



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Exploring the Design of Conversational Agents for Mental Health Self-Monitoring

Robert Jacob Bowman

School of Computer Science and Statistics
Trinity College Dublin

A thesis presented for the degree of:
Doctor of Philosophy

March 2, 2025

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

A handwritten signature in black ink that reads "Robert Bowman". The signature is written in a cursive style with a large, prominent 'R' and 'B'.

Robert Jacob Bowman

March 2, 2025

Abstract

An enduring aspiration is to use conversational agents (CAs) to provide mental health support (MHS) in order to tackle the persistent mental health treatment gap. CAs are computer interfaces that interact with people in a manner resembling human conversation. They are an alluring means of providing MHS as they can be designed to provide human-like support in increasingly accessible, engaging, and personalised ways. The aspiration of using CAs to provide MHS is starting to be realised with CAs available today as smartphone applications.

We need to understand how to design effective MHS CA systems so as to realise the potential to provide people with much needed support. If we do not learn how to design these systems effectively we risk missing out on the potential benefits and, worse, creating systems that cause harm. The design of CAs for MHS is a complex task with both CAs and MHS being innately complex. CAs are complex because of their human-likeness. Our understanding of how people interact with and desire to interact with these human-like computer interfaces is nascent. Mental health support is complex because people have diverse perspectives on the nature of mental health and how people should be supported with it. To navigate the complex design of CAs for MHS we need to develop understanding of the ways they could be designed to support people with their mental health.

To develop understanding of how CAs could be designed to provide MHS, this thesis examines the design of CAs for the MHS activity of mental health self-monitoring (MHSM). MHSM is used to support people experiencing a range of mental health conditions including depression, anxiety, and bipolar disorder. It is however an innately challenging activity that many people struggle to do. CAs appear a promising technology to support people with MHSM and its challenging nature. If CAs can be designed to support people to do MHSM more effectively, they could benefit the mental health of many people.

This thesis presents a series of three stands of work. In an effort to consider advanced utilisations of CAs for MHSM we first propose a plausible vision of future systems which we term pervasive therapy. This vision describes the general design of CA systems for the delivery of ecological momentary interventions (in the moment mental health support) as a component of blended care support (support combining human and digital support). We use this vision as a basis for discussion of three key design challenges: 1) intervention and conversation planning, 2) CA interaction design, and 3) clinician involvement. Furthermore we use this vision to discuss the ethical concerns and risks of pervasive therapy systems.

Next this thesis examines the challenge of CA interaction design for MHSM. In order to conceptualise and explore the design of these interactions we draw on pertinent theories of human-human interaction, namely joint action theory and politeness theory. Building upon this we conducted a mixed-methods interview-based study ($n=39$) to investigate: 1) how the use of politeness by CA impacts the user experience, and 2) people's perspectives on the use of CAs for MHSM more broadly. This study demonstrates that the use of politeness by CAs for MHSM can both positively and negatively impact the user experience. Our analysis of participants' perspectives makes apparent conflicts between the perceived

opportunities and risks of CAs for MHSM. This study raises questions about the ways in which it is appropriate and desirable for CAs to support people with their mental health.

This thesis goes on to investigate the ways CAs could be designed to support people with MHSM. We draw on human-human social support theory in order to conceptualise and explore the design of CA support for MHSM. We conducted a qualitative interview study (N=24) building on this theory-based work to investigate people's perspectives on the ways CAs could be designed to support people with MHSM. Our interview design asked participants to consider the use of MHSM CAs by both themselves and a person close to them (e.g., partner, sibling) which enabled us to develop a nuanced understanding of their perspectives. We argue that participants reject CA support that they view as replacing human support, but are receptive to CA support that they view as augmenting or facilitating human support.

To conclude, this thesis reflects on what has been learnt about the design of CAs for MHSM. This thesis ultimately argues that to design CA for MHSM responsibly we should move from dyadic views of individual users and CAs to a community view of users and CAs as part of larger support networks.

Associated Publications

The following publications present work that this thesis also presents:

1. **Robert Bowman**, Orla Cooney, Anja Thieme, Benjamin R Cowan, and Gavin Doherty. Pervasive Therapy: Designing Conversation-Based Interfaces for Ecological Momentary Intervention. *IEEE Pervasive Computing*, 21(4:55-63, 2022. <https://doi.org/10.1109/MPRV.2022.3209068>
2. **Robert Bowman**, Benjamin R Cowan, Anja Thieme, and Gavin Doherty, Beyond Subservience: Using Joint Commitment to Enable Proactive CUIs for Mood Logging. In *Proceedings of the 4th Conference on Conversational User Interfaces (CUI '22)*, pages 1-6, 2022. <https://doi.org/10.1145/3543829.3544512>
3. **Robert Bowman**, Orla Cooney, Joseph W Newbold, Anja Thieme, Leigh Clark, Gavin Doherty, and Benjamin R Cowan. Exploring How Politeness Impacts the User Experience of Chatbots for Mental Health Support. *International Journal of Human-Computer Studies*, page 103181, 2024. <https://doi.org/10.1016/j.ijhcs.2023.103181>

The following publication presents work that was conducted alongside the thesis:

4. **Robert Bowman**, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. Using Thematic Analysis in Healthcare HCI at CHI: A Scoping Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, pages 1-18, 2023. <https://doi.org/10.1145/3544548.3581203>

Work from this thesis is reported in two further papers that are currently unpublished:

5. **Robert Bowman**, Anja Thieme, Benjamin Cowan, Gavin Doherty. Exploring the Design of Conversational Agents for Mental Health Self-Monitoring: A Social Support Framework
6. (Under Review) **Robert Bowman**, Anja Thieme, Benjamin Cowan, Gavin Doherty. Mental Health Management as a Social Endeavour: Challenges and Opportunities for Conversational Agent Design

Acknowledgements

This research was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally Enhanced Reality (d-real, grant 18/CRT/6224) and Microsoft Research through its PhD Scholarship Programme.

