Reflection also helped some participants consider their higher-level goals at their current workplace: *"Reflection questions lead me to think about what brings me satisfaction, what I have learned. It was helpful for considering my goal at [company]."* (P7).

Challenges with Reflection Questions: Not all the reflection questions were seen as equally valuable. A number of questions were considered too abstract and hard to even answer: *"The questions are too general and sometimes hard to have a specific or informative answer."* (P10). The flexible and unscheduled nature of some participants' work made questions about planning and organization irrelevant. A participant whose main job is to offer technical support for others said: *"So far, I haven't found it very useful to do work reflection, mainly because my daily task(s) are pretty ad hoc and the question posted to me may not be very relevant."* (P1). Four participants appreciated questions that explicitly referenced their logged activities: *"My favorite reflection questions were the ones specific to my daily log."* (P2). However, personalized questions may sometimes incorrectly ask about tasks that are not as meaningful: *"I felt that some questions were too specific and I often didn't have anything meaningful to reflect on related to the question asked."* (P2).

Designing for Voice vs. Chat:

A key goal of our work was to explore the specific value and limitations of voice and chat modalities in the workplace. Looking at self-report measures, a paired samples t-test shows that responding through voice was seen as less easy($M=2.6$ vs. $M=4.0$; $t(9)=5.62$, $p< .001$) and more annoying ($M=4.3$ vs. $M=3.2$; $t(9)=-2.28$, $p=0.05$). Participants' complaints

about voice modality mostly stem from (known) limitations of voice-to-text transcription and limitations of the Dash Wand in particular. Nevertheless, a number of comments revealed a potential value of voice modality that looks past current technical limitations.

Advantages & Challenges of Voice Modality

Separate Channel for Reflection Valuable: Four participants considered the ability to use a separate voice channel for reflection useful, mainly due to being able to quickly capture some of their thoughts: *"It's good to have another means to quickly capture some useful points or thoughts."* (P1). Three participants also considered interaction via voice as being more like having a personal conversation with someone that cares about them: *"[voice] has a slightly more personal feel to it"* (P4), *"This interaction is nice. I felt like Robota is caring about me."* (P7). This feeling even led two participants to consider the voice-based agent as more of a counselor or even a machine they could share with: *"It does make it feel more, it makes me feel more reflective. Almost like a counselor or a therapist."* (P4), *"At the moment I am unhappy. That's the moment I want to complain and the machine gives me an opportunity to complain and that's very good."* (P8).

Easier to Answer Questions with Voice: Two participants felt they could generally answer questions faster with voice. They appreciated that they didn't need to type anything while answering: *"It doesn't take much time to answer, is easier than writing report on Slack."* (P7).

Interactive, fun and engaging: Still, the fact that the reflection questions were revealed only after interacting with the Wand had the potential to be more engaging and even fun: *"It was kind of neat to use the wand and have the voice reveal to me what the mystery reflection question was."* (P6), *"Talking to a machine is somehow fun."* (P10).

Perceived pressure to respond immediately: Although participants were told they could listen to a reflection question and then call the skill again after some time to respond, most felt the pressure to respond immediately after being asked: *"While using voice, it seemed to encourage me to answer right away, which is a bit stressful"* (P10). Such need to respond quickly made people feel they had less time to think about their answers: *"You also have less time to think while speaking it aloud. So I'm not sure if the essential points are captured."* (P4).

Listening to own responses inconvenient and uncomfortable: Two individuals felt that reviewing voice-based responses afterwards was not ideal: *"It is not transcribed and listening to what I said many times is somehow troublesome."* (P9). There was also a dislike for hearing one's own voice played back: *"Chat-robota was easier to review my answers after logging. (Sorry I felt uncomfortable to listen to my voice...)"* (P7).

Advantages & Challenges of Chat Modality

Easier to read questions, think about response: Half of the participants felt that it was generally easier and faster to read the question: *"Reading is much faster than listening."* (P9). They also felt they could take more time to re-read the question if needed, think about it, and then respond: *"It was easier to read the question and think about it"* (P2).

Easier to reply in own time and describe details: Seven participants felt that chat-based interaction allowed them to enter their responses at their own pace: *"As you type in, you can pause and think."* (P4). They further felt that typing makes it easier to describe the details. As most of our participants were non-native English speakers, this perceived ease of typing sometimes came from the contrast with having to describe things in voice in a foreign language: *"It's easier to answer than explaining in a voice. Since my English is not so good, I couldn"t answer to a question immediately if I have to speak."* (P9).

Easier to review and change responses: Three participants liked how typed responses were editable: *"I also could more easily change my response with the chatbot before submitting."* (P2). Also, having their reflections in text made it easier to review afterwards using the dashboard.

Typing is time consuming: Still, needing to type responses made some participants write more concisely: *"Sometimes the answers to the questions are a bit complex, but I write something that is simpler and reductive because I don't want to spend time detailing it out on slack."* (P6)

Slack seen as less personal: Two participants mentioned that reflecting on Slack, as compared to voice, felt less like having a conversation and more like formal reporting of activities: *"It is slightly less personal [Slack], maybe the voice felt a bit more personal"* (P4), *"Typing on slack is slightly more formal I guess, it is something that goes into the record"* (P7).

4.2.5 Discussion

The field study provided some initial insights into workers' behaviors and reactions to using a conversational agent via different modalities. Participants generally appreciated having a structured way of reflecting on their activities for planning and goal-setting. Unlike many existing workplace reporting tools, my design supported workers' individual work styles by including journaling prompts for different parts of the workday. Some participants chose mid-day journaling to encourage themselves to be more active.

Interacting with the agent via chat (as designed in my system) made non-native English speakers feel they could more easily read and respond to the questions. At the same time, interacting with the agent via a separate voice channel had the potential to be more engaging and personal (e.g. voice modality seems more suited for complaining and being more reflective). These add new dimensions to consider when designing for behavior change. Here

we provide further design considerations for future work based on the findings from our field study:

Combining the benefits of both modalities: For the purposes of the study, I limited users to only interact with the voice or chat modality for one week each, and saw that each modality had pros and cons. However, outside of a controlled study environment, users could be provided the opportunity to choose which modality they wished to use on a day-by-day basis, based on their current context at the time of journaling. Additionally, the system could rely on contextual cues to prompt the user to log and reflect in one modality versus another based on what it infers to be the most appropriate form. My findings further suggest that certain reflection questions may also be better suited for certain modalities. For example, questions that are more personal or require a deeper level of reflection may result in more valuable reflection activities when using voice-based input.

Integration with the work setting: Many participants mentioned that one benefit of using text interaction within Slack was that it was seamlessly integrated with a tool and platform (on their computer) where they were already doing much of their work. Perhaps for this reason, using a personal device that takes a person away from their desk to speak out loud and reflect upon personal topics may be less-suited for information workers whose day is primarily carried out on a computer in a public or semi-public space. The benefits of a mobile or portable solution for journaling and reflection may be greater for different types of workers, where daily activities are more mobile and occur in different settings; for example, people who engage in site visits or inspections, or frequently travel to visit customers on sales calls.

4.2.6 Conclusion

I introduced Robota, a conversational agent for workplace journaling and reflection that combines chat and voice interaction using a common backend. The three-week long deployment of Robota with knowledge workers, revealed numerous benefits and challenges of conversational reflection in the work setting. Robota successfully engaged workers in activity

journaling increasing their awareness of work progress, productivity, and helping them with composing reports. Robota's reflection questions helped workers improve work organization, gain new valuable perspectives around their everyday work tasks, and even engaged them in thinking about higher-level career goals. At the same time I identified several challenges with literature-based reflection question generation leading some users to consider questions to be too abstract, irrelevant for their specific context or relating to tasks they didn't find valuable to reflect on. These reveal context matching challenges and provide opportunities for future design improvements. Furthermore, the comparison of reflection via voice and chat modalities highlights tradeoffs between the modalities and points to areas likely to benefits from intelligent sensing. The chat modality was reported to support easier reading & thinking about the reflection questions, easier replying in one's own convenient time, an ability to describe details, as well as reviewing responses and making changes. On the other hand, this modality was considered more time consuming due to typing and less personal than voice. At the same time, the voice modality was reported as a valuable separate channel just for reflection that is also more personal and caring. Furthermore voice afforded an ability to quickly capture thoughts, respond faster to reflection prompt, and offer a generally fun and more engaging experience than chat. On the other hand, voice introduced more pressure to respond immediately as well as inconvenient and uncomfortable need for listening to one's own responses (when revisiting reflection). My work provides practical design for supporting conversational reflection at the workplace, tackling several challenges related to semi-public space and use of voice. I offer insight into benefits of reflection via different modalities and identify future improvement directions.

4.3 Discussion on Supporting Conversational Reflection in Both Settings

Reflection Companion and Robota were both conversational agents designed to engage users in reflection. While sharing a common purpose, they were each designed following a different content generation process, the were deployed in different settings and interacted with users via different modalities. These similarities and differences offer several opportunities to make

comparisons.

4.3.1 Comparison of Impact

Both agents were generally successful in engaging users in reflection on physical activity and workplace productivity respectively. Users of both Reflection Companion and Robota reported benefits of *increased awareness* of their activities specifically in relation to the progress they are making. In both settings conversational reflection also helped users *gain new perspectives*. In physical activity this related to critical examination of their current behavior and alternatives for physical activities, while in the work setting this related to new ways of thinking about work and value of work tasks they perform. With both agents, it seems that users also benefited from increased *mindfulness*. In the physical activity context this amounted to being aware of ones tendencies, motivation, inclination and barriers, while in work setting this related to reflecting on meaning behind work and higher-level carrier goals. Interestingly, some of the more nuanced use of the agent, e.g., venting, was shared across the settings as well.

4.3.2 Comparison of Conversational Reflection Design Approaches

The topics of conversation in Reflection Companion were generated via workshops with active users of physical activity trackers. The generation was guided by structured reflection model. Reflection questions in Robota on the other hand, relied on past literature on work reflection as well as on multiple less formal counseling and interviewing resources. While it is hard to make strong comparison claims in terms of quality impact these different generation processes might have had on user engagement, it is worth noting that several users of Robota complained about questions being too abstract to answer or non-applicable to their type of work. Such issues could have been due to the conversational reflection prompts not being sourced withing the target user population, but relying on on-size-fits all generalizations about the work context.

4.3.3 Impact of Modality & Interaction Channel

Robota explored slack and voice based interaction, while Reflection Companion relied on mobile SMS/MMS text messages. Users in the work setting considered voice interaction to be more personal & feeling more like talking to someone than slack. In physical activity setting the mobile SMS interaction was also reported by many as personal, due to the channel being used mostly for communicating with friends & family. On the other hand slack was considered much more formal and led users to spend more time on 'refining' their answers. Despite the fact that typing, rather than speaking afforded by slack, seems like an interaction more similar to mobile text-messaging, the impact seemed different. Slack was more associated with work tasks & within office communication. It seems that perception of a separate channel dedicated to personal reflection was the important distinction, not necessarily the interaction style itself. There were, however, differences in how people interacted with voice, specifically they felt a pressure to respond immediately, but they were also more spontaneous in what they shared. Finally voice had a quality of being perceived as more interactive, fun, and more engaging. Typing on the other hand offered more of a commitment in text and something that goes 'on-the record'.

4.3.4 Private vs. Semi-public Space

The physical activity reflection via Reflection Companion took place on user's personal mobile device, withing the hours specified by the user (which could be during or outside of work hours) and revolved around individual health-related goals and activities. On the other hand, reflection on work activities via Robota, was much more tied to the work context - interaction took place withing office hours, on work related tasks, and partially on work related medium (i.e., Slack). In the design of Robota, I was aware that given such setting, users might be less inclined to treat the reflection as a personal benefit rather than another work chore. On the other hand, they may also want the reflection time to be tangibly beneficial to their work (as it 'consumed' their work time). From the results it seems that having Robota

support work activity journaling & reporting (work support benefit) as well as providing reflection prompts of more personal nature via a separate voice channel (personal benefit) was a generally successful design strategy for engaging users in reflection at work.

4.4 Summary of Contribution

In this chapter I examined a conversational approach for engaging users in reflection on physical activity as well as on work tasks & productivity. Despite reflection being an important step of behavior change [142, 73] offering multiple benefits [23], the support for engaging users in reflection is limited in current technology [22]. To address this gap I designed two conversational systems - Reflection Companion & Robota. With Reflection Companion I demonstrate an effective process for designing reflection dialogues involving workshop-based question generation informed by structured reflection models [14]. With Robota I demonstrate the use of past literature and contextual resources for informing work related reflection dialogues. Both agents personalize their interactions by involving user fitness tracker based activity graphs & behavior change goals, in case of Reflection Companion, as well as user journaled work tasks, in case of Robota. Robota also explored the voice and slack modalities for engaging users in reflection. Both designs contribute to informing the design of technology to support reflection and reusable processes to follow to achieve similar designs in different settings. I further evaluate both systems in deployment studies in personal physical activity and workspace productivity settings for 2 and 3 weeks respectively. I demonstrate that both systems were successful in engaging users in interaction and meaningful reflection leading to increased awareness, critical thinking, prompting new behaviors, and increasing motivation. I also demonstrate that the benefits were directly linked to the specific aspects of conversational design. Furthermore, I compare the trade-offs in conversational reflection design for private and semi-public space as well as with voice and chat modalities. Designers and practitioners in health & behavior change could use the proposed strategies to effectively engage users in reflection. Also designers of conversational systems can leverage aspects of the proposed designs to inform dialogue design with external data (e.g., activity graphs, work

tasks), and for leveraging domain-specific conversation topic based content diversification (e.g., sourcing reflection topics).

Chapter 5

CONVERSATIONAL DATA COLLECTION: DESIGNING HEALTH & SOCIAL NEEDS CONVERSATIONAL SURVEY

This chapter aims to address the challenge of data collection in health & behavior changes settings. In this chapter I apply conversational design to help engage users with self-reporting their social needs & health data at hospital emergency departments. This data is collected for use by providers to offer patients assistance with supporting their social & health needs. While data collection in personal informatics relates to collection of data about oneself for personal use, my work in this chapter supports collection of personal data for the potential use by a health provider and not necessarily directly by an individual. While this is a bit different than the classic personal informatics model from Li et al [142], I note that: 1) Personal informatics already goes beyond just keeping the data personal, with users explicitly sharing their data with 3rd parties such as family and friends [4, 46], health providers [47], or implicitly with tracker manufacturers. In all these cases the challenges of trust and communication are present. 2) Challenges of engagement with data collection are prevalent in both settings, especially when collection relies on user self-reporting. Given these, I believe the setting in this chapter aligns well with the challenges of the 'collection' stage of Li's personal informatics model and my findings are similarly applicable to such settings.

5.1 Background

Accessing patients' social needs is becoming a critical challenge at emergency departments (EDs) [157]. EDs are designed to attend to acute conditions (e.g., heart attacks, accidents), but increasingly, especially in public safety-net hospitals, are the first point of contact for vulnerable populations with long-term social needs (e.g., homelesness, poverty, hunger) [157].

Unprepared for these kinds of challenges, the EDs have to devote valuable resources (hospital beds, medical staff time) to understand the needs of such visitors and connect them to social services that can offer much better help.

5.1.1 Challenges of Collecting Data From Vulnerable Populations

Unfortunately engaging vulnerable populations in sharing their social needs is hard to accomplish with traditional surveys. Such populations are also often of low health and general literacy, are wary of sharing personal information in formal, impersonal surveys, and are more willing to engage with and respond to an interviewer rather than to fill-out a survey [32]. However, most EDs do not have extra staff to administer survey-based screeners, and without personnel administration (such as research assistant, nurse, etc.), response rates for both paper and electronic surveys are low [32]. This is compounded by the fact that only 12% of Americans have proficient health literacy and it makes the response rates especially low for low health literacy patients [136].

5.1.2 Technology-based Data Collection vs. Human Interviewing

With the growing interests in clinical screening, research has examined the use of technology based solutions to support the self-administering of surveys via online survey platforms, mobile apps, and electronic kiosks to maximize scalability and speed of data collection while reducing cost [96]. Despite these advantages of existing technology-based solutions, face-to-face is still better when it comes to engagement and response rates (in one study, 92.8% face-to-face response rate compared to 52.2% web-survey response rate) [104]. These differences, also in ED context [44], have been linked to the motivating impact of interpersonal interactions, but reproducing such effects via technology is still a challenge.

5.1.3 Potential of Conversational Approach

Chatbots offer multiple potential benefits for social needs screening. Chatbots are systems designed to engage with users through natural language, mimicking a human-to-human interaction [127]. Popular examples of chatbots include Apple's Siri, Google's Now, and Microsoft's Cortana. Extended to the context of social needs assessment chatbot can support self-administering of social needs screeners to minimize personnel cost. In contrast to current form-based surveys, a conversational approach would be more "chat" like, potentially offering a sense of familiarity similar to mobile text messaging [40]. By creating a sense of interacting with another person, the chatbots may also increase participation engagement [124]. Furthermore, offering text-to-speech audio output can also facilitate comprehension [96].

5.2 Design Approach

I designed and implemented a custom chatbot called HarborBot to test conversational approach to survey administration. HarborBot interacts with users through chat and voice. It communicates via chat messages, which it can also read outloud, as if it is speaking. Users interact with the system primarily through buttons (for structured responses) and text (for text-based questions). HarborBot is implemented as a webapp and participants interact with it on tablets.

5.2.1 Design Process

To create HarborBot I followed an iterative design process in which a team of 2 senior HCI researchers and 6 design students followed three general design phases: 1) Requirements gathering - based on feedback from 3 ED practitioners & literature [44], 2) Design exploration - prototyped various low-fidelity versions and gathered feedback via small scale usability tests, and 3) Refinement - most promising prototype was developed further and refined with positive elements from others.

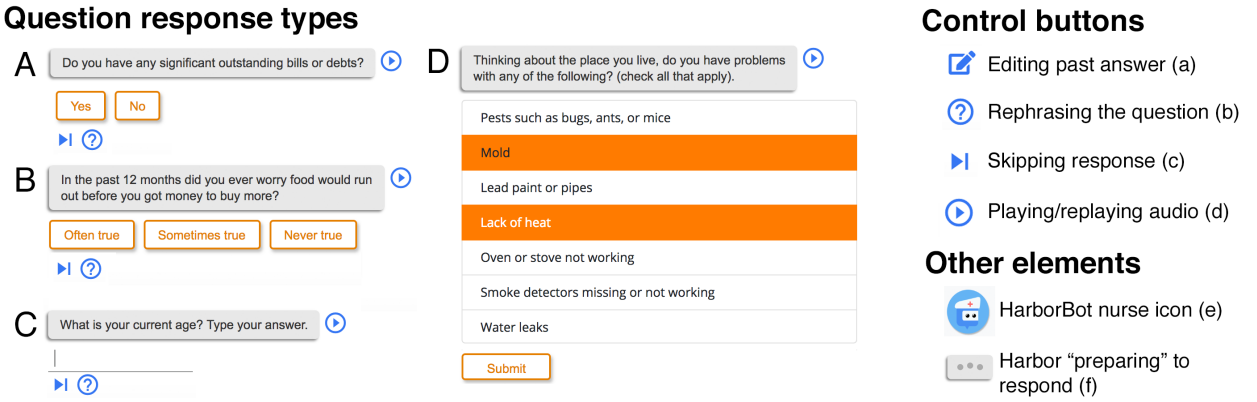

Figure 5.1: HarborBot GUI elements. On the left "Question response types" showing different types of responses users available. On the right "Control buttons" show the 4 controls associated with each question. "Other elements" show HarborBot icon and an ellipsis icon HarborBot used for mimicking writing by a person in chat interaction.

5.2.2 User Interface & Response Options

I used BotUI[1] - a Javascript framework to build conversational UIs. Users' messages are distinguished from the bot's by different colors. Animated ellipses are shown with a delay to denote that the bot is typing (Figure 5.1f). Interface supports different question types: *skip* - move to next utterance without responding (Figure 5.1c), *yes/no* (Figure 5.1A), *input* - free text response (Figure 5.1C), *options* (Figure 5.1B, or *many options* (Figure 5.1D).

5.2.3 Persona

I aimed for balance between serious and friendly tones to help users take the conversation seriously, but also provide comfort when answering personal questions. I avoided use of humor and sought to make HarborBot empathetic without sounding condescending. To accomplish this, HarborBot used occasional confirmatory phrases, such as: *"Okay, I'm getting a better idea of where you are at."*, *"Got it"*, and assurances, such as: *"The next questions*

[1]BotUI - https://botui.org/

are about your personal safety and may be tough to answer." The use of voice was important for understandability. HarborBot used a female voice taken from Microsoft's Bing Voices[2]. Users could adjust the volume of the voice or mute it entirely for privacy reasons or personal preference.

5.2.4 Dialogue-Based Interaction

Survey questions and user replies were presented as streams of messages in threaded conversation akin to chat messaging. Each question could be skipped (Figure 5.1c) and the conversation would continue. HarborBot supports rephrasing the question to offer its simplified version (fifth grade reading level) for low literacy individuals (Figure 5.1b). An edit button (Figure 5.1a) is present next to each past answer in case the user needs to change it. On top of these functional aspects, HarborBot would occasionally respond with conversational remarks. These utterances were essential to developing Harbor's personality, and engaging users in a conversation. Some of these interactions are dynamic based on a rule-based approach. For instance, if a user indicated they did not have a steady place to live, HarborBot would not ask the remaining housing questions. If the user response indicated a negative social situation, HarborBot would acknowledge it with a sympathetic affirmation, such as *"That must be stressful, I'm sorry to hear that."*

5.3 User Study

I conducted a within-subjects study with 30 participants (17 male, 10 female, 3 declined to answer, mean age: 39.63, SD=12.91) to compare the experience of answering a social needs survey using two different platforms: HarborBot (Chatbot) and a more traditional interface for taking surveys - Surveygizmo (Survey). I recruited participants with high (19 participants) and low (11 participants) health literacy at two study sites: 1) Seattle metropolitan area and 2) safety net hospital in Los Angeles (Harbor-UCLA). Chatbot interface was ex-

[2]https://docs.microsoft.com/en-us/azure/cognitive-services/speech/api-reference-rest/bingvoiceoutput

pected to be 1) more engaging, 2) more understandable, and 3) more comfortable to share information with, while 4) preserving response quality. I also expected these effects to be pronounced with low literacy users.

Procedure: Participants interacted with both survey interfaces using a tablet's web browser. After interacting with one interface, participants reported their perceptions and experience and then repeated the same procedure for the second interface. I randomized the order of interaction. After completing both, I conducted an interview. In both platforms users answered the social needs survey (LACHA) consisting of 36 questions related to demographics, financial situation, employment, education, housing, food, and utilities as well as questions related to physical safety, access to care, and legal needs. A number of questions can be considered sensitive, such as: *"Have you ever been pressured or forced to have sex?"*, *"Are you scared of being hurt by your house?"*, *"Did you skip medications in the last year to save money?"*

Measures: Participants evaluated interfaces on workload (NASA TLX survey [200]), engagement in the task (from O'Brian's engagement survey [180]), understandability of content, and willingness to share information. These measures have been commonly used in prior studies of chatbots [31]. Health literacy was measured using Rapid Estimate of Adult Health Literacy (REALM) [58] and Newest Vital Sign (NVS) health literacy scale [234]. During the interviews I asked about preferences for the two survey platforms, the specific features of the platforms, participants' comfort in sharing information in each platform, and perceptions of the personality of the chatbot.

Analysis: Analysis focused on descriptive statistics of user interactions, especially with Chatbot, and on comparison of answer equivalence for the two platforms. Differences in survey responses were assessed using paired t-tests and interactions between interface type and participant's health literacy levels were explored using linear mixed effects models.

The interviews took between 7 and 25 minutes (M=17.56, SD=9.21), conducted by three and analyzed by four of the authors. Each researcher wrote a detailed summary of interviews they had not conducted, including quotes. I then developed a codebook following a *top-down* and *bottom-up* approaches. Initial codes for the top-down pass were informed by the interview questions. I then refined the codes based on themes that emerged from the data in a bottom-up fashion. Each interview summary was coded by a researcher on the team (who had not conducted the interview, or written the summary). The coded interview summaries were used to identify themes. Three of the authors discussed the overall themes until consensus was reached. Researchers consulted with the audio and transcriptions of the interviews to ensure validity of the coding.

5.4 Results

5.4.1 Preferences

Low health literacy (LL) participants preferred using Chatbot over the Survey with 8 out of 11 expressing such preference. At the same time, 17 out of 19 high literacy (HL) participants preferred Survey. This difference was statistically significant (χ^2 (2, N=30)=12.5, p> .001).

5.4.2 Time to Completion

Participants had to respond to 36 questions in the social needs survey, but they could also skip answers. They spent significantly (t(27)=2.23, p> 0.05) more time answering questions via Chatbot (M=9 : 26 min; SD=3 : 14 min) than via Survey (M=6 : 48 min; SD=6 : 28 min). There was no significant difference between answering time (avg. of both interfaces) for LL (M=9 : 43 min, SD=3 : 23) and HL participants (M=7 : 36 min, SD=4 : 20). There was also no significant interaction between the interface and literacy level on time.

5.4.3 Equivalence of Responses

An important question is whether the two interfaces result in the same data quality. I explore two measures: *per-item response rates* and *data equivalence*. On average participants provided almost identical number of answers via the two interfaces: 32.93 (SD=3.48) questions answered with Chatbot and 33.00 (SD=2.95) with Survey. This suggests *comparable response rates*. In terms of data equivalence 87.0% (SD=11.6%) of the responses per user were the same across the two interface versions.

5.4.4 Reasons for Response Discrepancies

Skipping an answer in one interface, but not the other was the primary cause of answer discrepancy (48% of mismatches). There was, however, no significant difference between the two platforms in skipping behaviors. 25% of mismatches were a result of skipping a question in Chatbot only and another 23% due to the opposite. Furthermore, the order in which users encountered the interfaces had no significant impact on skip rates: 8.0% (SD=9.3%) when answering the survey the first time, and 7.8% (SD=8.2%) when answering the survey the second time. Hence the platforms are not different in this respect. One interesting finding from the explorations is that there seems to be an anchoring effect with users skipping more often when starting the study with Chatbot, for their responses to both platforms: Chatbot (M=9.8%, SD=29.7%) and Survey (M=9.8%, SD=27.2%) than when starting with Survey: Chatbot (M=5.2%, SD=22.3%) and Survey (M=4.9%, SD=21.6%). This is most likely due to the skip option being more explicit in the Chatbot and users wanting to be consistent in their answers.

Manual examination of the remaining mismatches revealed varied and non-systematic reasons for discrepancies such as: low equivalence only in the very first introductory question (53.3%), direct contradiction (e.g., user answered *"Yes"* in one interface and *"No"* in the other); similar, but not the exact same answers (e.g., answer: *"Yes, help finding work"* vs. *"Yes, help keeping work"*), ticking an additional option in a multi-choice answer (e.g.,

"Unemployed - looking for work" vs. *"Unemployed - looking for work, Disabled"*) and a possible misinterpretation of the question (e.g., when asked for income per month, user typed *"2000"* in one interface and *"24,000"* in the other).

5.4.5 Workload (NASA TLX)

Analysis of the NASA TLX survey responses revealed a difference in task load index (avg. of all items denoting workload, α=0.83) between Chatbot and Survey. Participants reported a higher workload when using Chatbot (M=2.460, SD=1.241), compared to Survey (M=2.167, SD=1.284; t(27)=−2.020, p=0.05). Given the scale from 1–lowest to 7–highest, this still represents a low perceived workload. There was also a main effect of literacy level: a higher perception of workload across both platforms by the LL participants (M=2.955, SD=1.335) than the HL ones (M=1.921, SD=0.948; t(27)=2.439, p> 0.05). The interaction effect was not significant.

5.4.6 Engagement, Understandability, and Comfort with Sharing

Analysis of the engagement index (average of O'Brian's engagement questions, α: 0.82), revealed a higher reported engagement for LL participants (M=3.920, SD=0.502) than HL ones (M=3.469, SD=0.402), (t(27)=2.672, p< 0.05). There was also a weakly significant interaction between interface and literacy with LL participants being more engaged with the Chatbot than HL ones, but less engaged with the Survey (Chatbot∗Low, β=0.485, SE=0.262, p=0.064). This represents a half a point increase on a 5-point likert scale for engagement. Trends in the same direction, but no significant differences were found for understandably and comfort with sharing information.

5.4.7 Interview Feedback

In this section, following mixed-methods approach, we complement and expand on the quantitative findings. Participants varied not only in their preferences for Chatbot or Survey, but

also in the particular aspects they liked about each, as well as in which design aspects were instrumental in creating particular perceptions and experiences. Participants valued the engaging conversational aspects of the Chatbot. Especially LL participants found the conversational interface more caring in the context of a sensitive topic. In contrast, HL valued the efficiency of the SurveyGizmo interface and felt slowed down by the Chatbot. Some participants found the Chatbot more robotic, disingenuous or pushy at times, but these seem to result from the particular way in which HarborBot implemented conversation.

Strength of Proposed Conversational Data Collection Approach

Engaging: Most participants, regardless of literacy level, found the chat more engaging than the Survey. Participants felt like they were having a conversation with a person when using the Chatbot. More than half the participants attributed such perception to the use of voice: *"she was reading the questions and I can answer it ... seemed like a conversation ... like someone was talking to me and it gave me the opportunity to answer back and then they answered back"* (H59). Other participants felt the *ellipses* made it feel like having a chat with someone (H76, L77), and even referred to the Chatbot as *"she"* (8 participants). Some participants valued that the Chatbot felt like a person: *"I liked... how it talked to you, reads you the questions ... it spoke directly at me"* (L60), *"I thought it was someone asking me those questions"* (L72).