I would like to thank my supervisor, Gavin Doherty, as well as my additional supervisors, Benjamin Cowan and Anja Thieme, for their exceptional support. Your guidance and encouragement were critical to the accomplishment of this thesis.

I would also like to thank my colleagues in the Health Technology Design research group and wider Irish HCI community. In particular, thank you to Orla Cooney, Camille Nadal, Séamus Ryan, Anna-Lisa Mann, Jacinta Jardine, Johanna Didion, Kellie Morrissey, and Doireann Peelo Dennehy.

Thank you to my parents, Angela and Mike Bowman. I could not have achieved this without your immense and enduring support.

Thank you to my partner, Lauren Clarke, for your companionship throughout this journey. I am grateful for your patience and encouragement.

Robert Bowman

Trinity College Dublin

March 2, 2025

Contents

List of Tables	14
List of Figures	16
1 Introduction	17
1.1 Motivation	17
1.1.1 Mental Health Self-Monitoring	17
1.1.2 Conversational Agents for Mental Health Support	18
1.1.3 Conversational Agents for Mental Health Self-Monitoring	19
1.2 Research Questions	19
1.3 Contributions	20
1.4 Thesis Structure	20
2 Background: Mental Health Self-Monitoring	24
2.1 Introduction	24
2.2 Collecting Data	25
2.2.1 Mental Health Data	25
2.2.2 Manual Self-monitoring	25
2.2.3 Recording Mental Experience	26
2.2.4 Data Validity	27
2.2.5 Prompts and Reminders	29
2.2.6 Passive and Semi-automated Monitoring	29
2.3 Reflecting on Data	30
2.3.1 From Description to Actionable Insight	30
2.3.2 Analytical Challenge	31
2.3.3 Emotional Challenge	31
2.3.4 Analysis Support	32
2.4 Intervention	32
2.5 Trajectories	33
2.5.1 Initiating Self-monitoring	33
2.5.2 Developing Self-monitoring Practices	33
2.5.3 Phases of Monitoring	34
2.5.4 Lapses	35
2.6 Support Relationships	35

2.6.1	Healthcare Professionals	35
2.6.2	Friends and Family	37
2.6.3	Relationship Changes	39
2.6.4	Supportive Tools	39
2.7	Conclusion	40
3	Background: Conversational Agents	41
3.1	Introduction	41
3.2	Technical Implementation	42
3.2.1	User Input	42
3.2.2	CA Output	43
3.3	Self-Disclosure to CAs	43
3.3.1	CAs Versus People	43
3.3.2	CAs Versus Forms	44
3.3.3	CAs for Journaling	45
3.3.4	Reciprocal Self-Disclosure by CAs	46
3.3.5	CAs as Mediators for Self-Disclosure to People	47
3.3.6	CAs for Mental Health Assessment	47
3.3.7	Speech Interaction for Disclosure	48
3.3.8	CAs for Ethnography	48
3.3.9	Summary	49
3.4	Relational Agents	49
3.4.1	Trustworthy Agents	50
3.4.2	Therapeutic Bond	50
3.4.3	Long-term Use	51
3.4.4	Mobile Use	52
3.4.5	Mental Health Support	53
3.4.6	Summary	53
3.5	Therapy Agents	53
3.5.1	Design	53
3.5.2	Clinical Evaluation	54
3.5.3	User Experience Evaluation	55
3.5.4	Pocket Skills	56
3.5.5	Summary	57
3.6	Social Agents	57
3.7	Social Support from CAs	58
3.8	Human-CA Relationship Development	59
3.9	Using Theories of Human-Human Interaction	60
3.10	Ethics and Safety	60
3.10.1	Insensitive and Harmful Interactions	60
3.10.2	Dependency	61
3.10.3	Negative Social Impacts	61
3.10.4	Stigma	61

3.11 Conclusion	62
4 Exploration: Considering the Design Challenges of Future Systems	63
4.1 Introduction	63
4.2 Pervasive Therapy Systems	64
4.3 Intervention and Conversation Planning	67
4.4 CA Interaction Design	68
4.4.1 Initiation Prompts	68
4.4.2 Delivery Modalities	69
4.4.3 Conversation Design	69
4.4.4 Human-like Interaction	69
4.4.5 Patient Agency	70
4.5 Clinician Involvement	70
4.5.1 Configuration	72
4.5.2 Monitoring	72
4.5.3 Adjustment	73
4.5.4 Communication	73
4.6 Risks and Ethical Concerns	74
4.7 Conclusion	76
5 Theory-based Exploration of CA Interaction Design	77
5.1 Introduction	77
5.2 Joint Action	78
5.2.1 Designing to Proactively Engage Users	78
5.2.2 Joint Action Theory	79
5.2.3 Designing for Joint Commitment	81
5.3 Politeness	84
5.3.1 Encouraging Willingness Using Politeness	84
5.3.2 Politeness Theory	85
5.3.3 Recommendation of Politeness for CA Interaction Design	86
5.3.4 Applications of Politeness Theory to CA Interaction Design	87
5.3.5 Designing Polite CAs for MHSM	89
5.4 Discussion	91
5.4.1 Joint Action as Framework for CA Interaction Design	91
5.4.2 From Human-Human to Human-CA Interaction	93
5.5 Conclusion	93
6 Interview Study: Investigating How Politeness Impacts the User Experience	95
6.1 Introduction	95
6.2 Methodology	97
6.2.1 Participants	97
6.2.2 Experiment Design	97

6.2.3	Interview Design	99
6.2.4	Procedure	100
6.2.5	Analysis Approach	101
6.3	Quantitative Analysis	102
6.3.1	Manipulation Check - CAs Using Politeness are Perceived as More Polite	102
6.3.2	General Agent Rating Items - Personal Politeness Is Favourably Perceived	102
6.4	Qualitative Analysis	105
6.4.1	User Perceptions of CAs using Politeness	105
6.4.2	Using CAs for MHSM	108
6.5	Discussion	111
6.5.1	Politeness Can Positively and Negatively Impact the User Experience	111
6.5.2	Face Threatening Acts in CA Interaction	113
6.5.3	Tensions in Using CAs for MHSM	115
6.5.4	Future Directions: Towards Polite CAs	116
6.5.5	Ethical Implications	117
6.5.6	Limitations	118
6.6	Conclusion	119
7	Theory-based Exploration of CA Support Types	120
7.1	Introduction	120
7.2	Social Support	121
7.3	Support Exploration	124
7.3.1	Appraisal Support	125
7.3.2	Informational Support	127
7.3.3	Instrumental Support	129
7.3.4	Emotional Support	129
7.4	Support Combinations	130
7.5	Risks	132
7.5.1	Risks of Appraisal Support	132
7.5.2	Risks of Informational Support	134
7.5.3	Risks of Instrumental Support	134
7.5.4	Risks of Emotional Support	135
7.5.5	Risks to Support Ecology	135
7.6	Discussion	136
7.6.1	Framework for Considering Supportive CA Design	136
7.6.2	Design Considerations	136
7.6.3	Limitations and Future Work	138
7.7	Conclusion	139

8	Interview Study: Investigating Perspectives on CA Support Types	141
8.1	Introduction	141
8.2	Methodology	142
8.2.1	Participants	143
8.2.2	Interview Design	143
8.2.3	Design Cards	147
8.2.4	Analysis Approach	149
8.3	Analysis	150
8.3.1	Theme 1: Using CAs is Abnormal	150
8.3.2	Theme 2: CAs Could be Harmful Coping Mechanisms	152
8.3.3	Theme 3: CA Support Threatens Human Support	153
8.3.4	Theme 4: CAs Could Help with the Support Burden	154
8.3.5	Theme 5: CAs Could Fill Support Gaps	156
8.3.6	Theme 6: CAs Could Support People with Human Support	157
8.4	Discussion	159
8.4.1	A Social Approach to the Design of CAs for MHSM	159
8.4.2	Challenges & Opportunities for Effective and Appropriate CA Support in Mental Health	160
8.4.3	Reflections, Limitations, and Future Work	163
8.5	Conclusion	164
9	Discussion	166
9.1	Introduction	166
9.2	Research Questions and Contributions	166
9.2.1	The Design Space of CAs for MHSM	167
9.2.2	Utilising Theories of Human-Human Interaction	168
9.2.3	Perspectives on CAs for MHSM	169
9.2.4	Towards the Social Design of MHSM CAs	170
9.3	Future Work	170
9.3.1	Design Exploration	170
9.3.2	Theory	171
9.3.3	Perspectives	172
9.3.4	Evaluation	173
10	Conclusion	175
A	Further Quant. Reporting	177
A.1	Modelling	177
A.2	Polite	178
A.3	Easy	179
A.4	Express	180
A.5	Natural	181
A.6	Care	182

<i>CONTENTS</i>	13
A.7 Relationship	183
A.8 Trust	184
A.9 Continue	185
A.10 Like	186
A.11 Satisfied	187
B Interview Study: Support - Design Cards	188
Bibliography	190

List of Tables

4.1	Forms of clinician involvement in an AI-delivered, human-supervised Per- vasive Therapy system.	71
5.1	Key Definitions for Joint Action and Joint Commitment	80
5.2	Example Project Proposals	81
5.3	Example Pre-sequences	83
5.4	Example Alterations	84
5.5	Base MHSM Queries	90
5.6	MHSM CA Prompt Sequences Informed by Politeness	92
6.1	Experiment Scenarios	99
6.2	The General Agent Rating (GAR) items	100
6.3	Politeness LMEM tests compared against Direct.	103
6.4	Descriptive Statistics for Politeness rating and GAR items (Mean (SD)). . .	103
6.5	LMEM tests Personal compared to Direct	104
6.6	LMEM tests Personal compared to Passive	105
7.1	Social Support Definitions and Examples	122
7.2	Social Support Categorisations	122
7.3	Tabulation of Support Types and Levels.	125
7.4	Support Type Risks	133
7.5	Support Exploration Design Considerations	139
8.1	Summary of Themes	150
A.1	Polite: Direct Intercept - Fixed Effects	178
A.2	Polite: Direct Intercept - Random Effects	178
A.3	Polite: Passive Intercept - Fixed Effects	178
A.4	Polite: Passive Intercept - Random Effects	178
A.5	Easy: Direct Intercept - Fixed Effects	179
A.6	Easy: Direct Intercept - Random Effects	179
A.7	Easy: Passive Intercept - Fixed Effect	179
A.8	Easy: Passive Intercept - Random Effects	179
A.9	Express: Direct Intercept - Fixed Effects	180
A.10	Express: Direct Intercept- Random Effects	180