Aside from the voice and ellipses, the conversational utterances also contributed to the perception of interacting with a person (L75, L58, L72, H32, L60, L36, H41, H59). One participant found them motivating: *Saying 'you got it.' It's giving you motivation ... nice to hear that once in a while"* (H73). Another felt like the conversation was adapting to the answers to be more relevant: *"seem like they tried to give you a little positiveness based on your answer"* (H59).

Caring: Participants perceived the Chatbot as caring, particularly in the LL group. These participants had a generally positive attitude towards the social needs survey questions (L51,

L55, L58, H73) and this topic resonated with their personal experiences *"It felt like it was telling me about my life. That was really amazing, like woow"* (L71). Therefore, some of the perceptions of the Chatbot might have been accentuated by the positive perception towards the survey topic. Many participants described the personality of the Chatbot using terms such as: caring, kind, patient, helpful, calm, familiar, or concerned (H35, H41, L52, L55, L57, L61, L77). Participant also reported the voice of the Chatbot was aligned with this caring personality: it was soothing (H57), had cadence (H32), helped a nervous participant feel more comfortable (L55) and was *"nice and sweet made me feel relaxed"* (L77).

The Chatbot was designed to provide supportive utterances in response to some of the participants answers. Many participants liked these utterances (L60, L36, H32, H41, H58, H59). One participant though the utterances made him feel *"comfortable to answer the questions"* (L61), and that they provided a positive reinforcement to keep on answering (H59). Participants perceived Chatbot utterances such as *"I am sorry to hear this"* as the Chatbot *"trying to be understanding"* (H59). Some found these utterances to be very applicable to the conversation context. For example L61 considered the Chatbot response: *"That must be stressful"* to be a reaction to the information she shared: *"she probably said that because of my financial situation"* (L61), which she felt would be calming for people *"to not be stressed, I would think it would be helpful"* (L61). Other participants felt the supportive utterances gave them confidence: *"nice lady giving me confidence ... with good tone of voice"* (L75).

Understandable: Several LL participants (5 of 11) reported having trouble with reading and understanding the written questions in the Survey. They liked using the Chatbot because it facilitated their understanding, which they attributed to the audio feature: *"When I hear it I have a better understanding of the question"* (L61) or that *"just hearing it I could ... relate better to the question"* (L53). Some participants reported using the feature that replayed audio, to better understand a particular item (L51, L58, L61, L73). This was especially useful when they missed some words or did not fully comprehend some of the contents at

first: *"I didn't get it at first, so I wanted to go back and listen to it again before answering"* (L58). Several also mentioned that they would have liked it if the answers were spoken via audio as well, to make them more understandable (L61, L54).

Accessible: Some participants had particular needs that the Chatbot was able to satisfy much better than the Survey. One participant who reported vision problems, preferred having questions read to them: *"If it is too small I can't see it so I prefer to have the questions read to me anyways"* (H73). Another participant reported feeling very comfortable with the Chatbot because she was regularly experiencing panic attacks and considered ED stressful: *"I was thinking I was texting somebody ... that made me forget where I was at ... it was like texting my sister my mom and waiting for them to respond back. And that made me feel patient"* (L77). In contrast, she found it particularly difficult to take the Survey: *"by myself... it felt awkward and alone"* (L77).

Weaknesses of Proposed Conversational Data Collection Approach

Inefficient: HL participants cared about efficiency, primarily reflected in the speed of completing the survey. The majority of HL participants (17 out of 19) preferred to use the Survey because of that. Several mentioned that the traditional interface enabled them to be faster than the Chatbot (H21, H22, H24, H59), or to go at their own pace (H36). Participants attributed being slowed down to various conversational features of the Chatbot. Some felt the Chatbot was slower because they needed to wait for the ellipses before a new question would appear (H35). They were also able to read faster than the questions were read by the bot: *"when she was talking at me. I felt like I was going at a slower pace"* (H23). Also not having to engage with additional conversational utterances was seen as more efficient (H35, H56). The audio feature was perceived as interfering with reading and thinking (H23, H40, H70, H21, L71). One LL participant preferred the Survey because they could concentrate more: *"to read is better ... Because that way I could like concentrate more and think about more and you know ... I could read my letters more and makes it better for me."* (L58).

Pushy: Somewhat surprisingly, a few participants perceived Chatbot as being pushy, based on the tone and the speed at which questions were asked. Some participants felt the questions asked were very direct (H57, L72, H52). L72 felt like he was answering questions to a teacher, and had to provide correct answers. H57 and L72 thought there could be more utterances to help prepare the survey taker for some very sensitive questions in the survey. H57 also felt that some of the questions were trying to repeatedly get information that he had already declined to provide: *"if I say none of the above ... don't be pushy"* (H57). Others also felt rushed in providing the answers to the Chatbot. For example,the use of ellipses, and the short delay between its messages made it feel like the Chatbot was moving faster than the participants were comfortable with (H23, H63). Participant H63 felt like the questions kept coming and he had no control over when they would be read.

Robotic and Disingenuous Voice: Some participants, primarily in the HL group, perceived the Chatbot as being robotic. Some participants found the voice not sounding natural (H21, H22, H23, L58, H59, H63, H70, H76), for example sounding *"truncated .. monotone... seemed pretty artificial to me."* (H70). Some perceived the Chatbot as disingenuous when the utterances did not meet their intended purpose (H63, H40, H23, H52): *"I feel like they were trying [to make] the software to feel sympathetic, or empathetic, that was weird"* (H63). Another participant perceived utterances as defaults: *"it felt like defaults rather than someone 'feeling for you"'* (H40). The perception of artifical responses led another participant to perceive the Chatbot as fake, and was reminded of customer support: *"kind of just programmed, recorded in, to appear to be more personal... hell there's nobody there somewhat disingenuous ... It reminded of ... dealing with the phone company"* (H70).

Inconclusive Impact on Willingness to Disclose Information: Most participants, regardless of health literacy level, reported being comfortable sharing information asked by the survey questions. However, the human-like interactions of Chatbot did affect some participants' willingness to disclose information, although participants reported effects in

both directions. For some, if they thought they were interacting with a person, they felt more reluctant to share sensitive information, or tell the truth: *"I might be more honest if I'm reading [the question] ... if someone else ask me about them, I might lie"* (L72). Another participant showed concern about the identity of the potential conversational partner: *"it was a robot, I didn't mind, but I think if it was a human being I would mind... and you really don't know who's on the other end"* (H40). In contrast, some participants were more willing to disclose because of the human-like interactions. *"If it says 'I would like to more about you'. It gives me the confidence to open up, because each question that follow sounds so interesting and it gives me the opportunity to interact with the person on the other side ... it wettens my appetite to give out more information"* (L75).

5.5 Discussion

In this paper, I proposed the use of a chatbot (HarborBot) for social needs screening at emergency departments and compared it to a traditional survey tool (SurveyGizmo). Based on interviews, interaction logs, and survey responses we demonstrate that the conversational approach is perceived as more engaging by all the participants, and further as more caring, understandable, and accessible among the low health literacy (LL) ones. Importantly, I also demonstrate that the conversational approach results in similar response rates and 87% equivalence in the collected data. At the same time, I found the conversational approach to be more time consuming (in line with reports from prior work [164]) and prone to be perceived as somewhat pushy, robotic, and disingenuous which was, however, mostly the perception of participants with high health literacy (HL).

5.5.1 Positive Design Aspects

Numerous strengths of the conversational approach for LL population can be linked to conversational features. First, various features of the chatbot facilitate understanding. The audio output is especially valuable for participants who are less proficient readers. Second, the ability to ask the bot to rephrase the question offered a way to ask for clarification that

is currently not a feature in online survey platforms. Third, chatbots can create a sense of interacting with a human. The utterances can make the survey takers feel cared for and engaged. Such positive interactions made some participants feel relaxed and even motivated to answer more questions.

5.5.2 Challenging Design Aspects

Conversational features felt pushy for some, especially HL participants. Such perception was linked to the tone of the questions and to the speed of the interaction. In terms of tone it is possible that the literal use of the wording of the survey questions was not the most appropriate for creating a conversational feel. In terms of speed of interaction, the use of voice might be a contributing factor. As reported in prior work agent asking questions via voice can create a perception of response urgency[130]. This could be improved by adding assurances like *"please take your time."*, manipulating intonation, or making it more explicit that the ellipses represent someone is typing (rather than the system is waiting for a response). The second reason for pushy feel could be related to the fixed speed of conversation. Human-human conversation involves not only exchanging information, but also coordinating various aspects of the exchange, e.g., its speed [31]. If a participant needs more time to think, a real person, would pick it up from verbal and non-verbal cues and adjust the speed. Our HarborBot is currently incapable of making such adjustments. Such fixed speed may feel too fast or too slow for some users.

HarbotBot felt *"caring"* for LL and *"robotic"* for HL participants. This might be related to the different expectations and tolerance levels for voice quality and may be improved with use of a better quality text-to-speech service (technical challenge), human pre-recorded audio clips (which comes with limitations in flexibility), or modifications of intonation and prosody using approaches such as Speech Synthesis Markup[3]. Another way may be to generate more personalized and diverse utterances[131].

---

[3]https://www.w3.org/TR/speech-synthesis11/

5.5.3 Future Design Directions

Given the division of preferences for chatbot/survey between the HL and LL groups, one possibility for a real-world use could be to have two versions of the tool and either intelligently assign or have patients pick the version they would prefer to engage with. While, long waits in healthcare setting make it less of a problem, a number of design opportunities can still be explored to make the chatbot interactions more efficient, such as simplifying the script, or providing user control over time between messages. While we focused on examining the effects of the conversational approach for a LL population, our findings suggest a potential for accessibility-focused uses of the chatbot. Participants who were hard of seeing mentioned they appreciated the audio output. Further, one participant with anxiety attacks appreciated the human-like interactions, which made them feel like chatting with a loved one, at home.

Finally, it is not clear based on our results, how the conversational approach affects people's comfort in responding to questions, and any potential desirability biases. Prior work suggests that the self-administered screeners would reduce social desirability bias, and limit the under-response in sensitive issues[91]. This is because people will not feel like someone is monitoring or judging them. We thought the Chatbot may strike a happy medium between being perceived as human-like to enhance engagement, while not being perceived as a person for people to feel uncomfortable with disclosures. It is not clear if we were able to achieve that balance. Some participants who thought the Chatbot was human-like did not mind sharing and commented that it was more motivating, while others that thought the Chatbot was human-like were concerned with sharing. It is possible that the very initial greeting from the bot sets the tone for the rest of the interaction[118]. This requires additional research.

5.6 Summary of Contribution

In this chapter I demonstrated the use a conversational approach to engage users in sharing information about their health & social needs in emergency department setting. I specifically designed a mixed-modality (voice and text) chatbot called HarborBot for administering a

social needs screening survey in a conversational manner. HarborBot included specific conversational language features to boost user engagement: social phrases, empathetic reactions, as well as conversational style and etiquette. It also supported chat GUI specific human-likeness features, such as a response delay and 'ellipses' indicator. On top of that it included features specific for supporting low health literacy populations, such as voice readout, and question rephrasing. I evaluated HarborBot with 30 emergency department visitors (11 low literacy) and identified a number of positive design aspects of chat-based approach: 1) improved engagement due to increased perception of care, calm personality, and feeling of interacting with a human. 2) improved understandability due to audio support and question rephrasing features (especially valuable for LL users). I also identified, that the high health literacy users mostly preferred traditional survey as it was more efficient. My work advances the understanding of the role conversational agents can play in supporting collecting data from users at scale. It particularly offers insight into designing engaging and understandable conversational interactions for low literacy, vulnerable populations and on sensitive topics. Further it highlight the impact such interfaces can have on patient-provider information sharing. Designers of conversational interfaces can use my findings to inform impact of different features on low and high health literacy populations. Health professionals and social workers can also leverage my work to improve conversational agent adoption in hospital settings to reduce care costs.

Chapter 6

AUTOMATING THE DESIGN OF ENGAGING CONVERSATIONAL DATA COLLECTION

In the previous chapter I have demonstrated the benefits of conversational administration of a survey on social needs in the emergency department context for improving user engagement. In earlier chapters I have explored some intrinsic conversational design features contributing to improved engagement, such as contents & language diversification, contextual tailoring, socialization & empathy as well as general human-likeness principles. At the same time across all the chapters I have demonstrated the substantial effort and work required to design enticing conversational experiences, which has also been identified as a challenge in prior work [97]. In this chapter I explore which and how common design components can be reused and applied automatically to aid the design of engaging conversational experiences. I explore this design automation support focusing specifically on my work from the last chapter by attempting to automate the adaptation of survey-based data collection to a more engaging conversational form.

Engaging users in sharing information about their health, behavior, preferences as well as other aspects is important for successful behavior change [165], health interventions [50], and also in a broad range of other domains [2]. As demonstrated in the previous Chapter 5 and in other work [124, 239], conversational survey administration can increase user engagement resulting in numerous benefits. Yet beyond hand-crafted examples of conversions of specific surveys, there is no clear systematic way of adapting a survey to the conversational form. To address these challenges and to systematize the knowledge about conversational design developed in previous chapters I propose a systematic automated process for adapting any form-based survey to chat-based conversational form following 4 augmentation tasks

informed by engaging conversation design principles distilled from prior chapters.

6.1 Background

The value and means of performing an automated adaptation of surveys to conversational form are based on several areas of work. Prior work demonstrated the specific benefits of administering a survey in a conversational format [124, 129, 239]. These works also offered examples of how conversational adaptation may look like, even if only for a particular survey and domain. The linguistic literature offers insights into what elements a conversation is composed of and hence can inform the linguistic adaptations required. Finally, several applied approaches help inform the technical solutions that can aid in making an automated conversion feasible.

6.1.1 Engagement Benefits of Conversational Survey Administration

Recent work, including my own, has shown that survey administration in a more conversational form such as via chat has the potential to increase user engagement resulting in higher quality responses and lower drop-out rates [124, 129, 239]. These benefits have often been attributed to the chatbots' ability to naturally deliver human-like interactions [239].

Kim et al. investigated in an experimental study whether it is the chat-like GUI or the specific use of language that provides the conversational administration benefits. They found that for conversational surveys to be effective, having GUI (chat-like interactions) alone is not sufficient. The language needs to be in a conversational style as well [124]. In Chapter 5 I have shown the value of conversational administration of surveys for user engagement particularly for questionnaires involving sensitive topics (e.g., about sexual abuse, violence, or financial situation) and applied in sensitive settings (e.g., hospitals) [129]. Xiao et al. emphasized the benefits of natural and familiar ways in which conversational interaction allows users to express themselves [239]. Personified and anthropometric features have also been linked to increased user attention and trust [221, 77]. Even framing the questions as more personalized conversational messages has been shown to have the potential to improve

user engagement and response quality [131, 43].

Yet beyond hand-crafted examples of conversational adaptations of specific surveys, there is no well defined and systematic way of adapting any survey to conversational form. This increases the barrier to entry for survey administrators without a design background to make their surveys more conversational and engaging for their audiences. This therefore leads to my first research question:

- **RQ1:** How to support the systematic conversion of any survey to a more engaging conversational form?

6.1.2 Linguistic Elements of Engaging Conversational Design

Linguistic theories provide several language elements that could inform augmentation. Proposed by Austin [211] and further developed by Searle [212] speech acts theory differentiates between five types of phrases: representatives (e.g., claiming, reporting), directives (e.g.., advice, request), and expressives (e.g., promise, threat). Empirical resources such as [86] offer concrete examples of some common phrases for speech act subcategories.

In the conversational agent domain, past work emphasizes the importance of various elements, such as proper agent introduction [197], ending of the exchange, and a certain level conversational etiquette [112, 197]. Engaging relational agents incorporated social behavior such as social dialog, empathy or expressions of liking [30]. A recent review of social cues identified several elements used in prior work, such as thanking, praise, and many others [78]. Conversational survey administration work focused on a subset of these, such as response feedback [124, 239, 129], social acknowledgments [239], handling conversational flow and transitions [124, 239, 129], response prompting or probing [239] as well as survey question rephrasing [124]. Aside from concrete phrases, prior work also indicates that engaging conversation relies on some overarching language properties, such as 'conversational style' [124], avoidance of repetitions [30, 131], lexical diversity [78], degree of human-likeness [38, 39, 15], formality [78], and consistency [112].

Past work and linguistic theory provides common language elements and properties in broader conversation as well as specifically used for designing engaging conversational experiences. Not all of these are, however, easily and meaningfully applicable to survey administration context. Furthermore, while broader categories are defined, concrete phrases directly usable for survey augmentation are limited. Given these indication, my two subsequent research questions focuses on their impact on survey respondents and the ability to support these elements well:

- **RQ2:** What is the impact of a converted survey on user engagement and perceptions of concrete conversational design elements?

- **RQ3:** Which aspects of conversion are handled well, and which are still problematic for our approach to automation?

6.2 Making Survey Conversational - Design & Automation

I consider the whole text of a survey to be structured as a sequence of survey items. A survey item can be a question, which requires an answer, or it can be a non-answerable item, such as an instruction, section heading, etc. Each item is a piece of text. Question items can be answerable via selection of one or more of answer options from an associated set or via free-text input. The conversational adaptation involves: 1) the use of chat-like GUI and 2) a number of linguistic adaptations to the survey text and structure. Prior work indicated that in survey context chat-like GUI alone without adjustment to the survey language may not produce expected engagement benefits [124].

Given these design indication, linguistic adaptations in this work focus on two types of changes: 1) additions of "conversational" utterances in-between the original survey items (e.g., addition of acknowledgments, reactions, greetings) and 2) limited modifications to the existing survey question text to fit conversational administration context (e.g., phrasing utterances as questions and modifying the language to be consistently "conversational"). At the same time, survey context applies certain constraints, most importantly the need for

preserving the text of the original survey questions. This is to avoid changes in meaning and preserve survey validity (in case of validated instruments) [108]. The conversational augmentation of a survey in this work is composed of 4 major tasks:

- Adding introduction & closing for the conversational interaction
- Adding reactions to user answers in question context
- Adding conversation progress communication
- Modifying survey questions to fit conversational style

6.2.1 General Design Principles

Here I describe a number of guiding principles for design & automation of the conversational survey augmentation as informed by prior work on conversational engagement and the requirements of the survey context.

Avoiding repetitiveness I have shown the importance of diversification of language in the design of engaging conversational agents. In Chapter 3, where I described the use of mobile chat for exercise promotion, the repetition caused measurable drop in user engagement and compliance with exercise suggestions [131]. Similarly in conversational reflection work with *Reflection Companion* in chapter 4, as well as in conversational social needs screening with HarborBot, described in Chapter 5, repetition was often mentioned in qualitative evaluation as a cause of 'artificial feel' and a potential factor negatively affecting engagement. Prior work on relational agents have also reported that repetition can be severely detrimental to user engagement, even leading to drop-out [30]. The level of lexical diversity is also listed as an important social cue [77]

The augmentation process tries to maximise the diversity of phrases used for augmenting the survey. This is accomplished by: 1) providing several variants for each augmentation phrase, 2) tracking frequency of use of particular phrase variants to minimize repetition.

Minimizing changes to the original text Non-research surveys can be designed for a specific product or narrow one-time purpose and do not necessarily rely on a highly precise language to collect valuable information. However, many surveys used in academic research are validated instruments meant to measure specific latent variables predicted by some theoretical foundations. In such cases the exact phrasing of the questions has been carefully designed and tested to ensure validity and consistency [108]. Major changes in the language could invalidate internal consistency measures in such surveys.

I try to minimize the changes to the survey phrasing by: 1) applying only minimal changes that would render the survey item applicable for use in conversation (making it 3rd person & framed as a question), 2) prioritize additions (e.g., prefix "Could you tell me"), rather the deletions of the original question contents.

Using empathy only in appropriate context Prior work used varying degrees of emotional expressiveness. Kim et al. employed an expressive casual style, where the chat communicates enthusiasm (e.g. *"Way to go!"*, *"Let's go to the next step!"*) and politeness (*"Please go on to the next section"*) [124]. Survey questions in that work are related to demographics and product feedback. Xiao et al. use much more personal and emotionally expressive phrasing (e.g., *"I am very impressed by what you do"*, *"Thanks, I'm glad you are happy with me"*) in a free text survey asking for feedback on a game trailer. My own work on social needs survey indicates that neutral reactions in the context of sensitive questions can lead to the perception of chat as being 'robotic', 'fake' and reminiscent of customer support lowering engagement [129]. Relational agents successfully employed expressions of empathy and liking behavior for driving user engagement [30]. Yet, there are also indications that overly expressive bot can feel disingenuous [24] and lead to heightened expectations [156].

The automated conversational augmentation tries to maintain neutral style in most of the added utterances. For a survey to remain unbiased, the chat should not try to be to positive or negative. The expressions of empathy are reserved for the context of questions framed in a sensitive manner (e.g., *"Do you feel threatened by violence?"*), while the neutral

acknowledgements are used as reactions for questions on neutral topics and framed in a neutral way (e.g., *"What is your age?"*).

Audience-sensitive augmentation Study of HR chatbot use in a company setting found differences in the appreciation of socialization aspects of a chatbot among users [147]. In Chapter 5 I have shown that high-literacy populations tend to prioritize efficiency and perceive additional social phrases as an unnecessarily lengthening the interaction [129]. In a counseling setting users may prefer neutral interaction to avoid judgmental tone [97]. Language style used for survey augmentation can vary from polite formal [129] to expressive [239] to casual [124]. Therefore the extent of social chat, empathy, and the style of language may depend on the intended population and context of use for the conversational survey.

The amount of social chat (i.e., how many reactions, progress phrases should be used), the tone of the interaction (e.g., whether a neutral or empathetic tone should be employed), and augmentation style (e.g., formal, informal) needs to be adaptable to the audience or even an individual. This is accomplished through: 1) parametrized application of all the augmentation tasks (e.g., the frequency of reactions or progress phrases) and 2) use of a separate augmentation phrase repository (e.g., polite repository can rephrase *"What is your age?"* to *"Please tell me what your age is?"*, style used in [129]; while a casual repository can rephrase it with *"Plz, tell me what is your age?"*, style used in [124]).

6.2.2 Building a Repository of Augmentation Phrases

I defined adaptation of survey to conversational form as composed of 4 tasks (i.e., adding introduction & closing, adding reactions, adding progress communication, and question rephrasing) and also provided several design principles. In order to support these tasks, I need to create a repository of phrases constructed in advance and informed by speech act theory [212, 86] and prior work [78, 239]. I can pick phrases from this repository as needed and inject them between the survey items. Conversational question rephrasing can be accomplished in a similar fashion, by appending conversational prefixes to survey items

to turn them into chat questions. The selection of the best phrases to pick from the repository for a particular survey position can be determined dynamically based on local context (e.g., question and user answer to decide on the reaction). Such retrieval based approaches have already been successfully used in conversational context [244]. Augmentations such as introduction & closing as well as progress communication can be largely accomplished using simple hand-crafted rules, while addition of appropriate empathetic reactions as well as questions rephrasing require data-driven ML components. In both cases, the augmentations are retrieved from a prewritten repository. Here I describe how this repository is constructed and what elements it contains.

Augmentation Elements

As discussed in related work, various phrases and language adaptations have been used in general conversational agents and in conversational survey administration specifically. Here I focus on a subset of phrases to support via a repository. Prior work indicated the importance of a conversational agent being able to properly initiate and end the conversation [112, 97] and also to follow proper conversational etiquette [112, 197]. Several works indicated the importance of chatbot properly communicating its purpose and capabilities [112, 127]. Hence the repository needs to contain Introduction & Closing phrases. Furthermore, the need and expectation of response feedback and acknowledgments [239, 124, 129] as well as the demonstrated value of empathy or expressions of liking on user engagement in relational agent design [28] dictates that the repository needs to contain Acknowledgements & Empathetic Reaction phrases. Prior work has also shown proper 'conversational style' is needed for engaging conversational survey administration [124] and is important for communicating engaging social cues [78]. The repository supports this goal by providing Question Rephrasing prefixes. Specific to the survey task the repository also contains Progress communication Phrases. These phrases have dual purpose, the usability purpose is to communicate task progress (survey completion is ultimately a task), the engagement purpose is to provide a sense of accomplishment, acknowledge user effort [239] and thank the user for contribution

[78]. Finally to further reduce the repetitiveness [131] and monotony of the exchange mimicking phrases used in conversational survey administration work [239, 129, 124], the repository contains phrases for Topic continuation & topic switching.

In summary the augmentation repository needs to contain phrases for: 1) Introduction & Closing of conversation; 2) Acknowledgments & Empathetic Reaction phrases to user answers 3) Progress communication phrases, 4) Topic continuation & topic switching phrases, and 5) Phrases supporting conversational reformulation of survey items, in the proposed approach - Question Rephrasing prefix phrases.

Generation of Repository Phrases

In order to populate the repository with concrete examples, I extracted phrases from prior work on manual conversational survey adaptation [239, 124, 129], included phrases from linguistic repositories such as CARLA [86], and further, especially for empathy expressions, from sources such as motivational interviewing [198], counseling, and more casual interview resources. There were two main aspects that had to be addressed in the process of adapting the phrases from prior work: 1) consistency of language style, 2) generalizability of phrases to different survey contexts. The first point required rewriting the phrases taken from prior work such that they would maintain a consistent style. Xiao et al. used very expressive style, Kim et al. used casual, teenage-like style, while in my HarborBot work for social needs screening I used professional and polite style. The second point required removing or replacing any survey specific words used in the phrases to ensure they are applicable in various survey contexts. This common generation process has been used for all the repository phrases except for conversational prefix phrase generation, where a more empirical approach was used.

While the conversational prefix phrases were also inspired by the examples from prior work on conversational survey administration [239, 124, 129], the generation process relied more on an empirical and iterative process. For a set of survey items, I would try to find a prefix phrase that could change the item into a question form in common polite conversational

style. In case the survey item was not formulated as a question, e.g., *"Type in your gross income"*, I would try to create a prefix that would turn it into a question, e.g., *"Could you please..."* In case an item was already in a question form, but phrased in a too direct language (e.g., *"Are you married?"*, the prefix would add consistency of polite style as well as diversification of phrasing, such as *"Please tell me whether..."*. Given a new question I would try to match one of the existing prefixes and if none would much, a new prefix would be created. Prefixes that seemed interchangeable for the same question context (e.g., *"Can you tell me whether..."*, *"Please indicate if"*) would form a prefix group. Aside from prefix phrases themselves a prefix group would also include empirically generated replacement rules for the original survey item, e.g., *"are you"* → *"you are"* as well as *"I"* → *"you"*). The process was designed to make the question conversational by use of additions rather than deletions of the original survey item text.

For each category I provide several phrase variants to help avoid repetitiveness (*'avoiding repetitiveness'* design guideline) and also several categories of augmentations to match local context either from perspective of empathy (*'contextual use of empathy'* guideline) or best fitting minimal grammatical augmentations (*'minimal changes to the original text'* guideline).

Augmentation Phrases Repository

The repository contains a total of 118 different phrases distributed among different aspects of the four augmentation tasks.

Introduction & closing phrases: The repository contains 13 different templates for chatbot introduction such as *"Hi, my name is name. I would like to talk to you about topic."*, *"Hi, I am name. Let's talk for a moment..."* and 6 different templates for conversation closing, such as *"We've completed everything! Thanks a lot!"*, *"We are done! Thanks."* for closings. These templates represent a coherent polite and professional tone to ensure consistency important in conversational interactions [172]. The introductions also contain

slots with a chatbot name and survey topic to be instantiated for a particular survey.

Empathetic reaction phrases: The repository contains three classes of empathetic reactions: *'Neutral acknowledgment'* , *'Expression of satisfaction'*, and *'Expression of compassion'*. Neutral acknowledgments play a role of non emotionally expressive feedback to the user that the chatbot is "listening" and "receiving" user input. These reactions are meant for context where emotional expressions would not make sense or could lead to judgmental tone. The repository contains 7 different phrases for this class such as: *"Thanks for sharing"*, *"I took a note of that"*, *"Okay, I'm getting a better idea of your answers"*. Expressions of satisfaction are meant to communicate positive emotional valence, encourage the user, and share in the user's positive emotion in an appropriate context. This class contains 10 different phrases such: *"I am glad to hear that"*, *"That's good"*, *"That's really great!"*. Similarly Expressions of compassion are meant to express chatbot's concern and empathize with the user, especially in the context in which the user might be disappointed or otherwise disconcerted. Use of such emotional reactions is meant to make the chat more human-like and natural drawing from work on relational agents. This class contains 6 different phrases such as: *"I am sorry to hear that"*, *"That sounds stressful"*, *"That's hard to hear"*.

Question rephrasing prefixes: The repository contains 6 classes of prefixes used for prepending to the survey items to turn them into questions, make them more conversational, provide diversification of phrasing, and also consistency of tone. There are a total of 37 different prefix phrases among the 6 prefix classes such as: *"Can you tell me"*, *"Would you"*, *"Have you experienced"*, *"Can you share if"*, *"Could you say that"*. These also contain replacement rules meant to rephrase the remainder of the survey item text into 3rd person question form, such replacement rules are e.g., *"i"* → *"you"*, *"am"* → *"are"*, *"are you"* → *"you are"*. Appendix B presents the classes along with the example survey items.