A.11 Express: Passive Intercept - Fixed Effects	180
A.12 Express: Passive Intercept- Random Effects	180
A.13 Natural: Direct Intercept - Fixed Effects	181
A.14 Natural: Direct Intercept - Random Effects	181
A.15 Natural: Passive Intercept - Fixed Effects	181
A.16 Natural: Passive Intercept- Random Effects	181
A.17 Care: Direct Intercept - Fixed Effects	182
A.18 Care: Direct Intercept - Random Effects	182
A.19 Care: Passive Intercept- Fixed Effects	182
A.20 Care: Passive Intercept- Random Effects	182
A.21 Relationship: Direct Intercept - Fixed Effects	183
A.22 Relationship: Direct Intercept - Random Effects	183
A.23 Relationship: Passive Intercept - Fixed Effects	183
A.24 Relationship: Passive Intercept - Random Effects	183
A.25 Trust: Direct Intercept - Fixed Effects	184
A.26 Trust: Direct Intercept- Random Effects	184
A.27 Trust: Passive Intercept - Fixed Effects	184
A.28 Trust: Passive Intercept- Random Effects	184
A.29 Continue: Direct Intercept - Fixed Effects	185
A.30 Continue: Passive Intercept - Random Effects	185
A.31 Continue: Passive Intercept - Fixed Effects	185
A.32 Continue: Passive Intercept - Random Effects	185
A.33 Like: Direct Intercept - Fixed Effects	186
A.34 Like: Direct Intercept - Random Effects	186
A.35 Like: Passive Intercept- Fixed Effects	186
A.36 Like: Passive Intercept - Random Effects	186
A.37 Satisfied: Direct Intercept - Fixed Effects	187
A.38 Satisfied: Passive Intercept - Random Effects	187
A.39 Satisfied: Passive Intercept - Fixed Effects	187
A.40 Satisfied: Passive Intercept - Random Effects	187

List of Figures

4.1	Conceptual Architecture of a Pervasive Therapy System.	66
6.1	Image of CA Prototype	98
7.1	Example Support Combinations	131
8.1	CA Design Cards: MoodBot and Monitoring Widget	145
8.2	CA Design Cards: Alex 1 and 2.	146
B.1	Early Iteration of CA Design Cards - A	188
B.2	Early Iteration of CA Design Cards - B	189

Chapter 1

Introduction

1.1 Motivation

1.1.1 Mental Health Self-Monitoring

A mental health treatment gap persists whereby many people experience mental health difficulties yet do not receive sufficient support (World Health Organisation, 2022). This gap is understood to be the product of both attitudinal and structural barriers whereby people choose not to (e.g., because of stigma) or are unable to (e.g., because of cost) access support (Andrade et al., 2014; World Health Organisation, 2022). In order to address this treatment gap we need to develop new approaches for mental health support that overcome these access barriers. It has long been recognised that there is potential to use technology to not only make mental health support more accessible, but also more engaging and effective (Coyle et al., 2007; Doherty et al., 2010; Sanches et al., 2019).

Mental health self-monitoring (MHSM) is an activity that technology could be designed to support people with in order to help them with their mental health. MHSM involves monitoring a person’s mental experiences (i.e., thoughts, feelings, mood) and related factors (e.g., exercise, sleep, nutrition). It is used to support people experiencing a range of mental health conditions including depression, anxiety, and bipolar disorder (Matthews and Doherty, 2011). MHSM monitoring is a challenging activity that people often struggle to do. It requires the individual to regularly consider and record their mental experience and related factors which can be burdensome and emotionally challenging (Matthews and Doherty, 2011; Matthews et al., 2017). Prior research has investigated how digital tools could be designed to help people overcome the challenging nature of MHSM (e.g., Matthews and Doherty, 2011; Matthews et al., 2017; Murnane et al., 2016, 2018; Schueller et al., 2021). Digital tools, for example, can be designed to proactively engage users to help them maintain regular monitoring practices (Murnane et al., 2016). What is more, digital tools for MHSM could be used to not only assess the person’s mental health, but also to provide support to promote it during everyday life (Balaskas et al., 2021; Heron and Smyth, 2010; Schueller et al., 2017). MHSM can negatively impact a person’s well-being and therefore it is imperative that tools for MHSM are carefully designed. MHSM can be reactive (e.g., identifying a low mood can exacerbate it (Doherty et al., 2020;

Shiffman et al., 2008)) and it can be used to facilitate harmful behaviours (e.g., eating disorders (Kelley et al., 2017)).

1.1.2 Conversational Agents for Mental Health Support

A prominent vision of mental health technology is the use of conversational agents (CAs) to provide people with mental health support (Abd-Alrazaq et al., 2019, 2020, 2021). CAs are computer interfaces that interact with people in a manner resembling human conversation (Bickmore and O’Leary, 2023). The resemblance of CA interaction to human conversation means CAs can be designed to provide human-like support. As CAs are computer interfaces, this human-like support could be made widely accessible. An early CA was the ELIZA system of the 1960’s which was programmed to mimic a Rogerian psychotherapist (Weizenbaum, 1966). Anecdotal accounts report that people were highly receptive to using the ELIZA system and would quickly engage in deep and emotional interactions with it (Weizenbaum, 1976). In response to ELIZA, speculation sparked about the potential to widely deploy systems like it making mental health support more available in order to address the mental health treatment gap of the time (Colby et al., 1966; Weizenbaum, 1976).