Progress communication phrases: The repository contains 12 progress communication phrase templates, which contain slots for current and total survey questions or a progress percentage, such as *"We are currently at question d out of n."* or *"We are done with percent% of our questions."* There are also distinct progress phrases for use in the middle of the survey: *"We are now in the middle of the survey"*, *"We're halfway there, still l questions to go"* as well as close to the end of it: *"We are almost done, thanks for your patience"*, *"We are almost at the end, thank you for staying that long"*.

Topic continuation & topic switching phrases: The repository contains 12 topic continuation and topi switching phrases. These include the generic phrases without specific topic slots, such as *"Let me ask you some more questions..."*, *"Just few other things I wanted to ask you about..."* as well as templates with topic information, such as: *"Let's move on to questions about {section_topic}"*.

It is worth noticing that this repository represents a particular consistent augmentation style that is meant to be polite and professional. It is possible and quite easy with a repository-based approach, to create survey augmentation phrases that represent e.g., informal, teenage-like style such as used in [124].

6.2.3 Design of Augmentation Tasks

Here I discuss the design and automation details of each of the 4 conversational survey augmentation tasks. These tasks rely on dynamically retrieving the most appropriate phrases given the survey context from the repository described earlier. As a result of the automated application of these tasks the form-based survey is adapted to a conversational form.

Task 1: Chatbot introduction & conversation closing

The goal of this task is to augment the survey with introduction and closing phrases selected from the repository. The position of the phrases in the resulting dialogue is fixed as the first and last utterances of the conversationally adapted survey. As described earlier, introduction

phrases are designed as templates with a chat name and the domain of the survey as slots to be instantiated for a specific survey. Chat name is simply determined from the domain, by appending Bot to the domain name. The closing utterance indicates to the users that the interaction is complete and politely thanks the user for their involvement. Table 1 contains example instantiations for specific surveys.

The chatbot name itself can communicate its purpose and capabilities and possibly set the tone for the exchange [11]. Academic and commercial applications used varying approaches to naming their conversational agents. Social and relational chatbots commonly use human-sounding names such as Eliza, Alice, Mitsuku, Xiaolce, or arguably more modern-sounding Tey or Zo. Various digital assistants, whose main task is to provide functional help, while maintaining a limited degree of social interaction tend to be given less human-sounding abstract names such as Siri, Cortana, Swelly, WoeBot, Tido. On the other hand, several non-social service bots derive their names from the companies they represent (e.g., eBay, Duolingo, Sephora), or function they perform (e.g., PizzaBot). Overpromising on agent capabilities can potentially lead to user disappointment and decreased engagement [156]. Given the narrow purpose of survey administration, which is just to collect answers to survey questions rather than engage in free-form social chat, I chose a naming scheme that corresponds to the survey domain, such as SleepBot, SocialNeedsBot or FinanceBot (Table 6.1).

Task 2: Survey question augmentation

The goal of survey question augmentation is to: 1) rephrase the original survey question to make it amenable for use in chat context, 2) preserve consistency of utterance style across conversation, and 3) introduce variation to the phrasing, that would avoid repetitiveness that could decrease user engagement. At the same time I want to make sure that the original text of the survey questions is preserved as much as possible.

To transform the originally phrased survey questions into form amenable for use in conversational context I perform several steps. First, I classify the survey question text into 6 phrasing categories (Table 6.2) derived empirically as described in the repository build-

Type	Survey	Instantiation
Opening	Theory of Planned Behavior Survey for Physical Activity	"Hi my name is [HealthBot]. I'd like to ask you a few questions about [health]..."
Opening	Reflection	"Hi I am [ReflectionBot]. Let's talk for a moment..."
Closing	Sleep Survey	"We're done! Thanks for talking to me about your [sleep]."
Closing	PANAS	"We've completed everything! Thanks for your patience"

Table 6.1: Introduction & Closing template examples and their instantiations for specific surveys. Phrases in-between square brackets are survey specific slots that are filled-in dynamically.

ing section. Based on the detected phrasing class, a concrete prefix text in that class is probabilistically selected from several different prefixes so as to minimize repetition (i.e., not reusing the same prefix in consecutive utterances). Selected prefix is prepended to the question text. Each phrasing category may include sentence modification rules to ensure the original question text is correctly phrased in 3rd person form (e.g., changing *"are you"* → *"you are"* or *"I"* → *"you"*). In the final step, to further lower the sense of repetition and emphasize the natural progression of the conversation an additional phrase such as *"Moving on"*, *"Next"* is prepended to some of the utterances with a given probability. Several examples of the original survey items and the subsequent conversational phrasing resulting from the augmentation process are presented in Table 6.2.

Task 3: Reactions to user answers

The goal of this task is to match the most appropriate chat reaction to the used answer in the question context. There are three reaction classes in the repository used for this task, each containing several concrete text phrases: *Neutral acknowledgments* (e.g., *"Thanks for sharing"*, *"Got it"*), *Expression of empathy* (e.g., *"Sounds great!"*, *"I am happy that's the*

Phrasing category	Question example	Matching prefixes*	Modification rules*
Adverb-based question	What gender do you identify as?	Can you tell me... Could I ask you...	
Verb-based question	Are you married?	Could I ask you whether... Please tell me if...	are you → you are i → you
Verb-based statement	Feeling tired or having little energy.	Have you experienced... Did you experience...	. → ?
Noun-based statement	This course requires us to understand concepts taught by the lecturer.	Would you say that… Is it true that...	are you → you are i → you us → you . → ?
Request-action	Indicate your current age.	Can you... Can I ask you to...	. → ?
None	If you've had any days with issues above, how difficult have these problems been?	N/A	N/A

Table 6.2: Examples of original survey items and the rephrasing resulting from the augmentation process. Phrases in-between square brackets have been added or modified.

case."), and *Expression of compassion* (e.g., *"That sounds stressful."*, *"I am sorry to hear that."*).

To select the most appropriate reaction I perform several steps. First, I classify the question text into 3 empathy framing categories: Positive, Neutral or Negative. This classification is somewhat similar to sentiment, but relies on how the question is framed in order to match the empathetic reaction. Specifically questions related to demographics or potentially "judgmental" topics should be classified as Neutral for empathy matching perspective. In the second step I classify the answer option into similar 3 empathy framing categories: Positive, Neutral or Negative. Positive framing for an answer communicates that the user expressed agreement with the question, while the Negative framing expresses disagreement.

Neutral answer framing represents uncertain answer, mixed answer, or categorical option without any clear opinion or sentiment. Certain categorical answer options can have their own intrinsic valence for empathy matching purposes, e.g., "Eviction", or "Crack/Cocaine" from the social needs survey would be classified as Negative due to the negative meaning of the concepts themselves. In the third step, the results of the question and answer classifications are combined using a fixed rule that decides on the reaction category to match. Non matching question and answer framing (i.e., Pos & Neg or Neg & Pos) would match *Expression of compassion*, while matching framings (i.e., Pos & Pos or Neg & Neg) would match *Expression of satisfaction*. Presence of Neutral framing in either would match *Neutral acknowledgment*. Appendix A presents examples of question and answer context in which different reaction categories would be appropriate. Selection of concrete text from that reaction class is done probabilistically and with keeping track of use frequency in the same fashion as for prefix selection described earlier.

Task 4: Progress communication

The goal of this task is to communicate the progress of the exchange to the user and also to further break the repetitiveness by injecting additional phrases related to topic & section management. This task injects text from two repository classes: Progress communication phrases and Topic continuation & topic switching phrases Table 6.3. Addition of these phrases is not based on any data-driven ML components, but is probabilistic and injected every n-th survey item as controlled by a meta-parameter. Progress communication phrases have particular subclasses of phrases that only apply for middle and close to the end of the survey, this is to introduce additional novelty to the dialogue. Similarly to the other tasks, the selection of the concrete text being added from each class is probabilistic with keeping track of the frequency of use to minimize repetitiveness. It is important to note that topic switching phrases are not necessarily well aligned with any actual change in the survey section or topics of questions.

Type	Concrete survey Instantiation
Progress - any place	"We are currently at question [4] out of [11]." "We are done with [20%] of our questions." "We have completed [70%] of the survey."
Progress - middle of the survey	"We are now in the middle of the survey" "Half of the survey is done, thanks for your patience" "We're halfway there, still [2] questions to go"
Progress - close to the end	"We are almost done, thanks for your patience" "We're mostly done, just [4] questions left" "We are almost at the end of the survey. I appreciate your patience."
Topic continuation	"Let me ask you some more..." "Just a few other things I wanted to ask you about..."
Topic switching	"Let's switch to something else..." "Let's move on to some other questions..."

Table 6.3: Template examples of progress communication and topic switching phrases and their instantiations for specific surveys. Phrases in-between square brackets are survey specific slots that are filled-in dynamically.

6.2.4 Automation of Augmentation Tasks

Out of the four conversational adaptation tasks described in the design section only progress communication does not rely on data-driven ML components (i.e., it adds progress phrases probabilistically). The automation, however, is also controlled by several meta-parameters defining the prevalence of different augmentations. Non-repetitive selection of concrete text phrases to be injected is based on keeping track of frequency of use and not learned from data. Here, I describe the datasets used for training, testing and validating the ML components. I also describe the data-driven ML components themselves as well as non-data driven automation.

Development Survey Dataset

The dataset used for initial training and testing the ML components included 16 surveys - 4 demographic surveys, 2 social needs surveys, reflection survey (Kember's reflection), stress survey (PANAS), physical activity motivation survey (TPB), workload survey (NASA TLX), depression survey and a few others (see Appendix D). Most of the surveys were related to health, wellbeing and behavior change given the nature of my work. The dataset included both validated instruments used in research as well as informal questionnaires. All the surveys have been represented in a common JSON format adapted from their original sources (PDF, Website, Word document). This adaptation step is still manual at this point. The text of questions and answers has been extracted from each survey and manually labelled by the author to provide the data for training and evaluation of the ML components. The 269 extracted survey questions were labelled for rephrasing tasks (6 phrasing categories) and for empathy framing tasks (3 categories). The phrasing labelling resulted in 138 (51%) of the questions labelled as 'adverb-based question' and the rest as other phrasing categories (Figure 6.1).

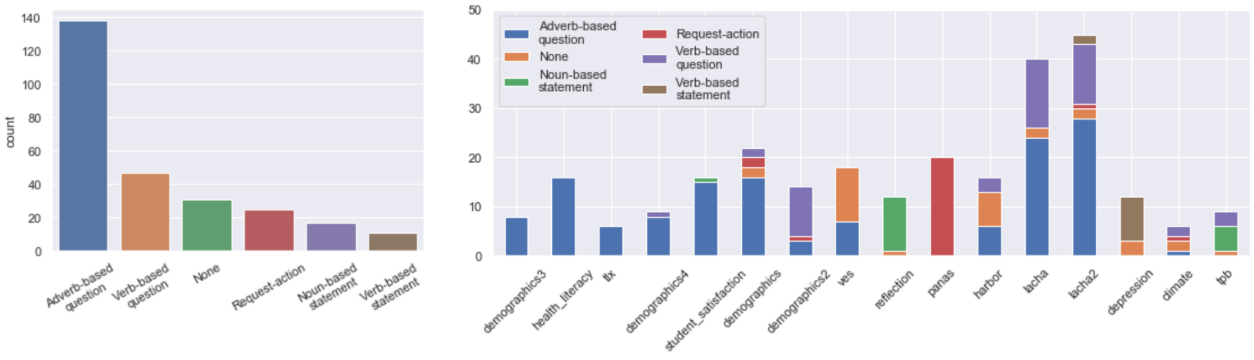

Figure 6.1: Distribution of 6 question phrasing categories in general and across the 16 development surveys.

Question labelling for the purpose of empathy question framing resulted in 106 (39%) of the questions receiving a Negative label, 85 (32%) receiving a Neutral label and 78 (29%)

labelled as Positively framed (Figure 6.2).

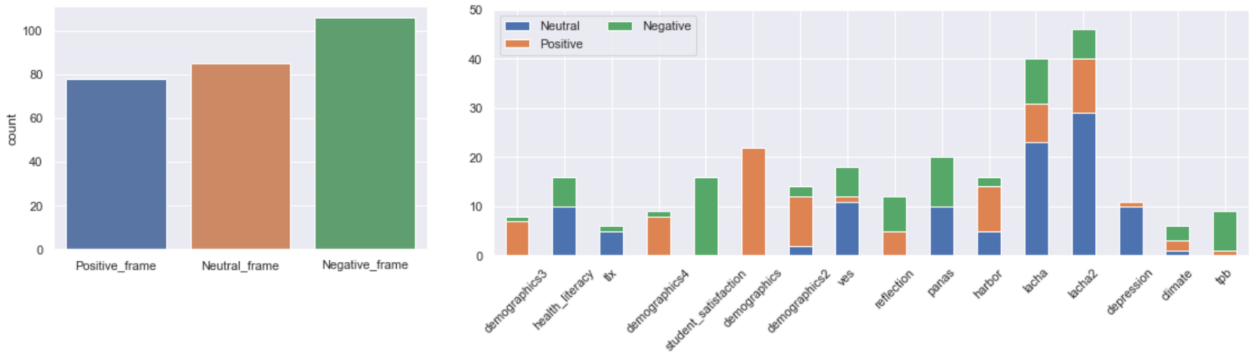

Figure 6.2: Distribution of 3 empathy question framing categories in general and across the 16 development surveys.

The answer dataset was composed of 577 answers extracted from the 16 surveys and 138 answers representing standard likert-scales extracted from [228]. The labeling of these answers for empathy framing resulted in 382 (53%) answers labelled as Neutrally framed, 186 (26%) labelled as Negatively framing, and 147 (21%) labelled as Positively framed (Figure 6.3).

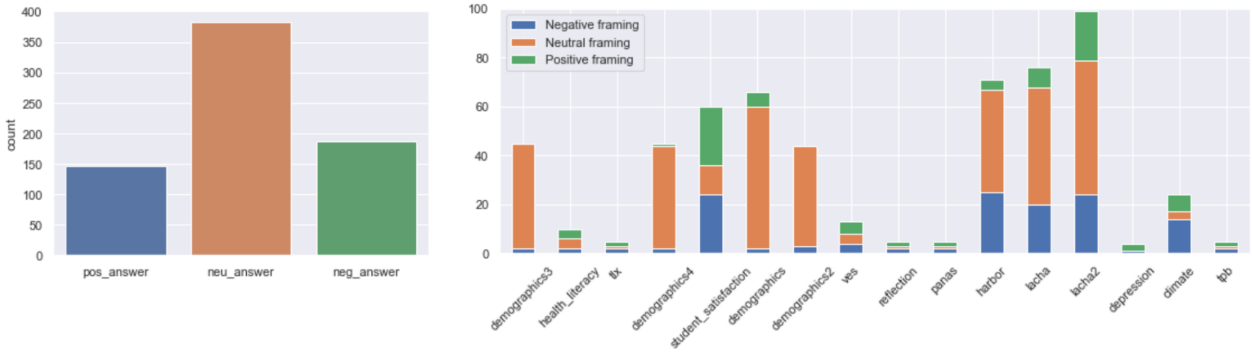

Figure 6.3: Distribution of 3 empathy answer framing categories in general (this includes the 138 answer examples extracted from common likert-scales [228]) and across the 16 development surveys.

Hold-out Survey Dataset

I selected 6 additional hold-out surveys after the conversion approach was finalized to evaluate the conversion performance on a range of surveys with potentially challenging properties (see Appendix C). Two of the surveys are informal, while others have been featured in published research. The surveys also employ different question phrasing (e.g., 1st person, 3rd person or mixed) and rely on different answer options (likert-scale-based vs custom scales). In summary the hold-out surveys are used to: 1) evaluate the amount of manual user corrections needed to make them applicable to the end-user scenario, and 2) collect user feedback on self-reported engagement, usability and quality of conversational elements in the user study.

I employed the same labelling process for the hold-out surveys. The 88 questions extracted from these 6 surveys were labelled for phrasing and question empathy framing. For phrasing, 51 (58%) of the questions were labelled as 'Noun-based statement' (Figure 6.4 A). For empathy question framing 37 (42%) were labeled as Positively framed, 34 (39%) as Neutrally framed, and 17 (19%) as Negatively framed (Figure 6.4 B). Labelling of the 97 answers resulted in 39 (40%) answers labelled as Neutral, 37 (38%) as Positive, and the remaining 21 (22%) as Negatively framed (Figure 6.4 C).

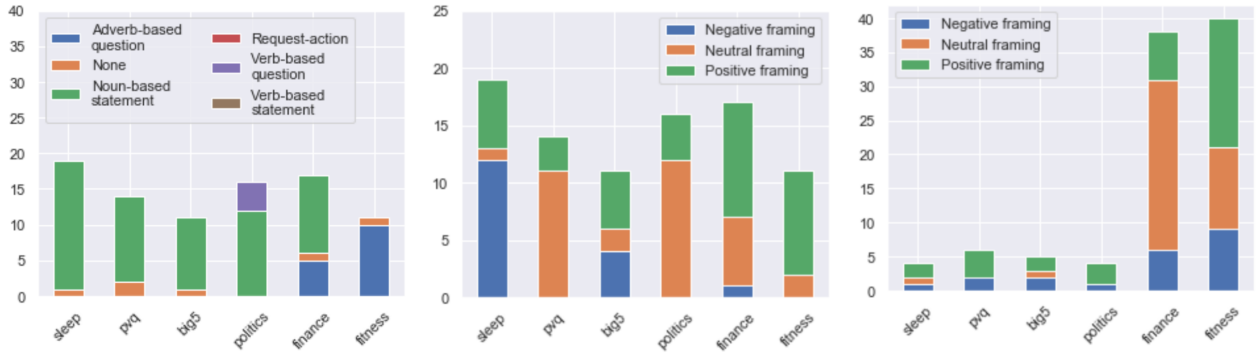

Figure 6.4: Distribution of labels in the 6 survey hold-out dataset. From the left: A) Distribution of question phrasing classes among surveys, B) Distribution of Empathy Question Framing classes, C) Distribution of Empathy Answer Framing classes.

Machine-Learning Components

The data-driven ML components fueling 3 of the 4 augmentation tasks (i.e., except for progress communication) have been framed as a text classification problem. I use the Scikit-learn ML library as well as Spacy NLP toolkit to process the data as well as train and evaluate the text classification tasks. While all the tasks have been framed as text classification problems, the specific nature of each task and the data pose different challenges. These properties result in different combinations of optimal features, preprocessing, and data augmentation (Table 6.4).

The survey domain detection presents several challenges: 1) survey can be on virtually any topic - open domain, 2) multiple suitable domains can be selected, e.g. sleep survey would work well with "health", "sleep" and "wellbeing" designations - hence the task can be seen as multi-label, 3) survey domain name must be suitable for use as name for the bot, e.g., HealthBot might be a more suitable name for a survey evaluating depression than DepressionBot. I address these challenges by defining a large, but curated, set of possible survey domains. I extract 90 different domain names from Wikipedia topics page (e.g., "culture", "health", "biology") and adjust the domain names to make them usable for the employed chat naming scheme. This set is likely to cover all possible survey domains at least at the high level of abstraction (e.g., may not contain "sleep", but will contain "health" domain). A similar approach to handling topics has been used in [75]. Given that the surveys in my dataset represent only a handful of domains, I seed the data for each domain with 30 most similar worlds generated by Spacy (split into 3-4 groups of words to provide few multi-word examples for each domain). During the survey classification I extract only nouns and verbs from a given survey (based on pos-tagging) and calculate an average of the embeddings from these extracted keywords to represent the whole survey (Table 6.4). Nouns and verbs are much more indicative of topical domain than other parts of speech [159] and embeddings are able to capture similarity of meaning [236].

Many, especially academic, surveys use variations of standardized likert-scales in their

answer options. I take advantage of this by adding a set of standard likert scale answers extracted from [228] to my training data. This subset is used in all the classification in addition to any survey specific data.

Through experimentation I have selected different sets of features for different classification tasks (Table 6.4). *Question Language Adaptation* benefits particularly from including part-of-speech tagging, while the use of word embeddings does not seem to be valuable. This is perhaps not that surprising given the language-structure specific nature of the task. *Question Empathy Framing* classification on the other hand benefits from embeddings and not from pos-tagging, which is not surprising as the meaning and contextual use of the words is important for this task. Value of word bi-grams is harder to explain. *Answer Empathy Framing* classification benefits from character level bi-grams, which makes sense given the short length of most answer options. In general effectiveness of combined use of word grams and embeddings is a bit surprising, perhaps both symbolic and neural representations convey valuable information. LogisticRegression also was more effective than SVM, when embeddings and n-grams were used together.

Conversion Meta-parameters & Non Data-driven Automation

Aside from data-driven ML automation, several components of the conversion approach are non data-driven. While the selection of augmentation category might be based on ML-based text classification, the concrete text to use from a repository is selected probabilistically with keeping track of the frequency of use to minimize repetitiveness. The following procedure is followed. A concrete text to use from the category is selected at random from a set of least frequently used texts (minus text used last time). Each use of the concrete text in the category is tracked.

Several metaparameters control the conversion process at the higher level (Table 6.5). The frequency of injection of the progress communication phrases as well as the frequency of topic phrases are both controlled by meta-parameters which define that these phrases would be injected every n-th survey item. The injection of the reactions and the use of empathetic

	Question Language Adaptation	Question Empathy Framing	Answer Empathy Framing	Survey Domain
Training data (source)	269 (16 surveys)	269 (16 surveys)	577 (16 surveys) + 138 (std. likert)	75 (16 surveys) + 340 (Spacy)
Categories	6 phrasing types	3 framings	3 framings	90 domain names [Wiki]
Classifier	Linear SVM with SGD training (penalty='l2')	LogRegression (penalty='l2', solver=liblinear)	LogRegression (penalty='l2', solver=liblinear)	SVM (C:0.5)
Preprocessing	lower-case	lower-case	lower-case number replace	noun, verb only extraction
Features	word bi-grams pos bi-grams	word bi-grams embeddings*	word bi-grams pos bi-grams char bi-grams embeddings*	embeddings*

* Embeddings are provided by Spacy (https://spacy.io/) from *en_core_web_lg* model with 300 dimensions

Table 6.4: Summary of the setup of between different classifiers supporting the augmentation tasks. The best setups have been determined in a limited parameter exploration on development set (however, no exhaustive grid-search has not been performed).

reactions (as opposed to always using only neutral acknowledgments) is controlled in a similar fashion, but by a parameter defining the probability of a survey item getting a reaction.

6.3 Evaluation

The purpose of the evaluation was to: 1) understand how well the proposed automated conversion can perform to support survey administrators in engaging their audience (**RQ2**) and 2) further identify and understand the aspects of the conversion approach which are handled well and the ones that are still problematic (**RQ3**). The evaluation is organized as a 3-step process: 1) evaluation of ML components performance, 2) manual correction effort and 3) user study based evaluation. I evaluate the performance of the ML components

Parameter	Description	Default value
Progress update frequency	Controls how frequently (every n-th item) should the progress utterance be injected into the dialogue	Every 5-th survey item
Reaction frequency	Controls how frequently (every n-th item) should the chatbot react to user answers	Every item
Reaction empathy probability	Controls the probability with which the reaction will be empathetic if possible (as opposed to always neutral)	100% - reacts to all answers
Question rephrasing probability	Controls the probability with which the questions will be rephrased to a conversational form.	100% - all questions

Table 6.5: Meta-parameters controlling the automated conversion

using performance metrics - classification accuracy, weighted F1 score in a cross-validation and leave-one-out evaluation setups (Figure 6.5 - ML performance). This captures the data-driven automation performance. Further I carry out a user evaluation. First on a hold-out set of 6 unseen surveys I evaluate the user correction effort (e.g., grammatical or other language issues that need to be corrected manually). This represents the additional effort a survey administrator would have to put in order to make the automatically converted surveys ready for end-user administration (Figure 6.5 - Correction effort). Then I evaluate the impact of the adapted conversational surveys (with minimal corrections) on survey respondents' self-reported engagement, usability and the quality of the conversational elements (Figure 6.5 - User study evaluation).

6.3.1 ML Performance

Out of the four conversational adaptation tasks described in the design section only progress communication does not rely on data-driven ML components (i.e., it adds progress phrases probabilistically). The remaining tasks of: 1) adding introduction & closing, 2) modifying questions to conversational form and 3) adding reactions to user answers, all rely on text classification ML components. Adding introduction & closing relies on domain classification

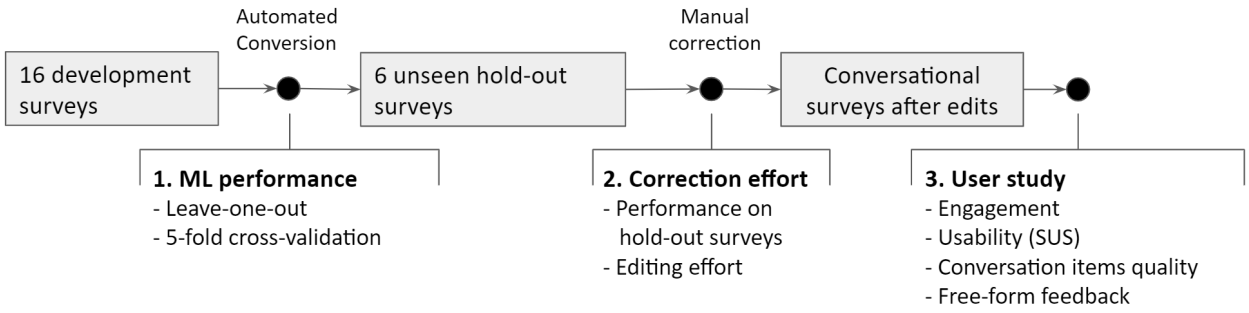

Figure 6.5: The 3-step evaluation process: 1. ML performance - evaluation via accuracy and F1 score in leave-one-out and 5-fold cross validation setups. 2. Correction effort - manual editor effort needed to correct basic issues (e.g, grammatical errors). 3. User study - impact of the adapted conversational surveys on engagement, usability and the quality of the conversational elements.

for filling-out the bot name and domain slots in the text templates. Modifying questions to conversational form relies on phrasing classification to select the most appropriate prefix and utterance rephrasing rules. Finally, adding reactions to user answers relies on the results of two text classifiers that classify empathy question and answer framing.

I evaluate the performance of the classifiers measuring accuracy and weighted F1 score metrics (due to class imbalances in some of the tasks) in a 5-fold cross-validation as well as leave-one-out evaluation setups. Cross-validation uses data across separate surveys and captures more of a within survey performance (i.e., items from the same survey are likely present in the training and testing data). Leave-one-out evaluation is more indicative of likely performance on new, unseen surveys. Specifics of how these metrics are calculated are given in the subsequent section on measures.

6.3.2 Correction Effort

Accuracy and F1 score capture the performance of ML components according to the provided problem scoping (i.e., text classification), but may not capture other potential issues. These could be related to grammatical mismatches (e.g., prewritten prefix template does not fit well

with the question despite correct classification), more nuanced aspects, such as "awkward" phrasing of a reaction in a particular context or other unforeseen issues outside of how the design was framed for automation. From a survey administrator's perspective correcting these mistakes would necessitate manual edits of the text. To evaluate such editing effort I calculate the edit distance (also known as Levenshtein distance) defined in the measures section. This is calculated between the phrasing resulting from automation and the corrected version that I have developed by hand to fix the grammar with minimal edits.

The applied corrections were limited to 2 aspects: 1) corrections of grammatical errors in the question rephrasing (see Appendix F), 2) corrections of mismatched empathetic reactions. In case of reactions the correction would involve replacing a mismatched reaction with a pre-written reaction text taken from the correct category from the repository (see Appendix G). I focused only on these corrections as they can be objectively evaluated and represent the minimal set of changes needed for making the conversational survey presentable to the end-users. This is essentially mimicking the edits that a survey administrator would need to apply at the very minimum. No other types of changes to the automatically generated conversational survey text are applied and any other potential "issues" are presented to the users in the subsequent user study.

6.3.3 User Study

The purpose of the user study was to: 1) capture user perception of conversationally adapted surveys, 2) evaluate the quality of specific conversational adaptation elements, 3) collect qualitative feedback. The study followed a between subject setup, where each participant was exposed to only one randomly selected condition (i.e., one of the 6 hold-out surveys). The setup has been approved by a university IRB and deployed on the Amazon Mechanical Turk (AMT) crowd-working platform.

Participants: 30 AMT participants were recruited (5 per each conversationally adapted survey). The participants were at least 18 years old and residing in the U.S. They were also

required to have completed at least 100 AMT tasks in the past with at least 90% approval rate. Each participant interacted with only one survey. One participant completed the task on a mobile device, while others used non-mobile devices. Participants spent on average 8:31 min (SD: 7.53 mn) on the task and the compensation was $1.5 (average rate: $10.50/hour).

Procedure: Participants were first asked to read and approve a consent form, which detailed study procedures, participant rights, compensation and research staff contact information as required by the IRB. They then completed a conversationally administered survey (each participant answered one survey). The chat interface was based on a BotUI framework with minor survey specific usability modifications introduced in the HarborBot study [129]. Following the chat interaction, the participants were asked 6 questions about engagement (detailed in the measures) and prompted for first-impressions free-form feedback. They then answered 10 usability questions (SUS survey detailed in measures) including an attention check question. Lastly, the participants were asked for feedback on the 4 conversational augmentations introduced to the survey. The page has been divided (see Figure 6.6). On the left side (top on mobile) the log of the conversation with red-highlighted phrases of interest was presented. The right side (bottom on mobile) included questions asking for quality evaluation and for free-form feedback. This setup was introduced to aid with recall and to present the conversational phrases in their context. Participants were asked to evaluate the overall quality and give free-form feedback for: 1) chat reactions to their answers, 2) progress communication, 3) introduction & closing, and 4) chat questions. The free-form feedback prompt changed depending on participant's quality score. This was done not to render one choice less effortful than other (AMT workers can be inclined to complete the task as fast as possible [60]). The final question asked for free-form feedback about anything "missing" that could improve the survey answering experience. The whole setup was tested for proper rendering on desktop and mobile devices.

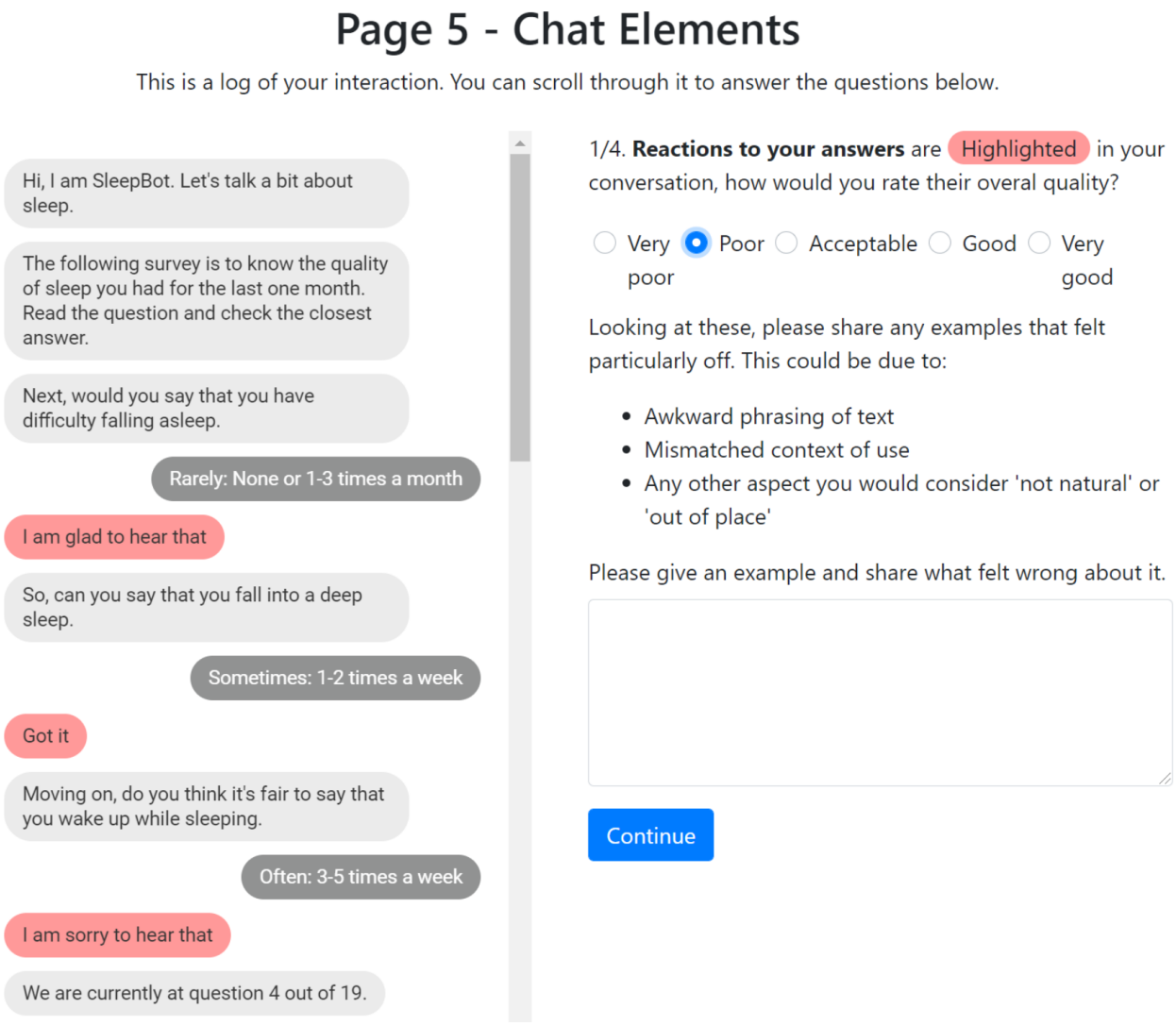

Figure 6.6: Last page of the AMT user study asking for feedback on particular conversational augmentation design elements. On the left participants were shown the log of their exchange with red-highlighted phrases of interest. On the right they were asked to evaluate the overall quality of the phrases as well as to give detailed free-form feedback. Pressing "Continue" would ask them to evaluate another aspect (red highlights in the conversation would change accordingly).

6.3.4 Measures

ML Evaluation Measures: All the ML components used are text classifiers. I evaluate their performance using standard metrics of Accuracy (fraction of correct category predictions) and weighted F1 score (a weighted average of the precision and recall) as provided by the Scikit-learn library [1]. I select the weighted variant of F1 score to account for label imbalance in some of the tasks (e.g., in empathy answer framing, 53% of the examples are labelled as Neutral out of 3 classes, consistently assigning a Neutral label could yield 53% accuracy).

Correction Effort Measures: I measure the manual correction effort by edit distance defined as a minimum number of single-character edits (insertions, deletions or substitutions) required to change one string into the other [246]. For example, a correction from *"Would you mind sharing do you have a workout buddy?"* to *"Would you mind sharing whether you have a workout buddy?"* would require 2 substitutions (*"d"* → *"w"*, *"o"* → *"h']]*) and 5 insertions ("ether") for a total edit cost of 7 characters.

User Study Measures: Participants evaluated the conversationally adapted surveys in terms of engagement using 6 questions adapted from O'Brien's engagement survey [180] (e.g., *"I was really drawn into answering questions"*, *"I felt involved in answering questions"*, *"This experience of answering questions was fun"*). The same engagement questions were used in my prior work on social needs screening with HarborBot described in Chapter 5. Usability was evaluated using System Usability Scale (SUS) [18] using 10 questions adapted to chat context (e.g., *"I think that I would like to use this chat interaction frequently."*, *"I thought there was too much inconsistency in this chat interaction."*, *"I thought the chat interaction was easy to use."*). Additionally the 4 design aspects (i.e., reactions, progress, introduction & closing, and question phrasing) were evaluated in terms of quality on a 5-point scale from (*"Very poor"* to *"Very good"*, with mid-point set to *"Acceptable"*).

[1] https://scikit-learn.org/stable/modules/model_evaluation.html

Several questions also asked for free-text feedback on overall aspects of the interaction (i.e., *"Please share any aspects of the interaction that you felt were particularly bad or good for your experience"*, *"Please share one aspect that was missing in the interaction and that you think would be valuable for improving your experience"*) as well as for specific conversational augmentation elements (e.g., regarding reactions to answers *"Please give an example and share what felt wrong about it."*).

6.3.5 Analysis

ML Evaluation Analysis: In a leave-one-out evaluation the accuracy and weighted F1 scores are calculated per validation survey (i.e., the survey not used for training) and then averaged across all the surveys. In the 5-fold cross-validation, the data is randomly split into 5 parts, with 4 used for training and the 5-th used for testing.

User Study Analysis: The 6 engagement questions showed high internal consistency (α=0.86) and were averaged to form an engagement score (same process as in [129]). Given the meaning on the 5-point likert scale, average values above 3 represent positive engagement. The 10 SUS items also showed high internal consistency (α=0.83). Scoring of the SUS survey involves adding up all the items and multiplying the result by 2.5 to form a score from 0 to 100. Past research indicates that SUS scores above 68 represent above average usability[2]. I also used a mixed-effect model with quality ratings for 4 design aspects (intro & closing, reactions, progress, question phrasing) as predictors and the engagement rating as a predicted outcome. I control for survey repetition by including survey id as a random effect. I used the model to examine the impact of augmentations on user engagement.

The qualitative feedback is coded and grouped into themes relating to strengths and weaknesses of applied conversational adaptations as well as based on feedback for particular conversational augmentation aspects.

---

[2]Scoring System Usability Scale (SUS) - https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

6.4 Results

The results are presented according to the 3-step evaluation process: 1) evaluation of ML components performance, 2) manual correction effort and 3) user study based evaluation. Conceptually the ML performance evaluation estimates how well the proposed conversational adaptation design can be executed automatically on unseen, but similar surveys. Manual correction effort estimates how much work is needed from a human survey administrator to correct the basic language mistakes and empathy mismatches resulting from the proposed automation (this is evaluated as if corrections were made on raw text output). Finally, the user study evaluation captures the impact of the automatically augmented surveys, after applying the minimal manual corrections, on survey respondents. I present the findings from each evaluation step, discuss insights and potential improvements for the future.

6.4.1 ML Performance

Table E.1 presents evaluation of the performance of different classifiers used for survey adaptation tasks. *Question Language Adaptation* classification selects the most appropriate prefix to be used for rephrasing the survey item and ensuring "conversational style" (3rd person and question form). The 83% accuracy in cross-validation indicates that the classifier performs much better than random (17%) or simple majority class selection (51%). Small discrepancy between the weighted F1 and accuracy scores indicates that class imbalance (51% of the data is labeled as only one of 6 classes - 'adverb-based questions') is not negatively affecting performance. The fairly large gap between cross-validation (83%) and leave-one-out evaluation (68%) performances suggests that some question phrasing categories are present in a small subset of surveys and not common across all the surveys (e.g., indeed PANAS survey [232] provides 19 of 30 examples for 'request-action' question phrasing).

Question Empathy Framing and *Answer Empathy Framing* are two text classifiers used jointly to decide on the empathetic reaction class to present to the user (detailed in the design section). Both classifiers select from 3 classes (positive, negative and neutral). In the

		Survey Domain	Question Language Adaptation	Question Empathy Framing	Answer Empathy Framing	Empathetic Reaction (derived)†
5-fold CV	**Acc**	0.64±.04	0.83±.07	0.81±.03	0.89±.04	0.73±.05
	F1	0.64±.04	0.82±.08	0.81±.03	0.89±.04	0.72±.05
Leave-one-out	**Acc**	0.62±.45	0.68±.33	0.69±.24	0.89±.13	0.57±.15
	F1	0.64±.45	0.68±.34	0.68±.25	0.89±.12	0.55±.14

† - Empathetic Reaction is derived from the Question Empathy Framing and Answer Empathy Framing using a fixed rule.

Table 6.6: Classification performance for the 4 text classification tasks (+1 derived) used in automated conversational survey adaptation. Question Empathy Framing and Answer Empathy Framing classifications are part of empathetic addition - the results of these two classifications taken together are used to decide on reaction class

case of *Question Empathy Framing* the classes are fairly balanced, which suggests that the achieved 81% accuracy is better than expected random (∼33%). Similarly to the *Question Language Adaptation*, the better performance in cross-validation (81%) than in leave-one-out evaluation (69%) suggests that a particular question framing tends to be overrepresented on a subset of surveys (indeed the demographics surveys have questions mostly labelled as neutrally framed). The *Answer Empathy Framing* classification also performs better than expected random (∼33%) or majority class selection (53%) with an average accuracy of 89% (answer classes are fairly imbalanced in the data), but lack of discrepancy between weighted F1 and accuracy scores does not suggest this to be a problem. Contrary to other classifiers, the *Answer Empathy Framing* classification performs similarly in cross-validation and leave-one-out evaluations, suggesting that answer framing is more reusable across surveys. This classifier also benefits from being seeded by additional external cross-survey data for common likert-scales containing 135 examples comprising ∼17% of the dataset. Removing this data from training slightly decreases performance for leave-one-out evaluation (Acc=$0.86 \pm .15$,

Hold-out Survey	Question Language Adaptation (% correct)	Question Empathy Framing (% correct)	Answer Empathy Framing (% correct)	Empathetic Reaction† (% correct)
Big 5	100%	55%	100%	60%
Fitness Survey	82%	64%	58%	35%
Personal finance	71%	71%	69%	56%
Political views	88%	44%	75%	33%
PVQ values	100%	93%	100%	92%
Sleep quality	67%	67%	75%	50%
Mean accuracy	**84.43%**	**65.34%**	**79.42%**	**54.24%**

† - Empathetic Reaction is derived from the Question Empathy Framing and Answer Empathy Framing using a fixed rule

Table 6.7: Classification performance on 6 hold-out surveys used for correction effort estimation and in user study.

F1=0.86±.17), but increases performance in cross-validation (Acc=0.90±.01, F1=0.89±.01) introducing some of the same imbalance present in other classifiers.

In general, the leave-one-evaluation yields lower performance than cross-validation suggesting unique aspects of specific surveys not necessarily shared across all the surveys in the dataset. Furthemore, large standard deviations of 0.33 for *Question Language Adaptation* and 0.24 for *Question Empathy Framing* in leave-one-out evaluation indicates that the performance varies a lot for different surveys.

Table 6.7 presents the classification evaluation results on the set of 6 hold-out surveys. Better accuracy in *Question Language Adaptation* suggests the hold-out surveys are very similar to most of the development dataset surveys in that respect. I explore some of the performance differences on hold-out surveys in the subsequent section on correction effort. Performance on the full dataset (combined 16 development surveys & 6 hold-out surveys)

can be found in Appendix E

In summary, the classifiers used for various conversational adaptation tasks perform better than random or majority class selection. *Answer Empathy Framing* classification seems more reusable across surveys than *Question Language Adaptation* or *Question Empathy Framing*. This classifier also benefits from additional cross-survey training data from standardized likert-scales, which contributes to its better performance in leave-one-out evaluation.

6.4.2 Correction Effort

Each of the hold-out surveys required some corrections (Table 6.8). Normalized by the characters automatically added in the conversational adaptation, the manual edits comprised 30.06% of the automatically added text on average. The Political views survey required most edits (42.12%), while the PVQ survey the least (18.75%). It is worth noting that while misclassifications almost certainly necessitate manual correction need (i.e., unless rephrasing resulting from misclassification is still grammatically correct from editor's perspective), some corrections might also be needed even if the classification is correct (i.e., in case rephrased question has grammatical issues despite seemingly correct classification). In fact, editing needs in spite of correct classification may signal a systematic issue in the problem definition. In total, question language adaptation corrections accounted for 13.65% of all the edits. This is a sum of 2 sources for such corrections: 1) language adaptation misclassification (5.15%) and 2) missing replacement rules (8.48%). Empathetic reaction corrections accounted for 84.05% of edits, caused by: 1) *Question Empathy Framing* misclassification (71.55%) and 2) *Answer Empathy Framing* misclassification[3] (12.48%). The remaining 2.31% of corrections were due to survey domain misclassification in the introductions.

[3]In case both the question and the answer were misclassified, the correction effort is counted under question misclassification to avoid double counting (from the editor's perspective a reaction would need editing only once).

Survey	Added (# chars)	Manual edits		Categories of manual edits (% of edits)			
				Question correction		Reaction correction	
		Edits (# chars)	Edits (% of added)	Phrasing misclassification	Missing replace rules	Question Empathy framing	Answer Empathy framing
Big 5	1410	496	35.18%	0%	8%	86%	0%
Fitness Survey	1312	453	34.53%	2%	0%	57%	38%
Personal finance	2755	601	21.81%	16%	0%	82%	0%
Political views	1847	778	42.12%	2%	0%	92%	5%
PVQ values	2341	439	18.75%	3%	43%	54%	0%
Sleep quality	2386	667	27.95%	6%	0%	58%	32%
Mean			**30.06%**	**5.15%**	**8.48%**	**71.55%**	**12.48%**

Table 6.8: Correction effort quantified as character edits per hold-out survey. The corrections represent minimal changes to the grammar and empathetic reactions needed from a survey administrator to present the conversational survey to end users.

Question Language Adaptation Corrections:

Question phrasing corrections were required when automated conversational rephrasing resulted in grammatical errors. Out of the 88 questions among the 6 hold-out surveys, 37 ($\sim$42.5%) required some form of editing (see Appendix F).

Question Phrasing Misclassification: The Personal Finance and Sleep Quality surveys required most editing effort due to phrasing misclassification (16% and 6% of the survey correction effort respectively). In the case of the Sleep Quality survey, 6 questions were

consistently misclassified as 'verb-based statement' instead of 'noun-based statement' resulting in rephrasing for e.g., *"I have difficulty falling asleep."* to *"Next, have you experienced i have difficulty falling asleep?"* instead of *"Next, would you say that you have difficulty falling asleep?"* Both phrasing classes are the least represented in the training data (only 11 and 17 example sentences respectively), which likely explains the misclassification. Similar situation takes place for the Personal Finance survey, where 4 questions are misclassified as 'noun-based statements' instead of more appropriate 'verb-based questions'.

Missing Text Replacement Rules: The PVQ survey required the most question edits due to missing replacement rules (43% of all the correction effort). The unique aspect of this survey was its 2nd person question framing, which was not the case for any other survey in the dataset. Lack of rules rephrasing 2nd person to 3rd person word use (i.e., *"he"* → *"you"*', *"himself"* → *"yourself"*) resulted in e.g. survey item *"It's important to him to show his abilities. He wants people to admire what he does."* being rewritten as *"Do you think that it's important to him to show his abilities. He wants people to admire what he does?"* which required 21 manual character edits (based on edit distance) to correct to *"Do you think that it's important to you to show your abilities? Do you want people to admire what you do?"*. It is important to note that the replacement rules are not learned from the data and require manual specification. There is, however, a finite set of replacement rules as they are based on personal, possessive and reflexive pronouns, which are a closed class. Also the replacements of 2nd person verbs (e.g., *"thinks"*, *"runs"*) can be accomplished via pos-tagging.

Verbosity & Multiple Viable Question Rephrasings: Two additional observations emerge from the question phrasing corrections. First, the conversational rephrasing might in some cases be unnecessarily verbose, for example *"I feel in control of my current financial situation."* is rephrased as *"So, is it fair to say that you feel in control of your current financial situation?"* while a more concise rephrasing would simply be: *"Do you feel in control of your financial situation?"*. This is partially an artifact of how the rephrasing was

designed to add diversification and ensure a consistently polite tone. An additional phrasing category could offer more concise rephrasing. Second, it seems that the same survey item could match more than one category of rephrasing, for example *"My finances are a significant source of worry for me."* could be rewritten using the 'verb-based question' category to *"Can you tell me whether your finances are a significant source of worry for you?"* or a 'noun-based statement' as *"Would you say that your finances are a significant source of worry for you?"*. Given that the answer options for this question are likert-scale from *"Not at all true"* to *"Very true"*, the second rewrite seems more appropriate. This in general might suggest that *Question Language Adaptation* classification might benefit from including the context of answer options into account.

Empathetic Reaction Corrections

Reaction corrections represented the largest portion of the correction effort (84.05%). On a per reaction basis, this represents 149 of 382 (39.0%) of the automatically provided reactions needing corrections.

Answer Framing Misclassification: Fitness and Sleep Quality surveys required the most rewriting effort due to answer misclassification (38% and 32% of all the rewrites respectively). The Sleep Quality survey uses a variation of a 4-point frequency scale for each question. One of the answer options (*"Sometimes: 1-2 times a week"*) has been consistently misclassified as Positive as opposed to Neutral, which resulted in the need for correction of subsequent empathetic reactions for every survey question. In the case of Fitness survey, the answer options are custom 4-point scales distinct for each question. Majority of misclassified answers have fairly ambiguous framing for empathetic reaction purposes. For example, the answer option: *"Yes but I don't always stick to it."* was misclassified as Positive, but in the context of the question: *"Do you have an exercise plan?"* this answer option would not go well with a chat reaction being either *"Great to hear that's the case"* (in case of Positive classification) nor with *"I am sorry to hear that"* (in case of Negative classification) and would better fit

a Neutral reaction such as *"Thanks for letting me know."* Several answer options have been misclassified in this survey due to answer options with mixed sentiment.

Question Framing Misclassification: Political Views survey scored particularly low for question framing classification (44% accuracy) and corrections of these misclassifications comprised 92% of editing effort in this survey. Indeed several neutrally framed questions were misclassified as negatively framed (e.g., *"A good government should aim chiefly at more aid for the poor, sick, and old."* likely due to the presence of keywords with negative sentiment such as "poor", "sick".) or positively framed (e.g., *"I would prefer a friend who is practical, efficient, and hard working."* likely due to keywords "efficient", "practical"). Framing classification is used for deciding whether chat reaction should be neutral or empathetic and is not equivalent to sentiment. In the context of a survey asking about political views, empathetic reactions would not be appropriate (i.e., judgmental). While the questions themselves contain keywords which can reveal something about the survey domain (e.g., "government"), lack of similar examples in the training data likely makes this challenging. The *Question Empathy Framing* in particular (as labelled in the data) implicitly relies on broader context and may benefit from explicit contextual information (i.e., explicit domain) or more training data (e.g., surveys representing different domains).

Nuance in Empathy Labelling & Oversized Impact of Certain Misclassifications: Two additional observations emerge from corrections of automatic empathetic reactions. First, the question and answer framing classification, as labeled in the data, rely on a more nuanced understanding of a broader context. With limited context and data it is hard to differentiate between classifying *"I feel vigorous after sleep."* in the sleep quality survey context as appropriate for an empathetic reaction, and recognizing that the question *"Someone who works all week would best spend the weekend trying to win at golf or other sport"* in the context of values survey is likely best matched with a neutral reaction (i.e., to avoid judgmental tone regarding someone's values). It is also worth noting that even appropriately labelling

such data for the empathy purpose can be challenging in itself. Secondly misclassification of question framing incurs a high correction cost in raw-text editing. For a misclassified question the reactions to all the answer options likely require rewriting. For example *"I feel vigorous after sleep."* classified as negatively framed, would result in reverse reaction valence for all the answer options (i.e., answer *"Rarely"* would result in *"That's great!"* and *"Almost always"* in *"So sorry about that"*). Similar situation happens if a survey heavily reuses a specific answer option that happens to be misclassified. While the editing cost is high with raw text, these misclassification scenarios offer an opportunity for an editing tool support. With such support relabelling the question or repeatedly used answer would require only one or two clicks (i.e. all the reactions under such a question could be automatically updated and changing the label for one answer could be automatically propagated to all the identical answers through the survey).

6.4.3 User Study: Quantitative Results

The user study results comprise quantitative self-reported evaluations of engagement, usability and quality of conversational augmentations as well as free-form qualitative feedback on general interaction experience and on specific conversational design elements.

Engagement

Participants reported positive average engagement with the conversational surveys of 3.73 (SD=0.90), where 3 represents the mid-point rating for 5-point likert scales used in engagement questions. For comparison, in my prior work the manual conversational adaptation of social needs screening survey (Chapter 5) received a mean score of 3.59 (SD=0.77) on the same engagement scale. Average reported engagement for all the surveys was above 3, with the conversational version of Big5 personality survey rated as most engaging 4.27 (SD=0.42) and Fitness survey rated the least 3.08 (SD=1.40). These differences were not statistically significant. Only 4 of 30 participants reported engagement lower than 3 and each for a different survey, suggesting no systematic issues with augmentation of a particular survey.

Usability

Participants reported an average usability of 70.17 (SD=19.11) for SUS survey on a scale from 0 to 100. Any score above 68 can be interpreted as above average according to [175]. The conversational version of Big5 personality survey received the highest usability score of 86.5 (SD=20.89), while the usability of Sleep quality survey was rated the lowest with score of 59.0 (SD=10.55). These differences were not statistically significant. 13 of 30 participants rated the usability of their surveys at below the average 68, indicating some potential usability issues. I look at qualitative feedback to understand these potential issues.

Quality of Conversational Augmentations

Participants were also asked to rate the quality of different augmentation utterances on a 5-point likert scale (from *"Very poor"* to *"Very good"*). These were highlighted in the context of their interaction (see Figure 6.6). All augmentations were rated high on the quality scale (see Figure 6.7). The introduction and closing phrases were rated the highest with 87% of users rating them as *"Good"* or *"Very good"*. The quality of reactions to user answers were rated the lowest, with 10% of the participants rating them as *"Very poor"* or *"Poor"* and only 70% as *"Good"* or *"Very good"*.

To check if added conversational augmentations are indeed positively impacting user engagement I used a mixed-effects model to predict engagement by user-reported quality ratings for different conversational elements (Table 6.9). To control for differences across surveys I include survey as a random effect. Empathetic reactions quality has a positive significant impact on engagement (β=0.34, p$<$ 0.05), while Introduction & Closing quality (β=0.34, p=0.067) as well as Question quality (β=0.34, p=0.071) are only weakly significant. The overall model fit is R^2=0.302. Given that engagement is measured on a 5-point likert scale, the effect sizes are all within half a point increase. It is hard to directly compare the effect sizes as reactions are much more frequent in a given survey than Introduction & Closing or Progress utterances.

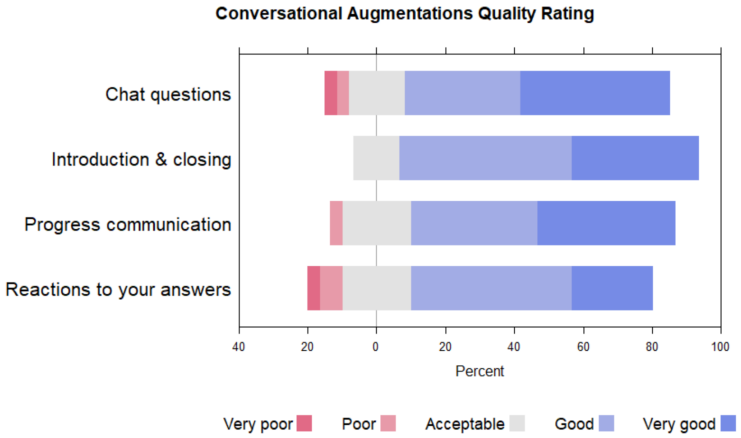

Figure 6.7: User rated quality of conversational elements in AMT study on a 5-point likert scale.

	Estimate(β)	S.E.
Reactions	**0.34***	.17
Progress	-0.31	.23
Introduction & Closing	0.50†	.27
Questions	0.27†	.15
Conditional R^2		.302

* p<0.05, † p<0.1

Table 6.9: Mixed-effects model predicting engagement by conversational element quality rating.

Survey respondents generally rated the quality of conversational augmentations high and the quality of these augmentations seems to positively impact engagement. It is interesting to note that usability was not correlated with engagement (r=0.125, p=0.51).

6.4.4 User Study: Qualitative Feedback

Here I present the themes from qualitative user feedback on the interaction as a whole as well as on specific conversational augmentation aspects.

Positive Perceptions

Most of the participants reported positive experience answering a survey with conversational chat. Several specifically described the chat as interactive & responsive (P1, P6, P15, P25), e.g., *"it felt very interactive did not really felt like chatting with a bot"* (P1). Participants also reported feeling comfortable and engaged in the interaction (P4, P9) e.g., *"All good. I felt comfortable and eager to engage."* (P4) and that the chatbot felt natural and easy to talk to (P5, P8, P10, P12, P18): *"It was a cool experience, the bot felt very natural and easy to talk too."* (P13). Several also described the interaction as easy to follow and straight-forward

(P16, P19, P29): *"I felt it was easy to follow along and answer the questions, this was good."* (P16). I further report the positive feedback for specific design aspects.

Empathetic Reactions Perceived as Good Quality: Several users perceived the empathetic reactions positively, reporting that they are of good quality (P5, P14, P26, P29), e.g., *"I think it is OK now"* (P5) and that they would not change anything (P4, P5, P14, P16, P18, P21, P25), e.g. *"I don't think they need to be improved upon."* (P16). One of the participants specifically described the reactions as natural, encouraging and pleasant: *"No it sounds natural and encouraging and pleasant really in my opinion"* (P15) and another considered them 'cute': *"I liked that there was a reaction, it was cute..."* (P9).

Progress Phrases Helpful & Appropriately Timed: Most participants considered the progress update phrases such as *"We are currently at question 4 out of 19."* not needing any improvements (27 of 30). Some specifically reported them as being natural, e.g. *"They sounded completely natural to me."* (P3) as well as helpful, e.g., *"Nothing needs to be improved, it is simple and helpful."* (P16) and reported them as concise and informative: *"This was on point and thoroughly appreciated"* (P12). They also reported the progress updates to be provided at just the right frequency: *"I thought it was just the right amount of updating."* (P27) or *"The progress info was provided timely and effectively, I liked it"* (P30).

Introduction & Closing Uniformly Perceived as Good Quality: The introduction and closing were considered of good quality by almost everyone. Several participants explicitly described them as appropriate: *"It was perfectly fine and appropriate."* (P27) and not needing any improvements: *"No need for improvement."* (P16). Only two participants suggested possible additions. One related to the name and more information about the bot: *"Any formal name? How the Bot was created?"* (P5) and the other comment suggested improved conversation closing: *"It could say goodbye."* (P7).

Conversational Question Adaptation Natural & Human-like: Participants gave feedback on survey questions in their final conversational form (i.e., 3rd person question rephrasing and prepended prefix), without being aware of how the question looked like in the original form-based survey. Several participants reported the questions being 'natural' and feeling like a part of the conversation (P3, P5, P14, P15, P30), e.g., *"They sounded completely natural and like a normal part of conversation"* (P3). and reported the questions to be well-formulated which suggests the adaptations blended in well with the whole questions text: *"The questions were well formulated and gorgeous, nothing to say about them"* (P30). One participant referred specifically to the conversational prefixes forming a good connection between questions: *"I thought the reaction into the next question was the more natural, human like of the whole thing"* (P15). Several other participants explicitly reported questions to be of 'good' quality (P8, P10, P17, P29) and others reported that no changes are needed (P4, P11, P18, P19, P21, P24, P25), e.g., *"Great questions, no improvement."* (P19).