Technology continues to advance and with it visions have advanced for how CAs could be used to support people with their mental health, physical health, and broader wellbeing. For example, in response to the emergence of mobile phones, Bickmore and Giorgino (2006) wrote of how CAs could be deployed using mobile phones to make support available anytime and anywhere. Furthermore, they envisioned systems utilising real-time sensing to provide proactive support at times of need. Visions for how CAs could support us with our mental health have also been contributed by science fiction. The film *Her*, for instance, tells a story of an advanced CA tending to the emotional needs of a lonely human character such that the character develops an intimate and human-like relationship with the CA (Jonze, 2013). CA technology has advanced rapidly over the last decade with CAs becoming an increasing common form of human-computer interaction (Brandtzaeg and Følstad, 2018; Følstad and Brandtzaeg, 2017). Visions of CAs for mental health support are starting to be realised with CAs such as Woebot, Wysa, and Replika widely available today as smartphone applications.

Although there is potential for CAs to support people with their mental health, there is also potential for CAs to negatively impact people’s mental health. CAs may fail to respond appropriately to users and instead respond in way that could cause users further distress or harm (Bickmore et al., 2018b,c; Miner et al., 2016). Incidents of existing CAs for mental health support responding inappropriately to users have already been reported in the media (Browne, 2022; White, 2018). Furthermore, there are concerns about the long-term and social impacts of CAs for mental health support. In particular, there are concerns about how the option of human-like support from computers may impact human-human support (Fiske et al., 2019; Miller and Polson, 2019; Morris et al., 2018).

1.1.3 Conversational Agents for Mental Health Self-Monitoring

CAs could be an effective technology for supporting with MHSM. Shiffman et al. (2008) describe how tools for MHSM could be like a “therapist in your pocket” (Shiffman et al., 2008, p. 18). With CAs, this metaphor could be even more apt. CAs could be designed to engage users in a continuous MHSM dialogue. Extending the opportunity of digital tools for MHSM, CAs could also support people with MHSM by providing human-like encouragement, comfort, and empathy (Bickmore and Giorgino, 2006).

How to design effective CAs for MHSM is unclear. The human-likeness of CAs makes them alluring for use to support people with their mental health, however it also makes their design complex. It is recognised that the human-likeness of CAs can lead to user expectations of a system that are misaligned with actual system capabilities which can hinder interaction (Cowan et al., 2017; Luger and Sellen, 2016). Furthermore it is unclear in what ways and what situations it is appropriate for CAs to behave in human-like ways (Clark et al., 2019b; Rapp et al., 2021). Indeed, our current understanding of how to design effective CAs is very limited with there being a little theoretical and design understanding of CA interaction (Clark et al., 2019a; Rapp et al., 2021).

To harness the positive potential of CAs to support people with MHSM, we need to understand how to design them effectively for this application. This requires us to navigate the complex human-likeness of CAs in order to understand how people could and want to be supported by human-like computers with their mental health. If we do not learn how to design CAs for MHSM effectively we risk missing out on the potential benefits, and worse, deploying systems that cause harm.

1.2 Research Questions

The aim of this thesis is to develop understanding of how to design effective CAs for MHSM. Accordingly, it examines the main research question (RQ) of:

Thesis RQ: How should CAs be designed to support people with MHSM?

The main research question is addressed through investigation of the three sub research questions (A, B, and C):

RQ-A: How can CAs be designed to support people with MHSM?

RQ-B: How can theories of human-human interaction inform the design of CAs for MHSM?

RQ-C: What are people’s perspectives on CAs for MHSM?

The approach of this thesis is breadth rather than depth focused. This means that this thesis explores a range of pertinent design aspects instead of investigating a single aspect more exhaustively. This approach was taken in order to develop understanding of the range of design opportunities and challenges of CAs for MHSM. By developing this understanding, this thesis aims to identify which design aspects are most important and warrant more in depth enquiry.

1.3 Contributions

This thesis contributes to the field of human-computer interaction (HCI). It contributes understanding of both how to design digital tools for mental health self-monitoring and how to design CAs. Principally, this thesis contributes understanding of how to design CAs for MHSM. The use of CAs to support people with their mental health is an enduring vision that is increasingly being realised. The contributions of this thesis will enable the creators of these systems to utilise the potential of CAs for MHSM.

Design Exploration Addressing RQ-A, this thesis contributes understanding of the ways CAs could be designed to support people with MHSM. It explores the design of CA interactions and support. It considers different ways CAs could support people alongside receiving support from human mental health professionals. Additionally, it considers different support roles CAs could take alongside human support sources including friends and family.

CA Theory Addressing RQ-B, this thesis investigates how theories of human-human interaction can inform the design of CAs for MHSM. Joint action theory (Clark, 1996), politeness theory (Brown and Levinson, 1987), and social support theory (House, 1981; House et al., 1985) are used to explore the different ways CAs could be designed. These theories are used as a theoretical basis for studies with participants. This thesis advances understanding of how these theories of human-human interaction relate to human-CA interaction and thus our theoretical understanding of CA interaction.

Understanding of Perspectives Addressing RQ-C, this thesis advances understanding of people's perspectives on the use of CAs for MHSM and mental health support more generally. This understanding was developed through two interview based studies (n=39, n=24). The understanding this thesis contributes challenges existing literature by making apparent tensions within participant's perspectives on the ways in which it is appropriate for CAs to support people.

New Direction Addressing its main research question, this thesis argues that to design CAs to effectively support people with MHSM they should be designed for being components of social ecologies of support. This thesis contributes understanding of the opportunities and challenges of designing CAs this way.

1.4 Thesis Structure

The organisation of the remaining chapters of this thesis is as follows:

Chapter 2 – Background: Mental Health Self-Monitoring Chapter 2 (p.24) is the first of two literature review chapters. It characterises the activity of mental health self-monitoring from a human-computer interaction perspective. This chapter describes

how MHSM is an intrinsically challenging activity that people often struggle to do. The opportunities of digital tools to support people are discussed and the design implications of prior work are highlighted. This chapter concludes by emphasising the opportunity for digital tools to facilitate effective MHSM by engaging proactively, flexibly, and encouragingly.