Remaining Challenges

While most users perceived their chat-based survey experience positively, those that did not pointed mostly to issues with empathetic reactions. One of the participants found the reactions to every answer a bit artificial and awkward when coming from a stranger in a personality survey: *"The feedback on every answer was the only thing that sounded artificial. 'That's hard to hear' is a somewhat strange response from a stranger when answering personality questions."* (P3). Other two felt they were judgmental: *"I hate that he would go 'so sorry to hear that' like buzz off with your judgemental self."* (P24) and *"I didn't like it when the bot said 'that's hard to hear.' Like gee thanks, it's good to hear parts of my personality disgust you."* (P27). The second issue is actually a mislabeling problem, rather than an intrinsic design issue. Two of the participants also pointed to the interaction still feeling a bit mechanical: *"I thought the chat was a bit mechanical and didn't feel personal."* (P11) and a little repetitive: *"It was a little repetitive, but not bad overall, and surely interesting"* (P30). In the subsequent sections I report more detailed improvement opportunities

for specific conversational aspects.

Reactions Suffer from Mismatched Empathy & Could Make Better use of Contents: A few participants reported wanting the reactions to be more elaborate (P8, P11), e.g., *"Make it so that the sentences seem more complete."* (P11). This is likely for short neutral reactions present in the repository such as *"Got it"* or *"Noted"*. Others felt that the reactions could make better use of specific answer context (P17, P2, P19, P25), e.g., *"It could mention that it received your answer (repeating the answer back to you)."* (P2) or *"Maybe a unique reply to my preference in relation to the topic asked."* (P19). The biggest reported issue with the reactions was when participants felt judged or patronized by an attempted empathetic reaction in the seemingly wrong context (P3, P7, P24, P30). Due to this perception the chat was described as over-familiar: *It's awkward and maybe a little over-familiar"* (P3) and *"judgey"* (P24). One user was specifically unhappy about an empathetic reaction in the context of a personality survey: *"I did not like the phrase "that's hard to hear" because it's who I am, why would that be hard to hear. Also, why is the bot sorry to hear that I'm not trusting?"* (P28). Other participant felt that there actually is no need for empathy and just acknowledgments would be sufficient: *"There's no need for too much empathy, acknowledging my replies is enough"* (P30) or that the empathetic reactions make the chat less natural by making it too polite: *"not interacting like a normal human too polite"* (P1).

Progress Repetitive & May Decrease Respondents Attention: Despite generally positive perceptions of progress communication, a few users reported specific improvement opportunities. For some the specific topic switching phrases such as *"Let's move on to talking about a few more things..."* felt out of place, e.g. *"it acts like it is going into something else but basically asks me what it should already know from the above"* (P7). These phrases are indeed treated as part of progress update and added probabilistically, not necessarily separating questions on different topics. For others the progress utterances felt a

bit repetitive: *"seems repetitive"* (P1) and one user reported that having such information in general can actually rush answering and negatively impact attention: *"I don't think these percentages should be used because it may rush workers on answering rather than giving their full and undivided attention."* (P22).

Conversational Question Adaptation Slightly Repetitive & Can be Personalized: There were very few negative perceptions of question phrasing from just a few users. One user reported the conversational survey questions still felt repetitive: *"questions are repetitive"* (P1) and another felt that the chat seems to ask for the same information repetitively: *"questions seem to ignore the preceding answer instead of relating to my input."* (P7). This is likely an artifact of a survey asking similar questions to measure the same latent construct. One participant also suggested further personalization of the questions with his/her name, e.g,. *"Possibly asking for a name and then personalizing the messages each time."* (P22).

6.5 Discussion

I first discuss the results in relation to my research questions. For **RQ1** about supporting conversational adaptation with automation, I proposed an automated process consisting of 4 augmentation tasks: 1) Adding introduction & closing, 2) Adding reactions to user answers in question context, 3) Adding conversation progress communication, and 4) Modifying survey questions to fit conversational style. These tasks rely on retrieval of phrases from a reusable augmentation repository. My approach led to conversational surveys that can be deployed with respondents after applying only minor tweaks. In relation to **RQ2** about the impact on survey respondents, I have used the proposed process to automatically generate conversational versions of 6 unseen surveys. I further quantified the remaining survey administrator's correction effort (to manually fix misclassfications & grammar issues) and evaluated the impact of such adapted surveys with 30 participants. Mixed-methods results demonstrated: 1) positive self-reported engagement (comparable to manual conversational survey adaptation in my prior work), 2) positive impact of conversational adaptation ele-

ments on engagement (via a mixed-effects regression model), and 3) nuanced understanding of the engagement impact of specific augmentation aspects based on thematic analysis of qualitative feedback. Finally, in relation to **RQ3** about conversion aspects handled well & the problematic ones, I have shown that the fairly simple approach involving a repository and trained on a limited set of 16 surveys can achieve reasonable results leading to positive user engagement with only about 30% of automated augmentations needing manual corrections. Further discussion focuses on: 1) design definition & manual correction effort improvement opportunities, 2) automation performance & capability improvements, 3) intrinsic trade-offs between survey requirements and an ideal conversational experience, and 4) augmentation tasks expansion as well as support for prototyping and tailoring.

6.5.1 Design Definition Improvement Opportunities

Evaluation results highlighted several opportunities for design improvements to the empathetic reactions specifically, but also to broader aspects.

Improvements to Empathetic Reactions: Empathetic reaction matching suffers from 3 major issues: 1) lack of appropriate reaction class for specific scenarios, 2) insufficient use of broader context, and 3) lack of specificity to survey contents. In relation to the first point, just 3 empathetic reaction categories might not be sufficient for some contexts. Prior work points to the richness of empathy expressions [243, 214, 215]. Some limitations are apparent in case an answer option is ambiguous (e.g., *"Yes but I don't always stick to it."*) or context implies a reaction beyond simple empathizing (e.g., Q: *"Do you want help with school or training?"*, A: *"Yes"*). These examples are, however, rare in the dataset and hence pose a challenge for automated matching. Secondly, proper matching of reactions may rely on broader context than just question and answer. Even the labelling itself implicitly incorporated the broader context, with demographics questions about age, gender, ethnicity and education being labeled as neutrally framed to ensure 'Neutral acknowledgments' will be matched as reactions. The situation becomes more difficult when questions on the otherwise

neutral topic, such as employment, are asked in the sensitive survey context, such as social needs (e.g., Q: *"Which of the following describes your employment situation right now?"*, A: *"Unemployed - looking for work"*) in which case an empathetic reaction might be appropriate. Thirdly, reactions could benefit from directly incorporating user answers (e.g., *"Thanks for saying yes"*) or specific mention of the question content (e.g., *"Thanks for letting me know about your housing situation"*). Directly referencing user input is in line with indications from prior work [78, 242] and also aligns with findings that users prefer sophisticated choices of words, as well as well-constructed and long sentences [154, 224].

Rushing Towards Completion, Limited Personal Feel & Language Verbosity: Other challenges could be grouped into: 1) issues with rushing towards completion 2) need for more personal interaction, and 3) verbosity of question rephrasing. In relation to progress repetitiveness and rushing towards completion, the informational part (i.e., what question is user at) could always be paired with interaction encouragement. For more personal interaction, use of person's name, sharing bot 'background', and improved politeness suggested in user feedback are in-line with relational agent design and could easily be included [28, 131]. Finally, the verbosity of question rephrasing is a by-product of addition rather than removal of contents. This is to avoid question text modification and also a technical limitation of 'deep' rewriting [133, 241]. Lengthening of the interaction and more reading effort can be detrimental to some users as shown in my prior work [129].

6.5.2 Correction Effort Reduction

More than 70% of the correction effort was due to question empathy framing misclassification alone. Such misclassification almost certainly invalidates all the empathetic reactions applied to all the answer options for a question. As shown on an actual example from the Sleep Quality survey in Figure 6.8-left, misclassifying a question *"Next, could you say that you fall into a deep sleep?"* as Negatively framed results in answer *"Rarely: None or 1-3 times a month"* being matched with an incorrect reaction: *"Sounds good"*. This is the case for all

the reactions even if all the individual answer options are classified correctly. Manual effort of rewriting all these reactions amounts to 66 character edits (rewriting to the reactions in Figure 6.8-right is assumed). A simple editing tool support, which would allow a survey administrator to correct the empathy framing classification via GUI drop-box could reduce such effort to just 2 clicks and also further provide the training data for improving future classification accuracy (see Figure 6.8-right).

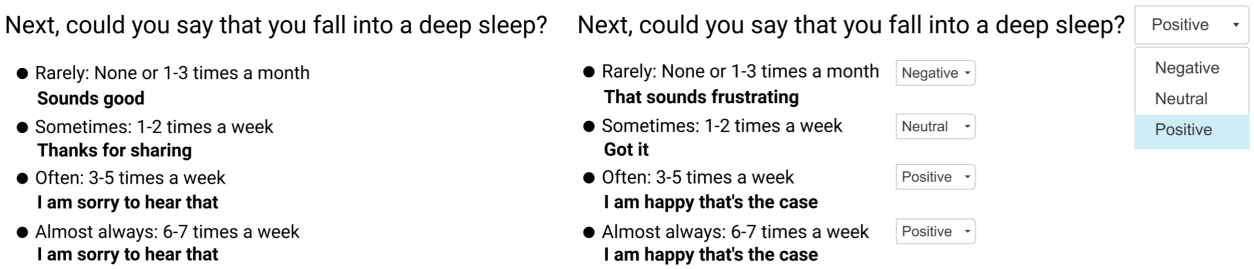

Figure 6.8: Correcting reactions - Left: manual correction of question misclassification results in the need of rewriting all the reactions (a cost of 66 character edits). Right: with GUI support from an editing tool, the correction could involve just re-labeling the question (a cost of 2 mouse clicks).

Similar magnitude in reduction of correction effort applies to scenarios when a particular misclassified answer option is frequently reused throughout the survey. This happened in the Sleep Quality survey when a misclassified answer option *"Sometimes: 1-2 times a week"* was used in all the 18 questions leading to a need for manual correction of all associated reactions (213 character edits representing close to 32% of all the manual correction effort for this survey). With a tool support such effort could be reduced to just 2 clicks as correction in one answer could be propagated to all the answers used in the survey.

6.5.3 Automation Performance and Capability Improvements

To enable ML automation with limited domain-specific training data I intentionally reduced the adaptation problem to a series of simple text classification tasks. Employed classical ML algorithms, with some custom feature tuning, are in line with performance that could

be achieved using some off-the-shelf end-user tools such as LUIS[4] or MonkeyLearn[5]. The automation can further be improved by 1) increasing the accuracy within the current problem framing (i.e., keeping tasks as text classification problems) or by 2) relaxing the problem beyond this definition (e.g., unconstrained text generation) to enable modeling more complex relations. I discuss these two ideas further.

Accuracy Improvements: Higher accuracy can be achieved by: 1) collecting more labelled data and 2) use of more capable algorithms. The current dataset (including development and hold-out) contains 22 surveys with 357 question examples labelled for phrasing and empathy framing, as well as 812 answer examples labelled for empathy framing. More survey examples could be automatically scraped from various sources such as SurveyMonkey templates[6] and SurveyPro templates[7]. These repositories, while convenient for scraping, contain only informal surveys. Validated tools included in research papers and white papers are harder to obtain at scale. The text classification simplification I used would make the labeling task easy to automate via crowd-sourcing approach. Beyond and in addition to obtaining more labelled data, more capable text classification algorithms can be employed. Unfortunately a more capable algorithm in presence of limited data is unlikely to yield better results [99] and can be prone to overfitting. A promising solution to this limitation is a domain adaptation approach, where a very capable language model comes already pre-trained on vast amounts of general language data [34]. An initial exploration of such approach using pre-trained bert model[8] fine-tuned on my limited dataset shows promising results increasing reaction matching accuracy from $0.57 \pm .15$ to $0.68 \pm .11$ and question language adaptation from $0.68 \pm .33$ to $0.81 \pm .24$ in a leave-one-out setup (see Table 6.10). Going forward, models

[4]LUIS - a machine learning-based service to build natural language into application - https://www.luis.ai/

[5]MonkeyLearn - machine learning service for designers and developers - https://monkeylearn.com/

[6]https://www.surveymonkey.com/mp/university-student-satisfaction-survey-template/

[7]https://www.questionpro.com/survey-templates/

[8]bert-base-uncased from Huggingface library: https://huggingface.co/transformers

pre-trained on sentiment analysis can be used as a basis for empathy framing classification [20] and survey domain detection can leverage topic models [245]. An empathetic reaction generation model from a mental health domain could also potentially be adapted [170].

	Leave-one-out		5-fold CV	
	Classic ML	**BERT***	**Classic ML**	**BERT***
Empathetic Reaction†	0.57±.15	0.68±.11	0.73±.05	0.83±.07
Question Empathy Framing	0.69±.24	0.78±.14	0.81±.03	0.85±.07
Answer Empathy Framing	0.89±.13	0.92±.09	0.89±.04	0.89±.01
Question Language Adaptation	0.68±.33	0.81±.24	0.83±.07	0.89±.04

† - Empathetic Reaction is derived from the Question Empathy and Answer Empathy Framing using a fixed rule.
** - Bert base uncased model, lr: $2e^{-5}$, eps: $1e^{-8}$, trained for 5 epochs*

Table 6.10: Comparison of accuracy for a classical ML model used and a pre-trained deep learning model fine-tuned on the task dataset

Capability Improvements: Improved automation capability (e.g., being able to include broader context or even generate reactions from scratch word by word) can be achieved by re-framing and relaxing the problem definition in various ways. For example, instead of separately classifying the question and answer for empathy framing and then selecting an appropriate empathetic reaction using a fixed rule, the classification could used combined features to directly select the best empathetic reaction class. Initial exploration of such joined classification showed an improvement in reaction classification accuracy from $0.57 \pm .15$ to $0.70 \pm .29$ in a leave-one-out evaluation. Additionally reaction selection context can include information about the survey domain in some form. Further expanded context, would unfortunately also require more training data and may not accommodate GUI supported corrections described in the previous section equally well. In the most unconstrained fashion the conversational question rephrasing could be seen as a translation task (similar to language translation [68]). Existing survey items would be treated as text in one language that needs

to be translated to a "conversational" language. Similarly empathetic reactions could be treated as any chat utterances generation problem [213]. Both approaches would not impose any design constraints, but would require large amounts of domain specific data and would likely provide less control over the output [233].

6.5.4 Intrinsic Challenges of Conversational Survey Adaptation

Several challenges seem to be particularly difficult due to the conflict with surveying practices, or because the type of data needed to address them is unlikely to ever be available.

Rephrasing 2nd and 1st Person Survey Items: Current design tries to make sure that all conversational utterances are questions in 3rd person form. This choice is, however, not the only possibility. In some cases survey questions are intentionally written in 1st or 2nd person to facilitate more honest answers [108]. In other cases the questions might have been tested in the exact form presented and changes could invalidate the survey. From a design perspective, it might make sense to provide alternative ways of rephrasing such questions, e.g., question *"I feel in control of my current financial situation."* could be rephrased as *"Is it fair to say that you feel in control of your current financial situation?"*, but also as *"How would you respond to the following statement: 'I feel in control of my current financial situation.'?"* This challenge requires further research.

Addressing Survey Intrinsic Question Repetition: Repetitiveness of conversational questions has been reported by some users. The design approach tried to minimize that by attaching dynamically varying conversational prefixes, but some repetitiveness is, however, likely intrinsic and intentional especially in validated surveys [108]. It is possible that users tolerate the repetition more in form-based surveys than in conversation where it is not 'natural'. It is also possible that the increased engagement and attention with conversational administration [124] is making users notice repetitions more. It seems that two approaches could be possible: 1) either a new standard for obtaining validity could be defined (as

suggested in [124]), or 2) chatbot could manage user expectations by explicitly announcing in the introduction, e.g., *"I will ask you some questions that might seem repetitive, but this is intentional to make sure we have a common understanding of the topic"* or preface such a similar question with *"I know I asked you a similar question before, but..."*.

Conversational Form of Tabular Questions: Several surveys organize questions in a tabular form to help streamline answering especially when all the questions are answered on a common scale (e.g., likert). A good example is Big 5 personality survey which in the PDF format is composed of a prefix phrase: *"I see myself as someone who..."* and then each table row refers to a specific concept, e.g., *"...is reserved"*, *"...is generally trusting"* answered on a common 5-point likert scale from *"Disagree strongly"* to *"Agree strongly"*. The current adaptation approach concatenates the prefix and each concept to form a separate survey item, e.g., *"I see myself as someone who is reserved"*, *"I see myself as someone who is generally trusting"*. Although this is an approach taken in prior work using manual adaptation [124], it may introduce additional repetition, even after the diversified conversational prefix is appended, rendering: *"Moving on, is it fair to say that you see yourself as someone who is reserved?"*, *"Do you think it's fair to say that you see yourself as someone who tends to be lazy?"* A different approach might be to support such questions with a rich GUI element rendered directly in the chat window. Rich GUI elements are suggested in prior work [127].

Matching Socialization and Empathy to User Characteristics: Some users did not seem to appreciate empathetic reactions in conversational surveys in general, indicating that *"There's no need for too much empathy, acknowledging my replies is enough"* (P30). This echoes general findings from prior work reporting that some users do not react well or expect socialization in chat [147]. Similarly some user groups may appreciate a different style of conversational adaptation (e.g., casual style for teenagers as in [124] or formal style for older audiences [129]). This is in principle supported through meta-parameters and a repository of phrases (I will discuss this in subsequent section), but an ability to automatically tailor

the style to an individual would require user specific data that might not be available.

6.5.5 Additional Augmentation Tasks; Prototyping & Tailoring Support

An automation approach I proposed in this work provides an opportunity to not only lower the conversational design effort, but also easily supports additional extensions and can facilitate research and practice of designing conversational interactions.

Expanding the Set of Augmentation Tasks: The initial exploration performed in this work limited the conversational adaptation aspects to a set of key tasks, which can arguably be seen as representing a minimal set of needed adaptations. In other chapters I have shown that different conversational design aspects can offer a tailored experience to serve the needs of different populations. The automation framework I proposed can easily be extended by such additional components. The text-to-speech generation can be added to provide voice capabilities as in Chapter 5 using commercial tools [9]. Similarly language paraphrasing for understandability employed in Chapter 5, could be integrated as an additional adaptation task using off-the-shelf models [248, 160]. Additional augmentation tasks such as adding a domain matching chat avatar icon could be supported as well by leveraging existing work on matching text and images [231]. These could be added as modules that can be turned on or off as needed to easily render conversational interactions with different properties.

Leveraging Repository for Altering Chatbot Language Personality: In a similar fashion more substantial changes to the language can be supported by leveraging the provided repository-based approach. The ML models used to retrieve phrases from the repository rely on survey content and general augmentation categories, but not on the concrete text of the repository utterances. What this means is that a different set of augmentation phrases can be provided without the need to retrain any of the models. This can be used to render a different chatbot language personality. Instead of current professional and polite style (i.e.,

[9] https://aws.amazon.com/polly/

"Hi, my name is {name}. I would like to talk to you about {topic}.") a different repository could be linked with e.g., informal teenage introduction style, such as *"Hey, awesome to meet you, I am {name}, let's chat about {topic}"*, and similar expressions of compassion (e.g., *"Wow, that really sucks :("*), and satisfaction (e.g., *"That's really awesome!"*). The augmentation categories provided by the current repository inform the types of phrases that need to be provided to render a consistently different language personality.

Support for Prototyping & Tailoring: Taken together, the ability to turn on and off different augmentation aspects as well as the ability to easily replace the chatbot language without the need for retraining any of the ML models offers several opportunities. It enables quick prototyping of different variations of conversational interactions for the purpose of design exploration. It can be used for quick A/B testing of different designs with different user populations to help answer some of the persistent questions in conversational design related to the level of socialization [146] and chatbot personality [41]. Fully automated approach helps ensure consistency of comparisons in such use cases. Finally, a detailed parametric control over augmentation aspect provides a step towards automated tailoring of conversational interactions to different populations in case user profiles are available.

6.6 Summary of Contribution

In this chapter I explored the feasibility of supporting the design of engaging conversational survey administration with automation in order to automate the design process, I had to perform manually in Chapter 5. To do that, I built up on the common linguistic and dialogue aspects of an engaging conversational interaction I identified in my prior work such as content diversification, conversational language style, contextual reactions, social dialogue & empathy, as well as conversational etiquette. I defined the conversational adaptation of a survey as composed of 4 tasks: 1) addition of introduction & closing, 2) addition of contextual empathetic reactions, 3) addition of progress communication & topic handling, 4) adaptation of question language to conversational style. This adaptation is also guided by 4 design

principles: 1) avoiding repetitiveness, 2) minimizing changes to the original survey items, 3) contextual use of empathy, 4) audience sensitive augmentation. The tasks rely on retrieving phrases from a reusable repository of 118 conversational augmentation phrases informed by prior work [129, 78, 239, 125] and linguistic resources [86]. I further employed data-driven machine learning (ML) techniques to enable the automation to learn from example survey data. In an evaluation with 30 respondents from a crowd-sourcing platform I show that the proposed approach can produce engaging conversational surveys (comparable to my manual design used in Chapter 5) with only 30% additional manual correction effort (as opposed to 100% effort that had to be put to make the adaptation from scratch). I also discuss the remaining issues and further improvements related to 1) design definition improvements, 2) user manual correction effort reduction opportunities, 3) automation performance & capability improvements, and 4) intrinsic challenges related to trade-offs between survey requirements and an ideal conversational experience. My work contributes to the understanding of what it means that the survey is conversational. It systematizes and automates the steps needed to make any survey conversational, advancing prior work which relied on one-time manual redesigns. I also highlight some of the intrinsic trade-offs between survey administration requirements (i.e., dictated by validity) and an engaging conversational experience. Outcomes of my work can directly support survey administrators, particularly without design background, in creating more engaging data collection experiences for their respondents with less effort. My work can also help engineers of data-driven conversational systems train their approaches with less data by leveraging survey domain knowledge insights I provided.

Chapter 7
DISCUSSION

Technology has been used to support health & behavior change applications, but despite offering valuable support for automating various aspects [165, 93], it generally struggled with supporting user engagement [45, 82]. Conversational agents, whose recently resurgent popularity has been fueled by advancements in technology, enable rethinking the support technology can offer in this context to engage and motivate users. Unique human-likeness aspects of conversational agents, can make them more engaging [27, 239], and coupled with the ease of incorporating additional application specific features (e.g., use of voice, adaptive behavior) make them valuable for supporting various user groups (e.g., low literacy) in different contexts (e.g., workplace) and for otherwise challenging purposes (e.g. reflection on behavior).

In this dissertation, I demonstrated how conversational interfaces can be designed to improve user engagement in the key health & behavior change challenges of activity promotion, learning form past behaviors (reflection), and data collection. My work further expands our understanding of user perceptions of such systems, as well as their strengths & weaknesses. I propose concrete design & implementation artifacts, as well as, several well documented reusable design processes for reproducing the proposed designs in different contexts. Finally, I also address the substantial effort required to design an engaging conversational experience, by exploring the use of automation to support non-designers in applying the findings of my work in their applications with ease. Through these findings I have demonstrated the claims made in my thesis statement. Here I summarize the key benefits and challenges conversational design can offer especially in the health & behavior change domain which I have identified in my work.

7.1 Benefits of Conversational Design in Health & Behavior Change

Across the conversational applications I have designed, implemented & evaluated in this work I identified several common benefits a well designed conversational interaction can offer. Through the exploration of these benefits I have confirmed my thesis statement showing that conversational interactions can be designed to support user engagement in health & behavior change applications.

7.1.1 Engagement in Interaction

Across the studies users consistently reported the benefits of increased engagement when interacting with conversational systems I designed. In Chapter 3 users reported engagement with conversational prompts' contents specifically crediting the *personalization* and *content diversification* aspects. Users reported increased attention (an engagement dimension [180]), informational value, and personal relevance. All of these factors led them to be more engaged in interaction with conversationally redesigned prompts as compared to the baseline non-conversational ones. Similar engagement was reported in Chapter 4 in the context of workspace reflection with *Robota*. Users reported being more engaged due to the perceived benefits such as increased awareness of work tasks, improvements in work organization & productivity, as well as new perspectives and understanding of their high-level career goals. In this context the engagement mechanisms based on perceived benefits of interaction echoes a dimension from relationship investment model [201]. Several participants also felt more engaged with the voice finding it fun and enjoying the 'surprise' aspects of changing voice interaction. Some of this might be associated with novelty. Field study with *Reflection Companion* in Chapter 4 provided a strong behavioral evidence of user engagement, when 11 of the 33 participants elected to use the system for additional 2 weeks without any compensation (this is on top of 2-week paid study period). They also used the system very actively putting the effort to type in free-text replies to 83% of the daily dialogues they received during this time. In Chapter 5 direct measurements with engagement survey from

O'Brien's engagement scale [180] revealed conversational engagement benefits for a vulnerable user group (low health literacy) in a particularly sensitive data collection setting. I complimented these findings with evidence of high engagement on the same scale across broader survey-based conversational data collection with general population in Chapter 6.

Engagement benefits or potential for such benefits has been reported by various studies, especially from Bickmore er al., using embodied conversational agents in medical health domain [30]. My work further expands on these findings in the context of health & well-being with non-embodied more cost-effective chat like interactions. Furthermore several studies reported the issues of interaction repetitiveness leading users to lose motivation to continue using the agent and follow recommendations [28]. Similar effect have been reported in computer counseling where the approach itself can be effective if users stay with the system, but high drop-out rates limit the positive impact [190]. My work contributes an important content diversification techniques that help keep users engaged. I also offer detailed understanding of the mechanisms leading to increased engagement in the novel context of reflection.

7.1.2 Motivation to Perform Activity

Across several studies users reported increased motivation to perform activities directly or indirectly promoted by the conversational agent. In Chapter 3 the diversified conversational prompts were promoting physical exercise challenges 4 times a day. I have shown through quantitative analysis that this significantly increased user activity, making users 3.7 times more likely to exercise in the 2 week study period as compared to baseline. In Chapter 4 although *Reflection Companion* did not directly promote activity, several user had their own goals related to being more active. These users reported that through conversational reflection they gained increased motivation thanks to the sense of accountability to the agent, and improved understanding of their behavior barriers. They also discovered small and concrete attainable steps, and were able to construct more well thought-out action plans. Several users directly reported new behaviors, usually small changes in routine or returning to posi-

tive past behaviors that have been abandoned. Users also reported increased additional side activities facilitating physical activity, such as wearing their fitness tracker more consistently or scheduling classes at a gym. In the workspace setting in Chapter 4 reflection with *Robota* also indirectly increased motivation and productivity. Users reported that awareness of limited progress they felt they have made, led them to try to be more productive. They reported being worried they will have nothing to write in their communication with the agent at the end of the day (this echoes accountability from physical activity setting). Users also reported that the interaction helped with the side tasks facilitating work. It specifically helped with composing reports due to the ability to quickly recall things and the fact that the interaction logs served as a source of concrete information. Several users also reported that work reflection helped them with organizing daily activity due to facilitating planning and making sure important things are not forgotten.

Prior work reported persuasive capabilities of conversational agents [29, 37, 207]. These works usually relied on direct persuasion techniques, which could be effective, but can also lead to undesired side effect such as reactance [238] (where user pressed with excessive persuasion acts the opposite) and with time have been shown to suffer from high drop-out rates [28]. My work shows an important indirect route to motivation and activity promotion through conversational reflection. Further an ability to indirectly promote an activity I introduced in Chapter 3 is valuable for sensitive settings where mentioning the topic of persuasion directly can be especially detrimental (e.g., antismoking campaigns talking about smoking may actually induce more smoking from smokers [95]). Finally, my work shows that some similar benefits to human counseling sessions, can be achieved with well designed conversational agents which can lower the cost and fill the gap in support for some users. My work also rises some ethical questions about the impact agents could have on user stress at work, which should be investigated further.

7.1.3 Accessibility: Familiarity & Understandability

Throughout my work I have demonstrated several accessibility benefits of conversational interaction. In Chapter 3 and Chapter 4 with *Reflection Companion* I used mobile text communication (SMS/MMS), which had the benefit of *familiarity* and *ease of use*. Users specifically reported the lower barrier to start using the system by not having to install additional applications on their mobile devices and not having to learn any new interface. These benefits have been reported in prior work [85]. In Chapter 4 on an example of workspace reflection with *Robota* I used company internal Slack chat-like platform which was also easy to use and familiar. In this work I also used voice-assistant-like interaction with an Alexa-enabled personal mobile device to support reflection. For the voice part users specifically commented on the ease and speed of answering questions using voice. They also praised an ability to quickly capture some points or thoughts with voice. Voice recording was also considered easier for non-native speakers than writing. Finally in Chapter 5 I augmented the standard chat interface with voice readout (which was also available on demand) and an ability to rephrase the question asked by the agent. These features were specifically praised by the low health literacy users who directly reported that the system facilitated their *understanding*. These features also occasionally helped high literacy users who had vision problems or felt fatigued and found audio feature lowering their interaction effort.