Chapter 3 – Background: Conversational Agents Chapter 3 (p.41) is the second of the two literature review chapters of this thesis. This chapter reviews research on CAs relevant to their design for MHSM. It discusses studies of CAs designed to support people with their health and mental health. Topics including disclosure to CAs, social support from CAs, and human-CA relationships are discussed. Concerns raised by this research about CAs for MHSM are emphasised. Additionally, this chapter discusses the opportunity demonstrated by research considered within this chapter to utilise theories of human-human interaction as a starting point for researching the design of CAs. This chapter summarises that CAs are an alluring technology to use to support people with their mental health but it is unclear how to design CAs effectively for applications like MHSM.

Chapter 4 – Exploration: Considering the Design Challenges of Future Systems Chapter 4 (p.63) proposes a vision of CAs for MHSM. Informed by our characterisation of MHSM (Chapter 2) and review of CAs for MHSM (Chapter 3), we envision the design of future CA systems for ecological momentary intervention (an advanced form of MHSM) which we term pervasive therapy systems. It considers the open challenges of designing pervasive therapy systems highlighting three major areas: 1) intervention and conversation planning, 2) CA interaction design, and 3) clinician involvement. Furthermore, it discusses the risks and ethical concerns. Through envisioning future MHSM CAs this chapter contributes to RQ-A. This chapter is based upon work published in the IEEE Pervasive Computing journal (Bowman et al., 2022a).

Chapter 5 – Theory-based Exploration of CA Interaction Design Chapter 5 (p.77) explores CA interaction design for MHSM. It utilises two theories of human-human interaction: 1) joint action theory (Clark, 1996), and 2) politeness theory (Brown and Levinson, 1987). First this chapter shows how these theories can be a useful basis for conceptualising human-CA interaction, particularly for MHSM. This chapter next demonstrates how these theories can inform designing for proactive, flexible, and encouraging MHSM CA interactions. By exploring how CAs can be designed to support MHSM using theories of human-human interaction, this chapter contributes to both RQ-A and RQ-B. This material is based on upon work published in the ACM Conversational User Interfaces (CUI) 2022 conference (Bowman et al., 2022b) and the International Journal of Human-Computer Studies (Bowman et al., 2024).

Chapter 6 – Interview Study: Investigating How Politeness Impacts the User Experience Chapter 6 (p.95) presents a mixed-methods interview-based study (N=39).

First, building on the work of Chapter 5, this study investigated how the use of politeness by CAs for MHSM impacts the user experience. The study included a within-participants experiment that involved participants interacting with three CAs that used different politeness strategies. It is demonstrated that politeness can both positively and negatively impact the user experience and therefore is important for CA interaction design. This part of the study addresses RQ-B. Second, this study investigated participant's perspectives on CAs for MHSM using semi-structured interviews that followed the experiment. Our analysis makes apparent tensions in participant's perspectives about the ways in which it is appropriate for CAs to support people with MHSM. RQ-C is particularly addressed by this component of the study. This chapter is also based upon work published in the *International Journal of Human-Computer Studies* (Bowman et al., 2024).

Chapter 7 – Theory-based Exploration of CA Support Types Chapter 6 raised questions about the ways in which it is appropriate for CAs to support people with MHSM. This chapter (7, p.120) explores the different ways CA could be designed to support people with MHSM. A theory of human-human social support is used as a framework for considering the different forms of MHSM support that CAs could provide. This framework centres on a tabulation of types and levels of CA support for MHSM. Implications for the design and evaluation of CAs for MHSM are identified based on this framework. In particular, this work argues that there is a need to increasingly consider CAs for MHSM as part of social ecologies of support involving human supporters such as healthcare professionals, friends, and family. As this chapter, like Chapter 5, uses a theory of human-human interaction to explore how CAs can be designed to support people with MHSM, it addresses both RQ-A and RQ-B.

Chapter 8 – Interview Study: Investigating Perspectives on CA Support Types Chapter 8 (p.141) presents a qualitative interview study of people's perspectives on the use of CAs for MHSM (N=24). Motivated by the tensions the previous interview study made apparent (Chapter 6), this study aimed to develop a more nuanced and in-depth understanding of people's perspectives on the ways in which it is appropriate for CAs to support people with their mental health. The study design was informed by my work drawing on social support theory of Chapter 7. In order to start considering CAs as part of social ecologies, this study asked participants to consider the use of CAs for MHSM both by themselves and a person close to them (e.g., a partner, friend, or family member). Participants were shown design cards describing different CA designs to facilitate consideration of the different ways CAs could be designed to provide MHSM support. My analysis argues that participants reject CA support that they view as replacing human support, but are receptive to CA support that they view as augmenting or facilitating human support. This work supports and advances the argument of Chapter 7 for the need to increasingly consider CAs as part of larger social ecologies of support.

Chapter 9 – Discussion Chapter 9 (p.166) revisits the research questions of this thesis and reflects on what has been learnt. Principally, this thesis argues that to harness

the potential of CAs for MHSM they should be designed for being components of social ecologies of support. Key directions and considerations for future work are identified.

Chapter 10 – Conclusion Chapter 10 (p.175) concludes this thesis by restating its argument for a social ecologies of support oriented approach to the design of CAs for MHSM.