In general the accessibility benefits of conversational interaction I explored can be categorized as *familiarity & ease of use* and *understandability* related. The first type of benefits have been indicated in prior work [13] praising frictionless and natural interaction that can replace mobile apps [127]. Naturally not all interactions are easier via mobile text and few reviews pointed to the difficulty of properly designing mobile text prompts to exploit the benefits and avoid the challenges [80, 102]. My work offers a confirmation of the familiarity benefits of chat-based interaction in new settings, particularly reflection. Prior work reported the potential for audio to improve the understandability among low literacy users in medical domain [96]. My work expands on these findings and improves our understanding

of the benefits of the use of audio for sensitive data collection in hospital setting. Further I contribute the conversational question rephrasing features as another understandability enhancing aspect.

7.1.4 Comfort & Sharing

In several applications users were willing to share and disclose personal information to the agent, even if not directly asked to do so. In interactions with *Reflection Companion* in Chapter 4 users shared various personal aspects with the agent in their daily mini-dialogues. They often described their personal and work plans, activities, and detailed schedules. They shared many personal aspect such as their relationships with friends and family. Often also freely sharing their emotional states 'telling' *Reflection Companion* they feel stressed, lazy, annoyed or even jealous of the physical fitness of their friends. Several users also felt comfortable to use the interactions for venting. It is worth nothing that the system did not directly asked about any of these personal aspects. Surprisingly some similar sharing also took place in the semi-public workspace setting with *Robota* in Chapter 4. This was especially the case with the dedicated voice channel, where several users felt the interaction was more personal and they reported feeling like *Robota* 'cares about them'. This feeling also made some users consider *Robota* more as a counselor or therapist to whom they can vent when they are unhappy and want to complain. While in reflection setting the agents never asked about specific personal or sensitive details, my work in Chapter 5 on social needs screening involved highly sensitive direct questions about 'sexual abuse', violence', and 'extreme poverty' asked to a vulnerable population. The agent also featured empathetic design. In this setting users' willingness to share was somewhat inconclusive with some users reporting that human-likeness features made them more comfortable to share, while others feeling just the opposite. Still even in this highly sensitive setting many participants described the agent as 'caring', 'helpful' and 'concerned'. Several also reported that empathetic aspects were 'calming' and gave then 'confidence'. These findings were echoed in Chapter 6 where I applied conversational empathy design to broader set of surveys, which revealed that impact

of empathy design on user comfort is highly contextual and possibly user specific.

These indications support the strong potential of the conversational approach to establish trust & encourage sharing. Prior work indicated such potential in controlled lab experiments with embodied agents [59], in job interview lab setting [145], as well as in stress relief in crowd-sourced setting [204]. My work build up on these findings by exploring the sharing and self disclosure in the novel reflection setting as well as in the semi-public space in a naturalistic settings of longer term field studies. I also provide novel insights into the complex mechanism in which human-likeness can affect willingness to share for different users. My findings suggest the need for understanding individual user characterises and the potential for tailoring empathy features. It also highlights the important ethical issues of user self disclosure which should be considered.

7.1.5 Guidance

Conversational interaction offers a natural support for sequential interaction, which could be particularly beneficial in health & behavior change. In Chapter 4 I leveraged this aspect by supporting reflection on physical activity. I designed the dialogue progression to mimic the progression of structured reflection process based on a theoretical model [14]. I found that the dialogue guidance encouraged deeper thinking, more meaningful answers in reflection and also extended the time users spent reflecting. Users also reported that having a bigger overwhelming reflection task split into small more manageable pieces guided by the dialogue lowered the effort of reflection for them. In that sense dialog can be used to decompose more complex task into smaller more manageable activities following the recommendations of goal-setting theory [152]. Another benefit of dialogue guidance is helping users avoid being stuck in negative perceptions [10]. This is something I have found in the pre-study workshop, when one of the participants reported being discouraged from looking at own activity graph due to fear of low performance. In the design of *Robota* I also used the dialogue progression to support reflection on workspace productivity. In this setting the dialogue was designed to connect separate activities (i.e., journaling & reflection). The

journaling was a task beneficial for work and report writing, while reflection was meant to engage users in personally meaningful task that could benefit their professional development. Users appreciated these benefits and also liked the connection between the dialogue stages by means of mentioning the work task they scheduled. Given that activity reporting is something user might need to do regularly, the dialogue connecting such aspects can help form a habit [227].

Several theoretical conceptualizations of user behavior in health & behavior change propose processes (e.g., personal informatics [142], structured reflection [14]) or cyclical processes (e.g., stages of change [191], lived informatics model [73]) to capture user journey. The sequential aspect of the conversational interaction seems particularly well suited to support such progression on a macro and micro scales. Furthermore, taking an inspiration from human-health coaches a conversational interaction can proactively guide users in specific beneficial directions [198]. My work specifically contributes to our knowledge of how to design a dialogue-based support for such processes and also make going through the cycles novel, personalized and engaging.

7.2 Challenges of Explored Conversational Design for Health & Behavior Change

My work uncovered several challenges that a designers of conversational systems would likely have to address in their designs. It is important to note that the challenges I identified are to some extent related to the application and the implementation of conversational interaction I provided in my work. I therefore relate these challenges to the broader literature on conversational interfaces.

7.2.1 Efficiency

Several conversational interactions I designed in this work revealed somewhat lower efficiently and higher effort of interaction conversational interfaces can introduce in some situations. In Chapter 4 users interacted with *Reflection Companion* by typing-in responses to the agent on

their mobile phone. While they perceived it as valuable for their engagement, they also felt this required additional typing effort. It is worth noting that this design choice is arguably intertwined with the reflection support purpose where free expression could be particularly valuable. In interactions with *Robota*, which offered a voice-based and text-based communication, users considered it easier to read than listen to voice, especially when the questions were long or complex. At the same time typing was considered more time consuming and effortfull than providing responses with voice. This shows how voice and text modalities can influence interaction efficiency. Finally the application of *HarborBot* with high and low literacy populations in Chapter 5 provided the most insights about the efficiency challenges. This application focused on data collection and hence required the highest amount of input from the users. While *HarborBot* supported structured graphical input to lower effort; the use of voice readout, reaction delays and additional socialization utterances lowered interaction speed which was negatively perceived by the high literacy users. It is worth noting that low literacy users did not mind the lower speed as understandability benefits outweigh these shortcoming for them.

The efficiency aspect being an issue is not uncommon for conversational agents. Prior work indicated that waiting for audio readout can be less efficient [223]. Several works specifically in conversational survey data collection reported longer completion times and lower perceived efficiency compared to form-based methods [124, 239]. Perceived efficiency in the broader context of conversational interaction is an important underlying theme of many task-oriented uses of such systems [156, 112, 97]. My work, however, shows that efficiency is not a universal problem for all the populations. Furthermore due to the detailed understanding of the specific causes of perceived inefficiency uncovered by my work it could be possible to optimize waiting times and tailor the interaction (specifically use of audio) to satisfy all user groups.

7.2.2 Artificial Feel

Across all the studies users reported certain aspects of interaction that felt 'artificial'. In Chapter 3 the conversational triggers relied on topic and lexical diversification mimicking conversational diversity reported in [78], which felt more natural in general. This diversification, however, invited higher scrutiny of content. When the content started repeating after some time, almost all the users noticed that and considered it artificial. This paired with the fact that people remember the negative more than the positive [21] led to quantitatively lower rating of helpfulness than a fixed repetitive prompt from baseline. This shows that the illusion of natural conversational interaction can easily be broken. In Chapter 4 the *Reflection Companion* mini-dialogues created an expectation of an 'intelligent' and 'meaningful' follow-up to user's free-text response in the first part of the dialogue. If this did not materialize to user's satisfaction, the follow-up was reported as 'generic' and 'computerized'. Similarly, several aspects of the *HarborBot* system for social needs screening in Chapter 5 felt artificial to the users. The biggest contributor to the artificial feel in that work was the text-to-speech voice used, which was described as 'truncated' & 'monotone'. Such limitations of voice combined with the sensitive questions made some users report the *HarborBot* as 'pushy' and interaction as coming form a teacher. Secondly due to the underlying survey contents, the users felt some information was asked repeatedly even after they declined to answer. Finally in this system issues with contextual empathy matching led users to see the agent reactions as 'defaults' and the agent to feel 'fake' and reminiscent of customer support. Several of these aspect have been echoed in Chapter 6 where users also complained about questionable use of empathy in some contexts, the artificiality of being asked the same or similar question repeatedly, and the repetitive nature of the progression communication utterances.

There are a few things here to consider. First, despite these negative perceptions, the conversational systems were still largely successful in accomplishing their goals of engaging users. Secondly, several of these challenges are related to the provided implementations and

most of them can feasibly be resolved with current technology for particular applications. Some are arguably more technically challenging, such as the quality of text-to-speech [193]. Thirdly, the resolution of the issues seem to be a trade-off between quality and cost & design effort (e.g., text to speech can be replaced with human voice recordings, richer content can be crowd-sourced to avoid repetition).

7.2.3 High Expectations, Contextual & Social Intelligence

While artificial feel I described earlier relates to relatively small aspects that felt unnatural or disappointing about the interaction, users also reported more fundamental issues related to agent's contextual knowledge and the fundamental acceptance of a computerized system to act socially or emotionally. In Chapter 3 the users expected the conversational prompts promoting physical activity to be somewhat 'intelligent' and 'meaningful'. They expected the agent to be aware of their status and the prompts' contents to fit the context of their activity, location and schedule. Furthermore they expected the contents of the conversational prompts to always supply new and unexpected information they could learn from. In the workspace setting, *Robota* asked several personalized questions which mentioned user previously journaled work tasks. While this specificity was appreciated for its personal focus, the participants complained that the agent picked tasks that were not meaningful for them (e.g., routine tasks or tasks that were not challenging). Users expected the system to be aware of the specifics of their work and also capable of deciding which tasks are the most meaningful for them to reflect on (it is worth noting that we used wizard-of-oz approach for this task, which shows that this can be fundamentally challenging for a human). The *HarborBot* agent for social needs screening I described in Chapter 5 employed the use of empathy to ease users into answering sensitive questions. Aside from challenges with matching empathy to context, several users reported that even if they felt the social and empathetic utterances were well designed, they would simply not subscribe to the 'illusion' that a computer system can or should exhibit such qualities. Some of the similar findings were echoed in Chapter 6.

There are several aspects here to consider. In some cases high user expectations of the

conversational system's capabilities could arguably exceed what a human could do, hence user expectations of conversational agents may go beyond human provided assistance. High expectations have been reported in conversational agent context [156]. While rich contextual sensing is possible in principle, although challenging in practice due to possible misinterpretations [203], such sensing could raise several ethical and data privacy issues users might not take under full consideration. Finally, regarding the users' fundamental acceptance of social aspect of the agents, some past works pointed to a possible fundamental individuals' preference for socialization in agent context [147, 146]. This aspect has yet to be well explored. My work contributes specific case studies further improving our understanding of the deeper challenges involved in designing conversational agents and user varying expectations of their performance.

7.2.4 Effort of Creating Engaging Content

I explored several approaches to generating enticing contents for conversational agent interaction in this work. In Chapter 3 I used crowd-sourcing, past literature, computational semantic relatedness, and theoretical models to create content that is diverse and novel to engage users in repeated interactions. I used value profiling to make the conversations personalized. Similarly in Chapter 4 I addressed the problem of diverse domain-specific contents with workshops, past literature, and informal resources. Supporting personalization in the dialogue required content generation to incorporate user goals, work tasks and fitness tracker data. In *HarborBot* design the core dialogue data relied on predefined survey, but the additional conversational content required design in consultation with domain experts and careful empathy crafting.

These examples show that creation of engaging conversational contents requires substantial effort which has been acknowledged in prior work [97]. Part of the challenge, especially in long-term behavior change domain, is the need for creating diverse and novel dialogues to keep users engaged as I have demonstrated in Chapter 3. Another challenge is making the interaction personalized and specific to the domain of application. In Chapter 6 I separated

some of the common reusable parts of a conversational experience (e.g., acknowledgments, transitions, introduction) to lower the design effort with automation, but this offers just a first step in lowering such effort. Fully data-driven approaches are hard to apply as they require substantial dialogue data in a particular domain [88], are hard to control [110], and can still suffer from repetition and consistency issues [240, 247]. Recent developments in data-driven approaches leverage models pre-trained on large scale generic dialogue data and fine-tune them to specific domain. While these approaches are still being researched for dialogue systems [184], their successful application relies on existence of small to medium scale domain-specific dialogue data. My work offers various design-driven processes to support data generation for such approaches.

Chapter 8

LIMITATIONS

There are several limitation of my work that apply to all of the applications as well as specific to the particular contexts. First of all, while I evaluated all of the systems I developed in field deployments, which boosts their validity, the deployments have been fairly short 2-3 weeks, which makes it hard to estimate their long-term impact. Somewhat associated with the study lengths is the issue of novelty. Conversational interfaces are still fairly new and could attract additional attention due to this aspect alone. The fact that in my studies I complimented the quantitative findings with interview feedback linking the impact to particular conversational design elements mitigates this worry to some extent. Another common limitation relates to the sizes of the user groups which varied from as few as 10 (*Robota*) to 33 users (*Reflection Companion*) in field studies. Small user groups could have limited my ability to statistically detect some true effects and could have introduced an outsized impact of outliers. Finally, while the challenges I addressed are general, I tested the conversational approaches on examples of particular applications in specific setting and with limited specific user groups. This rises the worry that the results may not generalize to other user groups. This can be a worry especially in workspace reflection which took place in a particular company. Similar limitation may apply to data collection in emergency department (although it has been conducted at two sites in different cities).

Chapter 9
FUTURE WORK

The findings in this dissertation reveal several possibilities for building up on the work as well as for new avenues for future research in the use of conversational interaction for health & behavior change:

Unified conversational support for multiple stages of behavior change: In this work I have explored applying conversational design to addressing several concrete applications in health behavior change loosely aligned with the personal informatics stages of data collection, learning from data (reflection), and activity promotion & maintenance [142]. While these applications share several common challenges such as repetition, engagement, I have applied separate conversational systems to support them. Naturally a complete behavior change support could benefits from a unification of these conversational systems under one coherent conversational agent driven support.

Exploring other challenges in behavior change: Similarly, while I explored the challenges and applications mentioned above, personal informatics models identified other areas that could benefits from conversational support, such as lapsing and re-engagement [73]. Similarly the stages of change model [191] identifies the 'precontemplation' and 'contemplation' stages in the behavior change cycle. In these stages people are unaware that their behavior is problematic or produces negative consequences. Conversational approach could try to engage users in these early steps as well. A particular challenges here would be to attract user attention, provide informational value to sustain user interest, and help guide the user to 'discover' a behavioral problem.

Studying long-term impact: Behavior is a complex, difficult, long-term process. While this dissertation shows that short-term conversational support can improve user engagement, motivation and even lead to increase behaviors, there is still a need to understand how to support this longitudinally. While this need not be a study of a couple of years to understand the efficacy of technology [126], the longitudinal nature of behavior change might surface different needs and support that users have as they work towards maintaining behavior [191]. How to design for longitudinal interactions that continuously keep people engage is an open question. Especially the aspects of content diversification and novelty would need to be addressed. One possibility could be to support forming a habit of interaction with conversational agent [225]. This requires future work.

Improving content diversification for long-term: One of the main challenges I identified in Chapter 3, and repeatedly encountered in other chapters is the challenge of repetitiveness in long-term interactions. While I introduced several successful strategies to increase both the diversity of topics as well as lexical diversity of the language, the problem was never entirely solved. Ultimately the number of unique topics that the user could be presented as motivations or prompted to reflect on is finite. As I have found in Chapter 3 when users recognize repetition it has detrimental impact on engagement. Future work could explore two different approaches to address this: 1) Given a sufficiently large set of topics, people might start forgetting past discussions. There might be an optimal threshold for total topic count dependent on frequency of interaction. 2) Interaction could be designed to reuse and build up on the same topics over time, possibly incorporating information user shared in the past. Remembering information from user past responses (e.g. a shared barrier of *"not having a person to run with"*), would allow the agent to bring back such information in future interaction, e.g., *"What could you do to try to find someone to run with?"* or *"Is not having someone to run with still an issue for you?"*.

Improving contextual understanding with sensing: In several of my deployments users expected a certain level of contextual awareness of the conversational agent. In Chapter 3 users expected the motivational prompts to tailor to their activity, location and schedule. In the workspace reflection setting in Chapter 4 intelligently sensing a worker's context, recent activity, and main accomplishments will help workers derive greater meaning and insights and will likely lead to improved productivity and work satisfaction. A future promising direction could try to incorporate such contextual sensing to enhance the conversation.

Improved automation for design support: In Chapter 6 I explored an automated approach to lowering the design effort of conversational survey-based data collection. I showed the basic feasibility of automating addition of some reusable components of conversational approach: social etiquette, acknowledgments, empathy, and conversational language style. Yet as I have demonstrated in other chapters a lot of effort is required to generate domain-specific data via workshops, literature search, and crow-sourcing. Future work could explore how such effort could be lowered further with use of modern deep learning or other automation technologies.

Chapter 10
CONCLUSION

Health & Well-being is increasingly important in modern society with aging populations, obesity, mental health issues, and multitudes of other challenges. Technology has successfully supported many crucial aspects of behavior change such as automated activity logging, visual analytics of behavior data, as well as facilitating health communication with peers and health professionals. Yet it had in many cases struggled with keeping users engaged, especially over longer time periods. Conversational interactions have demonstrated the potential for supporting user engagement, motivation and well as providing various accessibility benefits for vulnerable populations in need. In this work I take advantage of the technical advancement in conversational technology and growing popularity of conversational systems to explore how they could play a role in addressing various health & behavior change challenges.

In this thesis I designed and implemented four conversational systems: *Fitness Challeneges*, *Reflection Companion*, *Robota*, *HarborBot* and a process to lower the design effort of engaging conversational data collection: *Survey Converter*. I evaluated these systems in multiple deployment studies in personal, workspace, as well as, hospital settings. I demonstrate that the conversational design can be used to successfully engage people in various aspects of health & behavior change, such as physical activity promotion, reflection on behavior, and for increasing the understandability and comfort of sharing sensitive social needs data among vulnerable populations. Using mixed methods in my studies I also take advantage of qualitative findings to understand the mechanisms in which specific aspects of conversational interactions affect users.

My work identifies and provides evidence for several benefits of the use of conversational interactions in this space pointing to engagement in interaction, improved motivation for

performing activities, accessibility benefits related to familiarity, ease of use, comfort with sharing, and an ability to guide the users in the behavior change process via dialogue. I also identify several important challenges, such as perceptions of artificiality, managing high expectations of contextual knowledge and social intelligence, as well as lower efficiency that could negatively affect the experience for some user groups. I further investigate the concrete links between conversational design elements and these benefits and challenges. My thesis demonstrates various design processes that can lower the effort of designing conversational experiences. As technology progresses conversational interactions can offer valuable support complimenting the existing automated activity tracking and the efforts of health coaches. My work offers an important contribution to our understanding of how conversational interactions can play such a beneficial role.

BIBLIOGRAPHY

[1] The measurement of communication processes : Galileo theory and method. *Contemporary Sociology*, 11:328, 1982.

[2] Johan S Abildgaard, Per Ø Saksvik, and Karina Nielsen. How to measure the intervention process? an assessment of qualitative and quantitative approaches to data collection in the process evaluation of organizational interventions. *Frontiers in Psychology*, 7:1380, 2016.

[3] W. Abrahamse, L. Steg, C. Vlek, and T. Rothengatter. A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, 25:273–291, 2005.

[4] Elena Agapie, Lucas Colusso, Sean A Munson, and Gary Hsieh. Plansourcing: Generating behavior change plans with friends and crowds. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 119–133, 2016.

[5] Lionbridge AI. 15 Best Chatbot Datasets for Machine Learning. https://lionbridge.ai/datasets/15-best-chatbot-datasets-for-machine-learning/, 2019. [Online; Retrieved September 27, 2020].

[6] I. Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50:179–211, 1991.

[7] Dolores Albarracín, Kristina Wilson, Man-pui Sally Chan, Marta Durantini, and Flor Sanchez. Action and inaction in multi-behaviour recommendations: a meta-analysis of lifestyle interventions. *Health psychology review*, 12(1):1–24, 2018.

[8] Rusul Alrubail. Scaffolding student reflections + sample questions. https://www.edutopia.org/discussion/scaffolding-student-reflections-sample-questions, 2015. [Online; Retrieved September 28, 2020].

[9] Rusul Alrubail. Scaffolding student reflections+ sample questions. *Edutopia. Retrieved January*, 8, 2018.

[10] Jessica S Ancker, Holly O Witteman, Baria Hafeez, Thierry Provencher, Mary Van de Graaf, and Esther Wei. "you get reminded you're a sick person": personal data tracking and patients with multiple chronic conditions. *Journal of medical Internet research*, 17(8):e202, 2015.

[11] Otto Antikainen et al. Effective chatbot conversations: Experiments with bot identity and tone of voice. 2020.

[12] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[13] Julie A Ask, Michael Facemire, Andrew Hogan, and HB Conversations. The state of chatbots. *Forrester. com report*, 20, 2016.

[14] Sue Atkins and Kathy Murphy. Reflection: a review of the literature. *Journal of advanced nursing*, 18(8):1188–1192, 1993.

[15] Jeremy N Bailenson and Nick Yee. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science*, 16(10):814–819, 2005.

[16] John D Bain, Roy Ballantyne, Jan Packer, and Colleen Mills. Using journal writing to enhance student teachers' reflectivity during field experience placements. *Teachers and Teaching*, 5(1):51–73, 1999.

[17] A. Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84 2:191–215, 1977.

[18] Aaron Bangor, Philip T Kortum, and James T Miller. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, 24(6):574–594, 2008.

[19] Yehuda Baruch. Career systems in transition. *Personnel review*, 2003.

[20] Enkhbold Bataa and Joshua Wu. An investigation of transfer learning-based sentiment analysis in japanese. *arXiv preprint arXiv:1905.09642*, 2019.

[21] R. Baumeister, E. Bratslavsky, C. Finkenauer, and K. Vohs. Bad is stronger than good. *Review of General Psychology*, 5:323 – 370, 2001.

[22] Eric Baumer, Vera D. Khovanskaya, M. Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and G. Gay. Reviewing reflection: on the use of reflection in interactive system design. *Proceedings of the 2014 conference on Designing interactive systems*, 2014.

[23] Eric PS Baumer. Reflective informatics: conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 585–594, 2015.

[24] Austin Beattie, Autumn P Edwards, and Chad Edwards. A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication. *Communication Studies*, 71(3):409–427, 2020.

[25] E. Bessarabova, Edward L. Fink, and M. Turner. Reactance, restoration, and cognitive structure: Comparative statics. *Human Communication Research*, 39:339–364, 2013.

[26] T. Bickmore, A. Gruber, and Rosalind W. Picard. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient education and counseling*, 59 1:21–30, 2005.

[27] T. Bickmore, Daniel Mauer, F. Crespo, and T. Brown. Persuasion, task interruption and health regimen adherence. In *PERSUASIVE*, 2007.

[28] T. Bickmore and Rosalind W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Hum. Interact.*, 12:293–327, 2005.

[29] Timothy Bickmore and Toni Giorgino. Health dialog systems for patients and consumers. *Journal of biomedical informatics*, 39(5):556–571, 2006.

[30] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence*, 24(6):648–666, 2010.

[31] Timothy W Bickmore, Laura M Pfeifer, Donna Byron, Shaula Forsythe, Lori E Henault, Brian W Jack, Rebecca Silliman, and Michael K Paasche-Orlow. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. *Journal of health communication*, 15(S2):197–210, 2010.

[32] A. Bowling. Mode of questionnaire administration can have serious effects on data quality. *Journal of public health*, 27 3:281–91, 2005.

[33] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

[34] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[35] John T Cacioppo and Richard E Petty. Effects of message repetition and position on cognitive response, recall, and persuasion. *Journal of personality and Social Psychology*, 37(1):97, 1979.

[36] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2019.

[37] Valentina Carfora, Francesca Di Massimo, Rebecca Rastelli, Patrizia Catellani, and Marco Piastra. Dialogue management in conversational agents through psychology of persuasion and machine learning. *Multimedia Tools and Applications*, 79(47):35949–35971, 2020.

[38] Justine Cassell. Embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78, 2000.

[39] Justine Cassell and Kristinn R Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538, 1999.

[40] Pew Research Center. Demographics of mobile device ownership and adoption in the united states. https://www.pewresearch.org/internet/fact-sheet/mobile/, 2019. [Online; Retrieved September 29, 2020].

[41] Ana Paula Chaves and Marco Aurelio Gerosa. How should my chatbot interact? a survey on human-chatbot interaction design. *arXiv preprint arXiv:1904.02743*, 2019.

[42] Jilin Chen, Gary Hsieh, Jalal U Mahmud, and Jeffrey Nichols. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 405–414, 2014.

[43] Yukina Chen. *The Effects of Question Customization on the Quality of an Open-Ended Question*. Nebraska Department of Education, Data, Research, and Evaluation, 2017.

[44] H. Chiu, Nadia Batara, R. Stenstrom, L. Carley, C. Jones, L. Cuthbertson, and E. Grafstein. Feasibility of using emergency department patient experience surveys as a proxy for equity of care. *Patient Experience Journal*, 1:78–86, 2014.

[45] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A Kientz. Sleeptight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 121–132, 2015.

[46] Chia-Fang Chung, Elena Agapie, Jessica Schroeder, Sonali Mishra, James Fogarty, and Sean A Munson. When personal tracking becomes social: Examining the use of instagram for healthy eating. In *Proceedings of the 2017 CHI Conference on human factors in computing systems*, pages 1674–1687, 2017.

[47] Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine Zia, James Fogarty, Julie A Kientz, and Sean A Munson. Boundary negotiating artifacts in personal informatics: patient-provider collaboration with patient-generated data. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 770–786, 2016.

[48] Chia-Fang Chung, N. Jensen, Irina A. Shklovski, and S. Munson. Finding the right fit: Understanding health tracking in workplace wellness programs. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[49] 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.

[50] Andrew Clarke and Robert Steele. A smartphone-based system for population-scale anonymized public health data collection and intervention. In *2014 47th Hawaii International Conference on System Sciences*, pages 2908–2917. IEEE, 2014.

[51] Heather L Coley, Rajani S Sadasivam, Jessica H Williams, Julie E Volkman, Yu-Mei Schoenberger, Connie L Kohler, Heather Sobko, Midge N Ray, Jeroan J Allison, Daniel E Ford, et al. Crowdsourced peer-versus expert-written smoking-cessation messages. *American journal of preventive medicine*, 45(5):543–550, 2013.

[52] Mark Conner and Paul Norman. Health behaviour: Current issues and challenges, 2017.

[53] Sunny Consolvo, Predrag Klasnja, David W McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A Landay. Flowers or a robot army? encouraging awareness & activity with personal, mobile displays. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 54–63, 2008.

[54] Sunny Consolvo, Predrag V. Klasnja, D. W. McDonald, and James A. Landay. Designing for healthy lifestyles: Design considerations for mobile technologies to encourage consumer health and wellness. *Found. Trends Hum. Comput. Interact.*, 6:167–315, 2014.

[55] Sunny Consolvo, D. W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, B. Harrison, Predrag V. Klasnja, A. LaMarca, L. LeGrand, R. Libby, I. Smith, and James A. Landay. Activity sensing in the wild: a field trial of ubifit garden. In *CHI*, 2008.

[56] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2382–2393, 2017.

[57] R. Davis, R. Campbell, Z. Hildon, L. Hobbs, and S. Michie. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health Psychology Review*, 9:323 – 344, 2015.

[58] Terry C Davis, Sandra W Long, Robert H Jackson, EJ Mayeaux, Ronald B George, Peggy W Murphy, and Michael A Crouch. Rapid estimate of adult literacy in medicine: a shortened screening instrument. *Family medicine*, 25(6):391–395, 1993.

[59] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3):331, 2016.

[60] Xuefei Nancy Deng and Kshiti D Joshi. Why individuals participate in micro-task crowdsourcing work environment: Revealing crowdworkers' perceptions. *Journal of the Association for Information Systems*, 17(10):3, 2016.

[61] Laura Dennison, Leanne Morrison, Gemma Conway, and Lucy Yardley. Opportunities and challenges for smartphone applications in supporting health behavior change: qualitative study. *Journal of medical Internet research*, 15(4):e86, 2013.

[62] Beant Dhillon, Rafal Kocielnik, Ioannis Politis, Marc Swerts, and Dalila Szostak. Culture and facial expressions: A case study with a speech interface. In *IFIP Conference on Human-Computer Interaction*, pages 392–404. Springer, 2011.

[63] Giada Di Stefano, Francesca Gino, Gary P Pisano, and Bradley Staats. *Learning by thinking: Overcoming the bias for action through reflection.* Harvard Business School Cambridge, MA, USA, 2015.

[64] Giada Di Stefano, Francesca Gino, Gary P Pisano, Bradley Staats, and Giada Di-Stefano. *Learning by thinking: How reflection aids performance.* Harvard Business School Boston, MA, 2014.

[65] A. Dijkstra. The persuasive effects of personalization through: name mentioning in a smoking cessation message. *User Modeling and User-Adapted Interaction*, 24:393–411, 2014.

[66] J. Dillard and L. Shen. On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, 72:144 – 168, 2005.

[67] Leslie D. Dinauer and Edward L. Fink. Interattitude structure and attitude dynamics a comparison of the hierarchical and galileo spatial-linkage models. *Human Communication Research*, 31:1–32, 2005.

[68] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, 2015.

[69] Mateusz Dubiel, Alessandra Cervone, and Giuseppe Riccardi. Inquisitive mind: A conversational news companion. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–3, 2019.