Chapter 2

Background: Mental Health Self-Monitoring

2.1 Introduction

Chapter 1 introduced mental health self-monitoring (MHSM). To recap, MHSM is a widely used activity for mental health support that involves collecting and reflecting on data relating to a person’s mental health in order to inform its management. MHSM can empower individuals by helping them recognise triggers, early warning signs, and management strategies; ultimately contributing to a sense of control over their life (Ayobi et al., 2017; Matthews et al., 2017; Murnane et al., 2016; Potapov et al., 2021). However, people often find MHSM challenging and struggle to do it effectively. MHSM was originally a paper-based activity, but digital tools are now commonly used (Caldeira et al., 2017; Schueller et al., 2021; Widnall et al., 2020). Indeed, the affordances of digital technology (e.g., smartphones) present many opportunities for designing tools to help people to do MHSM more effectively. There remains opportunity to develop better digital tools to help people with MHSM and in turn with their mental health. In this chapter I characterise in greater depth the activity of MHSM with a focus on the design of digital tools to support people with it. The chapter is organised as follows:

- 1. Collecting Data:** In Section 2.2, I discuss the collection of MHSM data covering what, and how, data is collected.
- 2. Reflecting on Data:** Next, in Section 2.3, I discuss reflecting on collected data which includes consideration of analysis challenges.
- 3. Intervention:** I discuss in Section 2.4, how MHSM interactions can go beyond assessing a person’s mental health to providing interventions to promote it.
- 4. Trajectories:** Section 2.5 concerns MHSM trajectories. This section describes how MHSM can transition through different phases over time in response to a person’s changing support needs.
- 5. Support Relationships:** Lastly, in Section 2.6 I discuss the roles of people beyond

the ‘self’ in MHSM including healthcare professionals and the person’s friends and family.

I conclude this chapter by reiterating opportunities to design digital tools to better support people with the challenges of MHSM.

Credit Statement

This chapter is my own work informed by feedback from my primary supervisor Gavin Doherty.

2.2 Collecting Data

2.2.1 Mental Health Data

MHSM considers the self-monitor’s experience of mental health and things that may be impacting it. Monitoring the experience of mental health is complex as this involves considering the self-monitor’s mental experience (i.e., their thoughts, feelings, and mood). Mood is often a focus of MHSM with terms like *mood monitoring* and *mood logging* being used to describe the activity (Matthews and Doherty, 2011). This focus on a person’s mental experience is argued to distinguish MHSM from self-monitoring more broadly as these tend to focus on people’s behaviours (Schueller et al., 2021).

A focus of MHSM can be the identification of symptoms associated with mental health conditions (e.g., depression and bipolar disorder). The identification and monitoring of symptoms can inform clinical understanding of an individual’s mental health condition and its severity, which can in turn inform their treatment (Kruzan et al., 2023).

A diversity of things can impact an individual’s mental health and can therefore be valuable to self-monitor. Mental health can be impacted by holistic health impacting factors such as exercise, nutrition, menstruation, substance use (e.g., caffeine and alcohol consumption), and sleep. It can also be impacted by stressful (e.g., work or family related difficulties) and relaxing (e.g., pleasurable activities like hobbies and socialising with friends and family) experiences. Furthermore, people may do things that target improving their mental health including the use of medication (e.g., antidepressants), meditation, and therapy (Matthews et al., 2008; Murnane et al., 2016).

2.2.2 Manual Self-monitoring

The central consideration of MHSM is the self-monitor’s mental experience. Monitoring this requires that the individual manually self-monitors their mental state. That is, to collect data on their mental experience, the self-monitor must actively consider and record their experience. The term *manual* monitoring is commonly used to describe monitoring where the self-monitor manually records data (Choe et al., 2017). Although inferences can be made about a person’s thoughts, feelings, and mood from their observable behaviour, it is not currently – if ever – possible to passively (i.e., automatically using digital sensing) monitor a person’s mental experience. Passive sensing can play a role in mental health

self-monitoring through collecting data on the individual's behaviours that can be related to their mental experience. I will discuss passive sensing in Section 2.2.6.

Manual self-monitoring of mental experience can be beneficial for a person's mental health in numerous ways. First, active consideration enables individuals to develop self-awareness of their mental health. This is a key principle of cognitive behavioural therapy (Beck, 2011). Second, self-expression where an individual expresses (e.g., writes about) their experiences is understood to have positive physical and mental health benefits (Pennebaker and Chung, 2011). The activity of recording one's mental experience can thus be therapeutic in itself. Third, the collected data can become a valuable resource for learning about the individual's mental health and how to support it (Kruzan et al., 2023; Matthews and Doherty, 2011).

Manual self-monitoring of one's mental experience can be challenging for the individual. Regularly taking the time to actively consider and record one's mental state is work; it takes time and effort and therefore can be experienced as burdensome. The nature of mental health difficulties can exacerbate this burden (Murnane et al., 2016). It can be emotionally challenging as considering one's mental experience can itself be a challenging mental experience (Matthews et al., 2017; Rooksby et al., 2019). Identifying and recording that one is having a difficult mental experience can worsen that experience (Shiffman et al., 2008).

Due to its challenging nature, people often struggle to effectively manually self-monitor their mental state. Through finding ways to support people with manual MHSM, there is opportunity to support people with their mental health.

2.2.3 Recording Mental Experience

People can use tools to record their mental experiences; this requires deciding how to represent mental experience. Experiences can be represented quantitatively using rating and multiple-choice questions (e.g., rate your mood on a scale of 1-5). Rating items limit the self-monitors ability to express themselves, but it does allow for numerical analysis. An issue with more constrained input is self-monitors can feel that it doesn't accurately or meaningfully represent their experiences (Schueller et al., 2021). This is an issue of both data validity and user experience.