[70] William Ebben and Laura Brudzynski. Motivations and barriers to exercise among college students. *Journal of Exercise Physiology Online*, 11(5), 2008.

[71] Ofer Egozi, S. Markovitch, and Evgeniy Gabrilovich. Concept-based information retrieval using explicit semantic analysis. *ACM Trans. Inf. Syst.*, 29:8:1–8:34, 2011.

[72] Daniel A Epstein, Felicia Cordeiro, James Fogarty, Gary Hsieh, and Sean A Munson. Crumbs: lightweight daily food challenges to promote engagement and mindfulness. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5632–5644, 2016.

[73] Daniel A Epstein, An Ping, James Fogarty, and Sean A Munson. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 731–742, 2015.

[74] Paul Falcone. *96 great interview questions to ask before you hire*. Amacom, 2018.

[75] Hao Fang, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mari Ostendorf, Yejin Choi, and Noah A Smith. Sounding board–university of washington's alexa prize submission. *Alexa prize proceedings*, 2017.

[76] Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A Smith, and Mari Ostendorf. Sounding board: A user-centric and content-driven social chatbot. *arXiv preprint arXiv:1804.10202*, 2018.

[77] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, 132:138–161, 2019.

[78] Jasper Feine, Stefan Morana, and Alexander Maedche. A chatbot response generation system. In *Proceedings of the Conference on Mensch und Computer*, pages 333–341, 2020.

[79] B. Fjeldsoe, A. Marshall, and Y. Miller. Behavior change interventions delivered by mobile telephone short-message service. *American journal of preventive medicine*, 36 2:165–73, 2009.

[80] Brianna S Fjeldsoe, Alison L Marshall, and Yvette D Miller. Behavior change interventions delivered by mobile telephone short-message service. *American journal of preventive medicine*, 36(2):165–173, 2009.

[81] J. Flavell. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34:906–911, 1979.

[82] R. Fleck and G. Fitzpatrick. Reflecting on reflection: framing a design landscape. In *OZCHI '10*, 2010.

[83] James Fogarty. Code and contribution in interactive systems research. In *Workshop HCITools: Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits at CHI*, 2017.

[84] B. J. Fogg. A behavior model for persuasive design. In *Persuasive '09*, 2009.

[85] Asbjørn Følstad, Petter Bae Brandtzæg, Tom Feltwell, Effie LC Law, Manfred Tscheligi, and Ewa A Luger. Sig: chatbots for social good. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2018.

[86] Center for Advanced Research on Language Acquisition. The center for advanced research on language acquisition (carla): Pragmatics and speech acts. http://carla.umn.edu/speechacts/thanks/american.html, 2020. [Online; Retrieved September 29, 2020].

[87] Mirta Galesic and Rocio Garcia-Retamero. Graph literacy: A cross-cultural comparison. *Medical Decision Making*, 31(3):444–457, 2011.

[88] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374, 2018.

[89] G. Gibbs. Learning by doing: A guide to teaching and learning methods. 1988.

[90] K. Glanz, B. Rimer, and K. Viswanath. Health behavior and health education : theory, research, and practice. 1991.

[91] Laura Gottlieb, Danielle Hessler, Dayna Long, Anais Amaya, and Nancy Adler. A randomized trial on screening for social determinants of health: the iscreen study. *Pediatrics*, 134(6):e1611–e1618, 2014.

[92] R. Gouveia, Fábio Pereira, E. Karapanos, S. Munson, and M. Hassenzahl. Exploring the design space of glanceable feedback for physical activity trackers. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016.

[93] Rúben Gouveia, Evangelos Karapanos, and Marc Hassenzahl. How do we engage with activity trackers? a longitudinal study of habito. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1305–1316, 2015.

[94] Jefferson Graham. Alexa is for fun, siri is because typing is hard: survey. https://www.usatoday.com/story/tech/talkingtech/2017/06/05/alexa-fun-siri-because-typing-hard-survey/102436072/, 2017. [Online; Retrieved September 28, 2020].

[95] Joseph Grandpre, Eusebio M Alvaro, Michael Burgoon, Claude H Miller, and John R Hall. Adolescent reactance and anti-smoking campaigns: A theoretical approach. *Health communication*, 15(3):349–366, 2003.

[96] James N Gribble, Heather G Miller, Susan M Rogers, and Charles F Turner. Interview mode and measurement of sexual behaviors: Methodological issues. *Journal of Sex research*, 36(1):16–24, 1999.

[97] Jonathan Grudin and Richard Jacques. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[98] Kent L Gustafson and Winston Bennett Jr. Promoting learner reflection: Issues and difficulties emerging from a three-year study. Technical report, GEORGIA UNIV ATHENS DEPT OF INSTRUCTIONAL TECHNOLOGY, 2002.

[99] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[100] A. Harrison and R. Crandall. Heterogeneity-homogeneity of exposure sequence and the attitudinal effects of exposure. *Journal of personality and social psychology*, 21 2:234–8, 1972.

[101] Matthias R Hastall and Silvia Knobloch-Westerwick. Severity, efficacy, and evidence type as determinants of health message exposure. *Health Communication*, 28(4):378–388, 2013.

[102] Katharine J Head, Seth M Noar, Nicholas T Iannarino, and Nancy Grant Harrington. Efficacy of text messaging-based interventions for health promotion: a meta-analysis. *Social science & medicine*, 97:41–48, 2013.

[103] health.gov. Physical activity guidelines for americans. https://health.gov/our-work/physical-activity/previous-guidelines/2008-physical-activity-guidelines, 2008. [Online; Retrieved September 27, 2020].

[104] Dirk Heerwegh and Geert Loosveldt. Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public opinion quarterly*, 72(5):836–846, 2008.

[105] Guillaume Hervet, Katherine Guérard, Sébastien Tremblay, and M. Chtourou. Is banner blindness genuine? eye tracking internet text advertising. *Applied Cognitive Psychology*, 25:708–716, 2011.

[106] Hyehyun Hong. Scale development for measuring health consciousness: Reconceptualization. *that Matters to the Practice*, page 212, 2009.

[107] Floris Hooglugt and Geke DS Ludden. A mobile app adopting an identity focus to promote physical activity (movedaily): iterative design study. *JMIR mHealth and uHealth*, 8(6):e16720, 2020.

[108] I-Han Hsiao, Shuguang Han, Manav Malhotra, Hui Soo Chae, and Gary Natriello. Survey sidekick: Structuring scientifically sound surveys. In *International conference on intelligent tutoring systems*, pages 516–522. Springer, 2014.

[109] Gary Hsieh, Ian Li, Anind Dey, Jodi Forlizzi, and Scott E Hudson. Using visualizations to increase compliance in experience sampling. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 164–167, 2008.

[110] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596. PMLR, 2017.

[111] Barry Hutchinson and Peter Bryson. Video, reflection and transformation: action research in vocational education and training in a european context. *Educational action research*, 5(2):283–303, 1997.

[112] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 895–906, 2018.

[113] K. Jolly, Amanda K. Lewis, J. Beach, J. Denley, P. Adab, J. Deeks, A. Daley, and P. Aveyard. Comparison of range of commercial or primary care led weight reduction programmes with minimal intervention control for weight loss in obesity: Lighten up randomised controlled trial. *The BMJ*, 343, 2011.

[114] D. Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93:1449–1475, 2003.

[115] Prashant Kale and Harbir Singh. Building firm capabilities through learning: the role of the alliance learning process in alliance capability and firm-level alliance success. *Strategic management journal*, 28(10):981–1000, 2007.

[116] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A Konstan, Loren Terveen, and F Maxwell Harper. Understanding how people use natural language to ask for recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 229–237, 2017.

[117] Evangelos Karapanos, Rúben Gouveia, Marc Hassenzahl, and Jodi Forlizzi. Wellbeing in the making: peoples' experiences with wearable activity trackers. *Psychology of well-being*, 6(1):1–17, 2016.

[118] Yuta Katsumi, Suhkyung Kim, Keen Sung, Florin Dolcos, and Sanda Dolcos. When nonverbal greetings "make it or break it": the role of ethnicity and gender in the effect of handshake on social appraisals. *Journal of Nonverbal Behavior*, 41(4):345–365, 2017.

[119] J. Kaye, Mary McCuistion, Rebecca Gulotta, and D. Shamma. Money talks: tracking personal finances. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[120] M. Kelly and M. Barker. Why is changing health-related behaviour so difficult? *Public health*, 136:109–16, 2016.

[121] R. Kelly, S. Zyzanski, and S. Alemagno. Prediction of motivation and behavior change following health promotion: role of health beliefs, social support, and self-efficacy. *Social science & medicine*, 32 3:311–20, 1991.

[122] David Kember, Doris YP Leung, Alice Jones, Alice Yuen Loke, Jan McKay, Kit Sinclair, Harrison Tse, Celia Webb, Frances Kam Yuet Wong, Marian Wong, et al. Development of a questionnaire to measure the level of reflective thinking. *Assessment & evaluation in higher education*, 25(4):381–395, 2000.

[123] Tom Kenter and M. Rijke. Short text similarity with word embeddings. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015.

[124] Soomin Kim, Joonhwan Lee, and G. Gweon. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[125] Young Hoon Kim, Dan J Kim, and Kathy Wachter. A study of mobile user engagement (moen): Engagement motivations, perceived value, satisfaction, and continued engagement intention. *Decision support systems*, 56:361–370, 2013.

[126] 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.

[127] 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.

[128] Ahmet Baki Kocaballi, Juan C Quiroz, Dana Rezazadegan, Shlomo Berkovsky, Farah Magrabi, Enrico Coiera, and Liliana Laranjo. Responses of conversational agents to health and lifestyle prompts: investigation of appropriateness and presentation structures. *Journal of medical Internet research*, 22(2):e15823, 2020.

[129] Rafal Kocielnik, Elena Agapie, Alexander Argyle, Dennis T Hsieh, Kabir Yadav, Breena Taira, and Gary Hsieh. Harborbot: A chatbot for social needs screening. In *AMIA Annual Symposium Proceedings*, volume 2019, page 552. American Medical Informatics Association, 2019.

[130] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 881–894, 2018.

[131] Rafal Kocielnik and Gary Hsieh. Send me a different message: utilizing cognitive space to create engaging message triggers. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2193–2207, 2017.

[132] Rafal Kocielnik, F. M. Maggi, and N. Sidorova. Enabling self-reflection with lifelogexplorer: Generating simple views from complex data. *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 184–191, 2013.

[133] Barbara Konat, Katarzyna Budzynska, and Patrick Saint-Dizier. Rephrase in argument structure. In *Proceedings of the Foundations of the Language of Argumentation (FLA) Workshop*, pages 32–39, 2016.

[134] Birgit R Krogstie, Michael Prilla, Daniel Wessel, Kristin Knipfer, and Viktoria Pammer. Computer support for reflective learning in the workplace: A model. In *2012 IEEE 12th International Conference on Advanced Learning Technologies*, pages 151–153. IEEE, 2012.

[135] Kimberly Kulavic, Cherilyn N. Hultquist, and J. McLester. A comparison of motivational factors and barriers to physical activity among traditional versus nontraditional college students. *Journal of American College Health*, 61:60 – 66, 2013.

[136] Mark Kutner, Elizabeth Greenburg, Ying Jin, and Christine Paulsen. The health literacy of america's adults: Results from the 2003 national assessment of adult literacy. nces 2006-483. *National Center for Education Statistics*, 2006.

[137] André Sousa Lago, João Pedro Dias, and Hugo Sereno Ferreira. Conversational interface for managing non-trivial internet-of-things systems. In *International Conference on Computational Science*, pages 384–397. Springer, 2020.

[138] Margaret D LeCompte. Analyzing qualitative data. *Theory into practice*, 39(3):146–154, 2000.

[139] Min Kyung Lee, Junsung Kim, Jodi Forlizzi, and Sara Kiesler. Personalization revisited: a reflective approach helps people better personalize health services and motivates them to increase physical activity. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 743–754, 2015.

[140] James Lester, Karl Branting, and Bradford Mott. Conversational agents. the practical handbook of internet computing. *Chapman & Hall. ISBN-10: 9781584883814*, 8:2–3, 2004.

[141] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[142] I. Li, Anind K. Dey, and J. Forlizzi. A stage-based model of personal informatics systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010.

[143] I. Li, Anind K. Dey, and J. Forlizzi. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *UbiComp '11*, 2011.

[144] I. Li, J. Forlizzi, and Anind K. Dey. Know thyself: monitoring and reflecting on facets of one's life. *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, 2010.

[145] Jingyi Li, Michelle X Zhou, Huahai Yang, and Gloria Mark. Confiding in and listening to virtual agents: The effect of personality. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 275–286, 2017.

[146] Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. What can you do? studying social-agent orientation and agent proactive interactions with an agent for employees. In *Proceedings of the 2016 acm conference on designing interactive systems*, pages 264–275, 2016.

[147] Q Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N Sadat Shami, and Werner Geyer. All work and no play? In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[148] James J Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B Strub. Fish'n'steps: Encouraging physical activity with an interactive computer game. In *International conference on ubiquitous computing*, pages 261–278. Springer, 2006.

[149] M. Lindström, A. Ståhl, K. Höök, P. Sundström, Jarmo Laaksolahti, Marco Combetto, A. Taylor, and Roberto Bresin. Affective diary: designing for bodily expressiveness and self-reflection. In *CHI EA '06*, 2006.

[150] Yang Liu and Mingyan Liu. An online learning approach to improving the quality of crowd-sourcing. *IEEE/ACM Transactions on Networking*, 25(4):2166–2179, 2017.

[151] I. Llovera, M. Ward, J. Ryan, Thalia LaTouche, and A. Sama. A survey of the emergency department population and their interest in preventive health education. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*, 10 2:155–60, 2003.

[152] Edwin A Locke and Gary P Latham. New directions in goal-setting theory. *Current directions in psychological science*, 15(5):265–268, 2006.

[153] Robert Loo and Karran Thorpe. Using reflective learning journals to improve individual and team performance. *Team performance management: an international journal*, 2002.

[154] Catherine L Lortie and Matthieu J Guitton. Judgment of the humanness of an interlocutor is in the eye of the beholder. *PLoS One*, 6(9):e25085, 2011.

[155] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.

[156] 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.

[157] P. Malecha, J. Williams, N. Kunzler, L. Goldfrank, H. Alter, and K. Doran. Material needs of emergency department patients: A systematic review. *Academic Emergency Medicine*, 25:330–359, 2018.

[158] Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. Mahi: investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 477–486, 2008.

[159] Fiona Martin and Mark Johnson. More efficient topic modelling through a noun only approach. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 111–115, 2015.

[160] Louis Martin, Benoît Sagot, Eric de la Clergerie, and Antoine Bordes. Controllable sentence simplification. *arXiv preprint arXiv:1910.02677*, 2019.

[161] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. Affectaura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 849–858, 2012.

[162] Graeme McLean and Kofi Osei-Frimpong. Hey alexa... examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior*, 99:28–37, 2019.

[163] Mary McMahon, Wendy Patton, and Mark Watson. Creating career stories through reflection: An application of the systems theory framework of career development. *Australian Journal of Career Development*, 13(3):13–17, 2004.

[164] David S Metzger, Beryl Koblin, Charles Turner, Helen Navaline, Francesca Valenti, Sarah Holte, Michael Gross, Amy Sheon, Heather Miller, Philip Cooley, et al. Randomized controlled trial of audio computer-assisted self-interviewing: utility and acceptability in longitudinal studies. *American journal of epidemiology*, 152(2):99–106, 2000.

[165] Jochen Meyer, Steven Simske, Katie A Siek, Cathal G Gurrin, and Hermie Hermens. Beyond quantified self: Data for wellbeing. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pages 95–98. 2014.

[166] Sallyanne Miller. What it's like being the 'holder of the space': a narrative on working with reflective practice in groups. *Reflective Practice*, 6(3):367–377, 2005.

[167] S. Milne, S. Orbell, and P. Sheeran. Combining motivational and volitional interventions to promote exercise participation: protection motivation theory and implementation intentions. *British journal of health psychology*, 7 Pt 2:163–84, 2002.

[168] J. Moon. Reflection in learning & professional development: Theory & practice. 1999.

[169] Y. Moon. Personalization and personality: Some effects of customizing message style based on consumer personality. *Journal of Consumer Psychology*, 12:313–325, 2002.

[170] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research*, 20(6):e10148, 2018.

[171] Andreea Muresan and Henning Pohl. Chats with bots: balancing imitation and engagement. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[172] Clifford Nass, Katherine Isbister, Eun-Ju Lee, et al. Truth is beauty: Researching embodied conversational agents. *Embodied conversational agents*, pages 374–402, 2000.

[173] Roni Neff and Jillian Fry. Periodic prompts and reminders in health promotion and health behavior interventions: systematic review. *Journal of medical Internet research*, 11(2):e16, 2009.

[174] Christine M Neuwirth, Ravinder Chandhok, David Charney, Patricia Wojahn, and Loel Kim. Distributed collaborative writing: A comparison of spoken and written modalities for reviewing and revising documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 51–57, 1994.

[175] Annie WY Ng, HW Lo, and AH Chan. Measuring the usability of safety signs: A use of system usability scale (sus). In *proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 2, pages 1296–1301. Citeseer, 2011.

[176] Hien Nguyen and J. Masthoff. Designing persuasive dialogue systems: Using argumentation with care. In *PERSUASIVE*, 2008.

[177] J. Norcross, Marci S Mrykalo, and M. Blagys. Auld lang syne: success predictors, change processes, and self-reported outcomes of new year's resolvers and nonresolvers. *Journal of clinical psychology*, 58 4:397–405, 2002.

[178] P. Norris. Digital divide: Civic engagement, information poverty, and the internet worldwide. 2001.

[179] Jekaterina Novikova, Oliver Lemon, and Verena Rieser. Crowd-sourcing nlg data: Pictures elicit better data. *arXiv preprint arXiv:1608.00339*, 2016.

[180] Heather L O'Brien and Elaine G Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2010.

[181] Shereen Oraby, Pritam Gundecha, Jalal Mahmud, Mansurul Bhuiyan, and Rama Akkiraju. " how may i help you?" modeling twitter customer serviceconversations using fine-grained dialogue acts. In *Proceedings of the 22nd international conference on intelligent user interfaces*, pages 343–355, 2017.

[182] Marcia G Ory, Matthew Lee Smith, Nelda Mier, and Meghan M Wernicke. The science of sustaining health behavior change: the health maintenance consortium. *American journal of health behavior*, 34(6):647–659, 2010.

[183] Mina Park, Milam Aiken, and Laura Salvador. How do humans interact with chatbots?: An analysis of transcripts. *International Journal of Management and Information Technology*, 14:3338–3350, 2018.

[184] Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*, 2020.

[185] J. Pennebaker, M. Mehl, and Kate Niederhoffer. Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology*, 54:547–77, 2003.

[186] Rifca Peters, Joost Broekens, and Mark A Neerincx. Guidelines for tree-based collaborative goal setting. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 401–405, 2017.

[187] R. Petty and J. Cacioppo. Attitudes and persuasion: Classic and contemporary approaches. 1981.

[188] Afarin Pirzadeh, Li He, and Erik Stolterman. Personal informatics and reflection: a critical examination of the nature of reflection. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 1979–1988. 2013.

[189] Mark Pope. A brief history of career counseling in the united states. *The career development quarterly*, 48(3):194–211, 2000.

[190] David B Portnoy, Lori AJ Scott-Sheldon, Blair T Johnson, and Michael P Carey. Computer-delivered interventions for health promotion and behavioral risk reduction: a meta-analysis of 75 randomized controlled trials, 1988–2007. *Preventive medicine*, 47(1):3–16, 2008.

[191] J. Prochaska and W. Velicer. The transtheoretical model of health behavior change. *American Journal of Health Promotion*, 12:38 – 48, 1997.

[192] Kiran Ramesh, Surya Ravishankaran, Abhishek Joshi, and K Chandrasekaran. A survey of design techniques for conversational agents. In *International conference on information, communication and computing technology*, pages 336–350. Springer, 2017.

[193] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*, 2019.

[194] R. Rickenberg and Byron Reeves. The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2000.

[195] J. Riedel. Using a health and productivity dashboard: A case example. *American Journal of Health Promotion*, 22:1 – 12, 2007.

[196] Verónica Rivera-Pelayo, Valentin Zacharias, Lars Müller, and Simone Braun. Applying quantified self approaches to support reflective learning. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 111–114, 2012.

[197] Susan Robinson, David R Traum, Midhun Ittycheriah, and Joe Henderer. What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In *LREC*, 2008.

[198] Stephen Rollnick, William R Miller, and Christopher Butler. *Motivational interviewing in health care: helping patients change behavior*. Guilford Press, 2008.

[199] Catherine A Roster, Robert D Rogers, Gerald Albaum, and Darin Klein. A comparison of response characteristics from web and telephone surveys. *International Journal of Market Research*, 46(3):359–373, 2004.

[200] Susana Rubio, Eva Díaz, Jesús Martín, and José M Puente. Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied psychology*, 53(1):61–86, 2004.

[201] Caryl E Rusbult, Stephen M Drigotas, and Julie Verette. The investment model: An interdependence analysis of commitment processes and relationship maintenance phenomena. 1994.

[202] K. Ryokai, F. Michahelles, M. Kritzler, and Suhaib Syed. Communicating and interpreting wearable sensor data with health coaches. *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 221–224, 2015.

[203] Günther Sagl, Bernd Resch, and Thomas Blaschke. Contextual sensing: Integrating contextual information with human and technical geo-sensor information for smart cities. *Sensors*, 15(7):17013–17035, 2015.

[204] Shruti Sannon, Brett Stoll, Dominic DiFranzo, Malte Jung, and Natalya N Bazarova. How personification and interactivity influence stress-related disclosures to conversational agents. In *companion of the 2018 ACM conference on computer supported cooperative work and social computing*, pages 285–288, 2018.

[205] Donald A Schon. *The reflective practitioner: How professionals think in action*, volume 5126. Basic books, 1984.

[206] D. Schulman and T. Bickmore. Persuading users through counseling dialogue with a conversational agent. In *Persuasive '09*, 2009.

[207] Daniel Schulman and Timothy Bickmore. Persuading users through counseling dialogue with a conversational agent. In *Proceedings of the 4th international conference on persuasive technology*, pages 1–8, 2009.

[208] David Schumann, R. Petty, and D. Clemons. Predicting the effectiveness of different strategies of advertising variation: A test of the repetition-variation hypotheses. *Journal of Consumer Research*, 17:192–202, 1990.

[209] S. Schwartz, Jan Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J. Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103 4:663–88, 2012.

[210] A. Schwerdtfeger, C. Schmitz, and M. Warken. Using text messages to bridge the intention-behavior gap? a pilot study on the use of text message reminders to increase objectively assessed physical activity in daily life. *Frontiers in Psychology*, 3, 2012.

[211] John R Searle. Austin on locutionary and illocutionary acts. *The philosophical review*, 77(4):405–424, 1968.

[212] John R Searle, Ferenc Kiefer, Manfred Bierwisch, et al. *Speech act theory and pragmatics*, volume 10. Springer, 1980.

[213] Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.

[214] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *arXiv preprint arXiv:2101.07714*, 2021.

[215] Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*, 2020.

[216] Miriam Sherin and Elizabeth van Es. Using video to support teachers' ability to interpret classroom interactions. In *society for information technology & teacher education international conference*, pages 2532–2536. Association for the Advancement of Computing in Education (AACE), 2002.

[217] Ingo Siegert. "alexa in the wild"–collecting unconstrained conversations with a modern voice assistant in a public environment. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 615–619, 2020.

[218] P. Slovák, C. Frauenberger, and G. Fitzpatrick. Reflective practicum: A framework of sensitising concepts to design for transformative reflection. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[219] Kenny Smith, Amy Perfors, O. Feher, Anna Samara, K. Swoboda, and E. Wonnacott. Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372, 2017.

[220] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

[221] Lee Sproull, Mani Subramani, Sara Kiesler, Janet H Walker, and Keith Waters. When the interface is a face. *Human-computer interaction*, 11(2):97–124, 1996.

[222] Jessica Stillman. Hiring a Remote Worker? 7 Interview Questions to Ask. https://www.inc.com/jessica-stillman/hiring-remote-workers-interview-questions-to-ask.html, 2013. [Online; Retrieved September 28, 2020].

[223] Victor J Strecher, Saul Shiffman, and Robert West. Randomized controlled trial of a web-based computer-tailored smoking cessation program as a supplement to nicotine patch therapy. *Addiction*, 100(5):682–688, 2005.

[224] Nina Svenningsson and Montathar Faraon. Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, pages 151–161, 2019.

[225] Robert Tobias. Changing behavior by memory aids: A social psychological model of prospective memory and habit development tested with dynamic field data. *Psychological review*, 116(2):408, 2009.

[226] US VA. Sample pbi questions - performance based interviewing (pbi). https://www.va.gov/PBI/Questions.asp, 2018. [Online; Retrieved September 28, 2020].

[227] Aukje AC Verhoeven, Marieke A Adriaanse, Denise TD De Ridder, Emely De Vet, and Bob M Fennis. Less is more: The effect of multiple implementation intentions targeting unhealthy snacking habits. *European Journal of Social Psychology*, 43(5):344–354, 2013.

[228] M Vagias Wade et al. Likert-type scale response anchors. *Clemson international institute for tourism & research development, department of parks, recreation and tourism management, clemson university*, 2006.

[229] Harald Walach, Nina Buchheld, Valentin Buttenmüller, Norman Kleinknecht, and Stefan Schmidt. Measuring mindfulness—the freiburg mindfulness inventory (fmi). *Personality and individual differences*, 40(8):1543–1555, 2006.

[230] G. Walsh and J. Golbeck. Stepcity: a preliminary investigation of a personal informatics-based social game on behavior change. *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, 2014.

[231] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia*, pages 12–20, 2019.

[232] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[233] Wei Wei, Bei Zhou, and Georgios Leontidis. A hybrid natural language generation system integrating rules and deep learning algorithms. *arXiv preprint arXiv:2006.09213*, 2020.

[234] Barry D Weiss, Mary Z Mays, William Martz, Kelley Merriam Castro, Darren A DeWalt, Michael P Pignone, Joy Mockbee, and Frank A Hale. Quick assessment of literacy in primary care: the newest vital sign. *The Annals of Family Medicine*, 3(6):514–522, 2005.

[235] J. Weizenbaum. Eliza — a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 26:23–28, 1983.

[236] Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. How well sentence embeddings capture meaning. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8, 2015.

[237] WHO. Who — prevalence of insufficient physical activity. http://www.who.int/gho/ncd/risk_factors/physical_activity/en/, 2008. [Online; Retrieved September 27, 2020].

[238] Robert A Wicklund. *Freedom and reactance.* Lawrence Erlbaum, 1974.

[239] Ziang Xiao, M. Zhou, Q. V. Liao, G. Mark, Changyan Chi, W. Chen, and H. Yang. Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys. *arXiv: Human-Computer Interaction*, 2019.

[240] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510, 2017.

[241] Qiongkai Xu, Chenchen Xu, and Lizhen Qu. Alter: Auxiliary text rewriting tool for natural language generation. *arXiv preprint arXiv:1909.06564*, 2019.

[242] Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. Better conversations by modeling, filtering, and optimizing for coherence and diversity. *arXiv preprint arXiv:1809.06873*, 2018.

[243] Özge Nilay Yalçın. Empathy framework for embodied conversational agents. *Cognitive Systems Research*, 59:123–132, 2020.

[244] Rui Yan, Yiping Song, and Hua Wu. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64, 2016.

[245] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. Diverse few-shot text classification with multiple metrics. *arXiv preprint arXiv:1805.07513*, 2018.

[246] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

[247] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

[248] Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*, 2017.

[249] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.

[250] Maurizio Zollo and Sidney G Winter. Deliberate learning and the evolution of dynamic capabilities. *Organization science*, 13(3):339–351, 2002.

Appendix A

EXAMPLES OF REACTIONS MATCHED TO QUESTION-ANSWER CONTEXT

Reaction class	**Survey question context (framing class in [])**	**Possible reactions within class**
Neutral acknowledgment	Q: Would you mind sharing what gender you identify as? [neu] A: Female [neu]	"Thanks for sharing." "Got it!" "I took a note of your answer." "Got it! Thanks for letting me know."
	Q: What kind of financial accounts do you currently have? [neu] A: Money Market [neu]	
Expression of satisfaction	Q: Would you say that exercising for at least 20 minutes, three times per week for the next three months would be pleasant? [pos] A: Strongly agree [pos]	"I am glad to hear that." "Sounds great!" "That sounds positive, I am glad." "I am happy that's the case."
	Q: Could you please indicate how much difficulty, on average, do you have with walking a quarter of a mile? [neg] A: No difficulty [neg]	
Expression of compassion	Q: Can you tell me how well-maintained are the facilities at this university? [pos] A: Not so well-maintained [neg]	"I am sorry to hear that." "That sounds stressful." "That's hard to hear." "That must be frustrating." "So sorry about that."
	Q: Have you experienced feeling down, depressed, or hopeless? [neg] A: Several days [pos]	

Table A.1: Examples of empathetic reactions matched to local question-answer context. Phrases in-between square brackets have been added or modified.

Appendix B

PHRASING CATEGORIES USED FOR QUESTION REPHRASING IN AUTOMATION

Phrasing category	Question example	Matching prefixes*	Modification rules*
Adverb-based question	What gender do you identify as?	Can you tell me... Could I ask you...	
Verb-based question	Are you married?	Could I ask you whether... Please tell me if...	are you → you are i → you
Verb-based statement	Feeling tired or having little energy.	Have you experienced... Did you experience...	. → ?
Noun-based statement	This course requires us to understand concepts taught by the lecturer.	Would you say that… Is it true that...	are you → you are i → you us → you . → ?
Request-action	Indicate your current age.	Can you... Can I ask you to...	. → ?
None	If you've had any days with issues above, how difficult have these problems been?	N/A	N/A

Table B.1: Phrasing categories for survey questions derived empirically from survey data. Each category is composed from a prefix that is preprended to the original text of survey questions and a set of modification rules which change the text of the question to fit the 3rd person & question form.

Appendix C

HOLD-OUT SURVEYS USED IN THE USER STUDY EVALUATION

- Big Five Inventory-10 (BFI-10)
- Informal Fitness survey
- Personal Finance Survey (SurveyMonkey)
- Values Survey
- Portrait Values Questionnaire (PVQ)
- Sleep Quality Scale (SQS)

Survey	Domain	# items	Question format	Answers format
Big 5	personality	11	1st person *"I see myself as someone who is reserved"*	5-point likert (5 different)
Sleep	sleep	19	2st person *"My sleep hours are enough."*	4 level custom frequency scale (4 different)
Personal finance	finance	17	1st & 3rd person "Saving and investing is important to me."	5-point likert + domain options (25 different)
Fitness	physical activity	11	3rd person "How hard do you work out?"	Informal 4 point scales (40 different)
Values survey	political views	16	1st person + passive voice "A good government should aim chiefly at more aid for the poor, sick, and old"	4 preference (4 different)
PVQ	personality	14	2nd person "Having a good time is important to him. He likes to 'spoil' himself."	6-point likert only (6 different)

Appendix D

SURVEYS USED FOR ML DEVELOPMENT

- Vulnerable Elders Survey (VES-13)
- Theory of Planned Behavior Survey
- Student Satisfaction
- SDOH Short Screener
- Kember's Reflection Survey
- Positive and Negative Affect Schedule (PANAS-SF)
- NASA TLX
- Health Literacy
- 3 Minute Depression Test
- Demographics #1
- Demographics #2
- Demographics #3
- Demographics #4
- Climate Change
- Harbor

Appendix E

ML PERFORMANCE ON THE FULL DATASET

		Survey Domain	Question Language Adaptation	Question Empathy Framing	Answer Empathy Framing	Empathetic Reaction (derived)†
5-fold CV	**Accuracy**	0.65±.06	0.85±.05	0.75±.07	0.86±.03	0.73±.08
	F1-score	0.64±.07	0.84±.05	0.75±.07	0.86±.03	0.71±.09
	Precision	0.67±.08	0.85±.05	0.77±.06	0.87±.02	0.76±.07
	Recall	0.65±.06	0.85±.05	0.75±.07	0.86±.03	0.74±.08
Leave-one -out	**Accuracy**	0.54±.41	0.74±.22	0.64±.19	0.82±.24	0.56±.17
	F1-score	0.60±.41	0.76±.22	0.65±.18	0.81±.25	0.54±.16
	Precision	0.73±.45	0.83±.21	0.75±.18	0.82±.25	0.70±.26
	Recall	0.54±.41	0.74±.22	0.64±.19	0.82±.24	0.63±.25

† - Empathetic Reaction is derived from the Question Empathy Framing and Answer Empathy Framing using a fixed rule.

Table E.1: Classification performance for the 4 text classification tasks (+1 derived) on a full dataset of 22 surveys (combined 16 development and 6 hold-out surveys). Question Empathy Framing and Answer Empathy Framing classifications are part of empathetic addition - the results of these two classifications taken together are used to decide on reaction class

Appendix F

MANUAL QUESTION CORRECTIONS IN CHAPTER 6

Sleep Quality - 6 of 19 questions corrected

#	Auto generated	After manual correction	Edit	Cause
1	Next, did you experience i have difficulty falling asleep?	Next, have you experienced difficulty falling asleep?	7	Wrong class
2	Let's carry on, can you share whether you have experienced poor sleep gives me headaches?	Let's carry on, can you share whether you have experienced poor sleep giving you headaches?	6	Wrong class
3	Please tell me if you have experienced poor sleep makes me irritated?	Please tell me if you have experienced poor sleep making you irritated?	6	Wrong class
4	Moving on, did you experience poor sleep makes me lose my appetite?	Moving on, did you experience poor sleep making you lose your appetite?	10	Wrong class
5	So, please tell me if you've experienced poor sleep makes me lose interest in work or others?	So, please tell me if you've experienced poor sleep making you lose interest in work or others?	6	Wrong class
6	Can you share whether you've experienced my fatigue is relieved after sleep?	Can you share whether you've experienced your fatigue being relieved after sleep?	8	Wrong class

Big 5 survey - 10 of 11 questions corrected

#	Auto generated	After manual correction	Edit	Cause
1	Moving on, is it fair to say that you see myself as someone who is reserved	Moving on, is it fair to say that you see yourself as someone who is reserved	4	No rule: 'myself' → 'yourself'
2	Further, please indicate the extent to which you see myself as someone who is generally trusting?	Further, please indicate the extent to which you see yourself as someone who is generally trusting?	4	No rule: 'myself' → 'yourself'
3	Do you think it's fair to say that you see myself as someone who tends to be lazy?	Do you think it's fair to say that you see yourself as someone who tends to be lazy?	4	No rule: 'myself' → 'yourself'
4	Next, does it make sense to say that you see myself as someone who is relaxed, handles stress well?	Next, does it make sense to say that you see yourself as someone who is relaxed, handles stress well?	4	No rule: 'myself' → 'yourself'
5	Let's carry on, do you think it's fair to say that you see myself as someone who has few artistic interests?	Let's carry on, do you think it's fair to say that you see yourself as someone who has few artistic interests?	4	No rule: 'myself' → 'yourself'
6	Let's carry on, is it true that you see myself as someone who is outgoing, sociable	Let's carry on, is it true that you see yourself as someone who is outgoing, sociable	4	No rule: 'myself' → 'yourself'
7	Do you think that you see myself as someone who tends to find fault with others?	Do you think that you see yourself as someone who tends to find fault with others?	4	No rule: 'myself' → 'yourself'
8	Moving on, would you say that you see myself as someone who does a thorough job?	Moving on, would you say that you see yourself as someone who does a thorough job?	4	No rule: 'myself' → 'yourself'
9	Is it fair to say that you see myself as someone who gets nervous easily?	Is it fair to say that you see yourself as someone who gets nervous easily?	4	No rule: 'myself' → 'yourself'
10	Further, does it make sense to say that you see myself as someone who has an active imagination?	Further, does it make sense to say that you see yourself as someone who has an active imagination?	4	No rule: 'myself' → 'yourself'

Finance survey - 5 of 17 questions corrected

#	Auto generated	After manual correction	Edit	Cause
1	So, can you say that you have an emergency savings fund established to cover 3 to 6 months of expenses should you lose the ability to work?	Do you have an emergency savings fund established to cover 3 to 6 months of expenses should you lose the ability to work?	19	Wrong class
2	Going forward, can I ask you to i feel capable of handling my financial future overall?	Going forward, do you feel capable of handling your financial future overall?	17	Wrong class
3	Further, please indicate the extent to which you believe you have adequate information to help make the best financial decisions for you and your family?	Further, do you believe you have adequate information to help make the best financial decisions for you and your family?	33	Wrong class
4	Continuing, could you say that you feel you have a good grasp on the importance of insurance in all of its forms. (life, health, disability, Long Term Care)?	Continuing, do you feel you have a good grasp on the importance of insurance in all of its forms. (life, health, disability, Long Term Care)?	16	Wrong class
5	Do you feel that you feel comfortable about your financial future because you have adequately planned for it?	Do you feel comfortable about your financial future because you have adequately planned for it?	14	Wrong class

Values & Politics survey - 2 of 16 questions corrected

#	Auto generated	After manual correction	Edit	Cause
1	Continuing, could you tell me a good government should aim chiefly at introducing the highest ethical principles into its policies?	Continuing, could you say a good government should aim chiefly at introducing the highest ethical principles into its policies?	7	Wrong class
2	Moving on, can I ask you a good government should aim chiefly at introducing the highest ethical principles into its policies?	Moving on, should a good government should aim chiefly at introducing the highest ethical principles into its policies?	12	No rule: 'I' → 'you' for sentence start

Fitness survey - 2 of 11 questions corrected

#	Auto generated	After manual correction	Edit	Cause
1	Would you mind sharing do you have a workout buddy?	Would you mind sharing whether you have a workout buddy?	7	Wrong class
2	Would you mind sharing how do you feel after a workout?	Would you mind sharing how you feel after a workout?	3	Wrong class

PVQ values survey - 12 of 14 questions corrected

#	Auto generated	After manual correction	Edit	Cause
1	Would you say that thinking up new ideas and being creative is important to him. He likes to do things in his own original way.	Would you say that thinking up new ideas and being creative is important to you. You like to do things in your own original way?	12	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
2	Next, is it fair to say that it is important to him to be rich. He wants to have a lot of money and expensive things.	Next, is it fair to say that it is important to you to be rich. You want to have a lot of money and expensive things?	8	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
3	He thinks it is important that every person in the world should be treated equally. He believes everyone should have equal opportunities in life.	Do you think it is important that every person in the world should be treated equally. You believe everyone should have equal opportunities in life?	12	Wrong class + No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
4	Do you think that it's important to him to show his abilities. He wants people to admire what he does.	Do you think that it's important to you to show your abilities. You want people to admire what you do?	17	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
5	Does it make sense to say that it is important to him to live in secure surroundings. He avoids anything that might endanger his safety.	Does it make sense to say that it is important to you to live in secure surroundings. You avoid anything that might endanger your safety?	12	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
6	Do you think it's fair to say that he likes surprises and is always looking for new things to do. He thinks it is important to do lots of different things in life.	Do you think it's fair to say that you like surprises and *are* always looking for new things to do. You think it is important to do lots of different things in life?	12	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd

#	Auto generated	After manual correction	Edit	Cause
7	So, could you tell me he believes that people should do what they are told. He thinks people should follow rules at all times, even when no-one is watching.	So, could you tell me whether you believe that people should do what they are told. You think people should follow rules at all times, even when no-one is watching?	15	Wrong class + No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
8	Then, could you say that it is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them.	Then, could you say that it is important to you to listen to people who are different from you. Even when you disagree with them, you still want to understand them.	14	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
9	Further, can you say that it is important to him to be humble and modest. He tries not to draw attention to himself.	Further, can you say that it is important to you to be humble and modest. You try not to draw attention to yourself?	14	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
10	Please indicate the extent to which having a good time is important to him. He likes to "spoil" himself.	Please indicate the extent to which having a good time is important to you. You like to "spoil" yourself?	12	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
11	Continuing, do you feel that it is important to him to make his own decisions about what he does. He likes to be free and not depend on others.	Continuing, do you feel that it is important to you to make your own decisions about what you do. You like to be free and not depend on others?	17	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd
12	Moving on, do you think it's fair to say that it's very important to him to help the people around him. He wants to care for their well-being.	Moving on, do you think it's fair to say that it's very important to you to help the people around you. You want to care for their well-being?	11	No rules: 'he'→'you', 'him'→'you', 'his'→'your', VERB→3rd

Appendix G

MANUAL REACTION CORRECTIONS IN CHAPTER 6

PVQ survey - 6 of 72 (8.3%) reactions corrected

#	Auto generated	After manual correction	Edit	Cause
	Q: He likes surprises and is always looking for new things to do. He thinks it is important to do lots of different things in life.		**236**	Wrong Q cat
1	**A:** Very much like me Okay, I'm getting a better idea of your answers	That sounds positive	40	
2	**A:** Like me Okay, I'm getting a better idea of your answers	That sounds positive	40	
3	**A:** Somewhat like me Okay, I'm getting a better idea of your answers	That sounds positive	40	
4	**A:** A little like me Okay, I'm getting a better idea of your answers	That sounds positive	40	
5	**A:** Not like me Okay, I'm getting a better idea of your answers	That sounds stressful	38	
6	**A:** Not like me at all Okay, I'm getting a better idea of your answers	That sounds stressful	38	

Sleep survey - 34 of 72 (47.2%) reactions corrected

#	Auto generated	After manual correction	Edit	Cause
	Q: I have difficulty falling asleep		**18**	
	A: Sometimes: 1-2 times a week			
1	That sounds stressful	Sure	18	Wrong A cat
	Q: I fall into a deep sleep		**66**	Wrong Q cat
	A: Rarely: None or 1-3 times a month			
2	Sounds good	That is frustrating	17	
	A: Sometimes: 1-2 times a week			
3	I am sorry to hear that	Got it	19	Wrong A cat
	A: Often: 3-5 times a week			
4	I am sorry to hear that	I am happy that's the case	15	
	A: Almost always: 6-7 times a week			
5	I am sorry to hear that	I am happy that's the case	15	
	Q: I have difficulty getting back to sleep once you wake up in middle of the night.		**12**	
	A: Sometimes: 1-2 times a week			
6	That's hard to hear	Thanks for sharing	12	Wrong A cat
	Q: I wake up easily because of noise.		**22**	
	A: Sometimes: 1-2 times a week			
7	So sorry about that	Thanks for letting me know	22	Wrong A cat
	Q: I toss and turn.		**21**	
	A: Sometimes: 1-2 times a week			
8	I am happy that's the case	Got it! Thanks for sharing	21	Wrong A cat

#	Auto generated	After manual correction	Edit	Cause
	Q: I never go back to sleep after awakening during sleep.		**15**	
	A: Sometimes: 1-2 times a week			
9	That is frustrating	Got it.	15	Wrong A cat
	Q: I feel refreshed after sleep		**67**	Wrong Q cat
	A: Rarely: None or 1-3 times a month			
10	That's really great!	I am sorry to hear that	17	
	A: Sometimes: 1-2 times a week			
11	So sorry about that	Got it! Thanks for sharing	20	Wrong A cat
	A: Often: 3-5 times a week			
12	So sorry about that	Sounds good	15	
	A: Almost always: 6-7 times a week			
13	So sorry about that	Sounds good	15	
	Q: I feel unlikely to sleep after sleep		**53**	Wrong Q cat
	A: Rarely: None or 1-3 times a month			
14	That's really great!	That's hard to hear	11	
	A: Sometimes: 1-2 times a week			
15	That is frustrating	Noted	18	Wrong A cat
	A: Often: 3-5 times a week			
16	That is frustrating	That sounds positive	12	
	A: Almost always: 6-7 times a week			
17	That is frustrating	That sounds positive	12	
	Q: Poor sleep gives me headaches.		**38**	
	A: Sometimes: 1-2 times a week			
18	That sounds stressful	Okay, I'm getting a better idea of your answers	38	Wrong A cat
	Q: Poor sleep makes me irritated.		**12**	
	A: Sometimes: 1-2 times a week			
19	That's hard to hear	Thanks for sharing	12	Wrong A cat

#	Auto generated	After manual correction	Edit	Cause
	Q: My sleep hours are enough.		**75**	Wrong Q cat
	A: Rarely: None or 1-3 times a month			
20	Sounds nice	That is frustrating	16	
	A: Sometimes: 1-2 times a week			
21	I am sorry to hear that	Sure	21	Wrong A cat
	A: Often: 3-5 times a week			
22	I am sorry to hear that	Okay, that's good	19	
	A: Almost always: 6-7 times a week			
23	I am sorry to hear that	Okay, that's good	19	
	Q: Poor sleep makes me lose my appetite.		**16**	
	A: Sometimes: 1-2 times a week			
24	That is frustrating	Thank you for your answer	16	Wrong A cat
	Q: Poor sleep makes hard for me to think.		**39**	
	A: Sometimes: 1-2 times a week			
25	Thanks for sharing that	Okay, I'm getting a better idea of your answers	39	Wrong A cat
	Q: I feel vigorous after sleep.		**62**	Wrong Q cat
	A: Rarely: None or 1-3 times a month			
26	Great!	So sorry about that	16	
	A: Sometimes: 1-2 times a week			
27	That sounds stressful	Got it	18	Wrong A cat
	A: Often: 3-5 times a week			
28	That sounds stressful	Sounds nice	14	
	A: Almost always: 6-7 times a week			
29	That sounds stressful	Sounds nice	14	
	Q: Poor sleep makes me lose interest in work or others.		**20**	
	A: Sometimes: 1-2 times a week			
30	I am sorry to hear that	Noted	20	Wrong A cat

#	Auto generated	After manual correction	Edit	Cause
	Q: My fatigue is relieved after sleep.		**65**	Wrong Q cat
	A: Rarely: None or 1-3 times a month			
31	That's good	That is frustrating	13	
	A: Sometimes: 1-2 times a week			
32	So sorry about that	Got it! Thanks for sharing	20	Wrong A cat
	A: Often: 3-5 times a week			
33	So sorry about that	Great!	16	
	A: Almost always: 6-7 times a week			
34	So sorry about that	Great!	16	

Big5 survey - 20 of 50 (40.0%) reactions corrected

#	Auto generated	After manual correction	Edit	Cause
	Q: I see myself as someone who is generally trusting.		**84**	Wrong Q cat
	A: Disagree strongly			
1	Thanks for letting me know	I am sorry to hear that	20	
	A: Disagree a little			
2	Thanks for letting me know	I am sorry to hear that	20	
	A: Agree a little			
3	Thanks for letting me know	I am glad to hear that	22	
	A: Agree strongly			
4	Thanks for letting me know	I am glad to hear that	22	

#	Auto generated	After manual correction	Edit	Cause
	Q: I see myself as someone who has few artistic interests.		**50**	Wrong Q cat
	A: Disagree strongly			
5	Sure	Sounds nice	8	
	A: Disagree a little			
6	Sure	Sounds nice	8	
	A: Agree a little			
7	Sure	So sorry about that	17	
	A: Agree strongly			
8	Sure	So sorry about that	17	
	Q: I see myself as someone who is outgoing, sociable		**76**	Wrong Q cat
	A: Disagree strongly			
9	Noted	That sounds stressful	18	
	A: Disagree a little			
10	Noted	That sounds stressful	18	
	A: Agree a little			
11	Noted	That sounds positive	18	
	A: Agree strongly			
12	Noted	That sounds positive	18	
	Q: I see myself as someone who tends to find fault with others		**66**	Wrong Q cat
	A: Disagree strongly			
13	So sorry about that	Great!	16	
	A: Disagree a little			
14	So sorry about that	Great!	16	
	A: Agree a little			
15	That sounds positive	Thanks for sharing that	17	
	A: Agree strongly			
16	That sounds positive	Thanks for sharing that	17	

#	Auto generated	After manual correction	Edit	Cause
	Q: I see myself as someone who does a thorough job		**70**	Wrong Q cat
	A: Disagree strongly			
17	Sure	That is frustrating	17	
	A: Disagree a little			
18	Sure	That is frustrating	17	
	A: Agree a little			
19	Sure	That sounds positive	18	
	A: Agree strongly			
20	Sure	That sounds positive	18	

Values survey - 38 of 64 (59.4%) reactions corrected

#	Auto generated	After manual correction	Edit	Cause
	Q: A good government should aim chiefly at more aid for the poor, sick, and old.		**51**	Wrong Q cat
	A: highest preference			
1	So sorry about that	That's really great!	17	
	A: second preference			
2	So sorry about that	That's really great!	17	
	A: third preference			
3	So sorry about that	Thanks for sharing.	17	
	Q: A good government should aim chiefly at the development of manufacturing and trade.		**9**	
	A: lowest preference			
4	Good to hear that	I am sorry to hear that	9	Wrong A cat
	Q: A good government should aim chiefly at introducing the highest ethical principles into its policies.		**15**	
	A: lowest preference			
5	Sounds good	Sorry to hear that	15	Wrong A cat

#	Auto generated	After manual correction	Edit	Cause
	Q: A good government should aim chiefly at establishing a position of power and respect among nations.		**16**	
	A: lowest preference			
6	Sounds nice	That is frustrating	16	Wrong A cat
	Q: Someone who works all week would best spend the weekend keeping up on the latest in scientific advances.		**84**	Wrong Q cat
	A: highest preference			
7	I am happy that's the case	Thanks for sharing	21	
	A: second preference			
8	I am happy that's the case	Thanks for sharing	21	
	A: third preference			
9	I am happy that's the case	Thanks for sharing	21	
	A: lowest preference			
10	I am happy that's the case	Thanks for sharing	21	Wrong A cat
	Q: Someone who works all week would best spend the weekend trying to win at golf or other sport.		**72**	Wrong Q cat
	A: highest preference			
11	That's really great!	Sure	18	
	A: second preference			
12	That's really great!	Sure	18	
	A: third preference			
13	That's really great!	Sure	18	
	A: lowest preference			
14	That's really great!	Sure	18	Wrong A cat

#	Auto generated	After manual correction	Edit	Cause
	Q: Someone who works all week would best spend the weekend going to a classical music concert or art museum.		**80**	Wrong Q cat
	A: highest preference			
15	I am glad to hear that	Noted	20	
	A: second preference			
16	I am glad to hear that	Noted	20	
	A: third preference			
17	I am glad to hear that	Noted	20	
	A: lowest preference			
18	I am glad to hear that	Noted	20	Wrong A cat
	Q: If I could influence the educational policies of the public schools of some city, I would try to promote the study of and participation in music and the fine arts.		**164**	Wrong Q cat
	A: highest preference			
19	That's good	Okay, I'm getting a better idea of your answers	41	
	A: second preference			
20	That's good	Okay, I'm getting a better idea of your answers	41	
	A: third preference			
21	That's good	Okay, I'm getting a better idea of your answers	41	
	A: lowest preference			
22	That's good	Okay, I'm getting a better idea of your answers	41	Wrong A cat

#	Auto generated	After manual correction	Edit	Cause
	Q: If I could influence the educational policies of the public schools of some city, I would try to encourage the study of social problems.		**84**	Wrong Q cat
	A: highest preference			
23	I am happy that's the case	Thanks for letting me know	21	
	A: second preference			
24	I am happy that's the case	Thanks for letting me know	21	
	A: third preference			
25	I am happy that's the case	Thanks for letting me know	21	
	A: lowest preference			
26	I am happy that's the case	Thanks for letting me know	21	Wrong A cat
	Q: If I could influence the educational policies of the public schools of some city, I would try to provide additional laboratory facilities.		**80**	Wrong Q cat
	A: highest preference			
27	Sounds nice	Thank you for your answer	20	
	A: second preference			
28	Sounds nice	Thank you for your answer	20	
	A: third preference			
29	Sounds nice	Thank you for your answer	20	
	A: lowest preference			
30	Sounds nice	Thank you for your answer	20	Wrong A cat
	Q: If I could influence the educational policies of the public schools of some city, I would try to increase the practical value of courses.		**68**	Wrong Q cat
	A: highest preference			
31	Good to hear that	Got it! Thanks for sharing	17	
	A: second preference			
32	Good to hear that	Got it! Thanks for sharing	17	
	A: third preference			
33	Good to hear that	Got it! Thanks for sharing	17	
	A: lowest preference			
34	Good to hear that	Got it! Thanks for sharing	17	Wrong A cat

#	Auto generated	After manual correction	Edit	Cause
	Q: I would prefer a friend who is practical, efficient, and hard working.		**20**	Wrong Q cat
	A: highest preference			
35	Great!	Got it	5	
	A: second preference			
36	Great!	Got it	5	
	A: third preference			
37	Great!	Got it	5	
	A: lowest preference			
38	Great!	Got it	5	Wrong A cat

Fitness survey - 21 of 40 (52.5%) reactions corrected

#	Auto generated	After manual correction	Edit	Cause
	Q: How often do you workout?		**54**	Wrong Q cat
	A: Not at all			
1	Great!	That's hard to hear	17	
	A: 2-3 times a week			
2	Got it! Thanks for sharing	That's good	20	Wrong A cat
	A: Every day			
3	Thanks for sharing that	That's good	17	
	Q: How healthy is the food you eat?		**47**	
	A: Not sure: I eat whatever is in front of me.			
4	Noted	I am sorry to hear that	20	Wrong A cat
	A: Okay: I count calories but I'm not too strict.			
5	That sounds stressful	Noted	18	Wrong A cat
	A: Excellent: I've adopted the perfect plate and feel great.			
6	Noted	Sounds nice	9	Wrong A cat

#	Auto generated	After manual correction	Edit	Cause
	Q: Do you know what the perfect plate is?		**65**	Wrong Q cat
	A: Yes: it balances food intake for I never use it.			
7	Thanks for letting me know	Got it	23	Wrong A cat
	A: Yes: 1/2 vegetables, 1/4 protein & 1/4 carbohydrates. I use it when I remember.			
8	Thanks for letting me know	Good to hear that	21	
	A: Yes: It sums up my approach to every meal.			
9	Thanks for letting me know	Good to hear that	21	
	Q: How hard do you work out?		**71**	Wrong Q cat
	A: Not very hard, it depends on the mood I'm in.			
10	Good to hear that	So sorry about that	12	
	A: I start hard but usually tail off part of the way through.			
11	That's hard to hear	Sure	17	Wrong A cat
	A: I make sure to feel the burn by the end.			
12	Thank you for your answer	Okay, that's good	21	Wrong A cat
	A: Hard enough to ensure I have given it my all. By the end I can barely stand.			
13	Thank you for your answer	Okay, that's good	21	Wrong A cat
	Q: Do you have a workout buddy?		**21**	
	A: Yes but I only see them once every month or two.			
14	I am sorry to hear that	Got it! Thanks for sharing	21	Wrong A cat
	Q: How do you feel after a workout?		**39**	
	A: Relieved that I made it through the session.			
15	Thanks for letting me know	That sounds stressful	21	Wrong A cat
	A: Shattered like I just pushed myself to the limit and maybe passed it.			
16	Thanks for sharing that	Got it	19	Wrong A cat

#	Auto generated	After manual correction	Edit	Cause
	Q: Do you have an exercise plan?		**63**	
	A: Yes but I don't always stick to it.			
17	Great!	Thank you for your answer	23	Wrong A cat
	A: Yes, it's designed to work all muscle groups with alternating exercises.			
18	Okay, I'm getting a better idea of your answers	That sounds positive	40	Wrong A cat
	Q: What would you say is your current level of fitness?		**59**	Wrong Q cat
	A: Pretty poor; I get out of breath just walking up the stairs.			
19	Sure	So sorry about that	17	
	A: Good; I exercise regularly and watch what I eat.			
20	Sure	I am glad to hear that	21	
	A: Excellent; I'm always at the gym and avoid all foods that are bad.			
21	Sure	I am glad to hear that	21	

Finance survey - 28 of 84 (33.3%) reactions corrected

#	Auto generated	After manual correction	Edit	Cause
	Q: As it relates to matters of personal finance, what topics do you feel you could use more information on?		**224**	Wrong Q cat
	A: Budgeting			
1	I am sorry to hear that	Thanks for letting me know	20	Wrong A cat
	A: Credit			
2	I am sorry to hear that	Thanks for letting me know	20	Wrong A cat
	A: Wills			
3	I am sorry to hear that	Thanks for letting me know	20	Wrong A cat
	A: Life Insurance			
4	I am sorry to hear that	Thanks for letting me know	20	Wrong A cat

	A: Disability Insurance			
5	I am sorry to hear that	Thanks for letting me know	20	Wrong A cat
	A: Health Insurance			
6	I am sorry to hear that	Thanks for letting me know	20	Wrong A cat
	A: Long Term Care Insurance			
7	I am happy that's the case	Thanks for letting me know	21	Wrong A cat
	A: Loans/Debt			
8	I am sorry to hear that	Thanks for letting me know	20	Wrong A cat
	A: Saving			
9	I am happy that's the case	Thanks for letting me know	21	Wrong A cat
	A: Investing			
10	I am happy that's the case	Thanks for letting me know	21	Wrong A cat
	A: Other (please specify)			
11	I am happy that's the case	Thanks for letting me know	21	Wrong A cat
	Q: I feel in control of my current financial situation.		**56**	Wrong Q cat
	A: Not at all true			
12	Sounds good	That is frustrating	17	
	A: Somewhat untrue			
13	Sounds good	That is frustrating	17	
	A: Somewhat true			
14	That's hard to hear	That's good	11	
	A: Very true			
15	That's hard to hear	That's good	11	
	Q: I feel capable of handling my financial future overall.		**70**	Wrong Q cat
	A: Not at all true			
16	I am glad to hear that	That is frustrating	17	
	A: Somewhat untrue			
17	I am glad to hear that	That is frustrating	17	
	A: Somewhat true			
18	That sounds stressful	Okay, that's good	18	
	A: Very true			
19	That sounds stressful	Okay, that's good	18	

#	Auto generated	After manual correction	Edit	Cause
	Q: I have the following types of insurance.		**99**	Wrong Q cat
	A: Life			
20	That is frustrating	Thank you for your answer	16	Wrong A cat
	A: Health			
21	That is frustrating	Thank you for your answer	16	Wrong A cat
	A: Auto			
22	That is frustrating	Thank you for your answer	16	Wrong A cat
	A: Homeowner's/Renter's			
23	That is frustrating	Thank you for your answer	16	Wrong A cat
	A: Disability			
24	That is frustrating	Thank you for your answer	16	Wrong A cat
	A: Long Term Care			
25	I am glad to hear that	Thank you for your answer	19	Wrong A cat
	Q: How necessary or important do you feel it is for you to work with a financial advisor?		**44**	Wrong Q cat
	A: Not important			
26	That sounds stressful	Got it	18	
	A: Somewhat unimportant			
27	That sounds stressful	Got it	18	
	A: Very important			
28	Sounds nice	Got it	8	