Experiences can also be represented more qualitatively using unconstrained text input (e.g. describe how you are feeling today). In contrast to quantitative representation, unconstrained text input allows for greater self-expression, but requires a more nuanced (qualitative) analysis (Ayobi et al., 2018; Lee and Hong, 2018; Thieme et al., 2015). An opportunity of less constrained input is that it allows for data to be collected, and patterns identified, that may not have been considered important prior. For example, Matthews and Doherty (2011) observed that a therapist was able to identify that their client was experiencing a common feeling of being physically sick through analysis of free-text diary entries. However, a challenge of less constrained input is that it puts greater onus on the self-monitor to decide what to record. To balance the strengths and weaknesses of each, a combination of quantitative and qualitative representation can be used.

Tools need to be flexible to the mental health self-monitoring needs of individuals. Informed by the analysis of a survey of 552 self-monitors living with bipolar disorder, Murnane et al. (2016, p. 483) suggest the design implication of:

Design Implication: “allow customisation”

This is in response to the articulated need to “[support] idiosyncratic circumstances, preferences, and goals, including how individuals’ conditions and management practices and needs evolve over time”. Tools however also need to be simple and ‘user friendly’ allowing the user to make entries easily, particularly when they are experiencing higher levels of symptoms (Murnane et al., 2016). Murnane et al. (2016, p.483), for example, present the participant quote of:

“I used to use a mood app I found on my phone but it was confusing so now I just use the notes section on my iPhone or an actual paper journal.”

In response to issues of available digital tools being too rigid and complex, some people resort to developing their own. For example, Matthews et al. (2017) describe a participant who created her own custom tool using a spreadsheet. People may also resort to using multiple tools to attempt to fulfill the MHSM needs (Ayobi et al., 2020; Murnane et al., 2016). The use of multiple tools may however complicate, or limit, the overall analysis of collected data.

2.2.4 Data Validity

The effectiveness of MHSM is impacted by validity of collected data. That is, how well does the data about the individual’s mental experience represent their mental experience. This is impacted by *which* experiences are recorded and *how* they are recorded.

Recall Validity

People’s recollection of experiences is understood to be inaccurate and systematically biased, despite people often being confident about the accuracy of their own memories (Doherty et al., 2020; Shiffman et al., 2008). Recall is understood to not be ‘recall’ but a construction processes whereby fragments are pieced together under the influence of various cognitive heuristics (Shiffman et al., 2008). For examples of studies demonstrating the inaccuracies of recall and discussion of the cognitive heuristics involved, including from the domain of mental health, see Shiffman et al. (2008) and Doherty et al. (2020). Due to the limits of recall, there is push for self-monitoring to focus on an individual’s experience in the moment.

Ecological Validity

The validity of experience data is further impacted by the context in which it is collected. To record representative data, there is push for self-monitoring to take place in the contexts we want to study (ecological contexts) (Doherty et al., 2020; Shiffman et al., 2008). For mental health self-monitoring, this means aiming to not just collect data in the same, easy,

or safe settings, but across a diversity of settings including those that are challenging for the individual’s mental health.

The validity of experience data can be impacted by the experience of recording the data. This is a consideration of *reactivity* which is the potential for the experience to be affected by the act of assessing it (Korotitsch and Nelson-Gray, 1999; Shiffman et al., 2008). If the experience of recording mental health self-monitoring entries is stressful, or in contrast calming, this may systematically impact the data collected. This is an important concern for the design of mental health self-monitoring tools. For instance, we may be inclined to design interfaces that are encouraging or calming to support people with the activity of mental health self-monitoring, and with their mental health more generally. Yet such design risks systematically influencing the collection of experience data.

Longitudinal Validity

With MHSM, we are not interested in single experiences or measures, but rather in experiences and measures over time. To produce an accurate characterisation over time, there is a need for repeated measurement with appropriate frequency (Doherty et al., 2020; Shiffman et al., 2008). The characterisation of their mood over time was a concern expressed by participants of an interview study on people’s use of mood monitoring apps by Schueller et al. (2021). The authors note that many mood monitoring apps allow users to record their mood once a day. Participants were concerned that they did not track their mood regularly enough to capture the depth and range of their emotional experience. This included concern that their data would be biased due to entries being at the same time each day and that it was missing particularly important moments. To improve the longitudinal validity of data, there is push for more regular data collection and for attempts to collect data at opportune moments (Doherty et al., 2020; Shiffman et al., 2008).

Ecological Momentary Assessment (EMA)

Motivated by the issues of recall, ecological, and longitudinal validity is the self-monitoring methodology of ecological momentary assessment (EMA), also referred to as experience sampling. EMA methods focus on the regular collection of experience data in real world settings (ecological) and in the moment (momentary) (Doherty et al., 2020; Shiffman et al., 2008).

Validity Trade-offs

While more regular, momentary, and ecological self-monitoring can promote more valid data collection, it may not be appropriate for the self-monitor. Firstly, it may be too burdensome or inconvenient (Schueller et al., 2021). Secondly, due to the reactivity of mental health self-monitoring it risks being too emotionally disruptive or even overwhelming. For example, with more regular self-monitoring people can experience feeling too aware of their mental experience.

“When I was tracking all that time on the excel sheet, I felt my moods constantly. When I stopped using excel and just started “checking in with myself each night” – I felt more normal.” (Matthews et al., 2017, p429)

To support individuals with MHSM, there is a need to balance encouraging more valid and demanding practices with encouraging practices that are less valid but are more achievable.

2.2.5 Prompts and Reminders

Digital tools are an opportunity to promote more valid data collection as they can prompt, or remind, the self-monitor to make monitoring entries regularly and at opportune times. Research has identified that mental health self-monitors do commonly use technology (e.g., alarms on their phone) to remind themselves to self-monitor. This use can be entwined with the self-monitor’s use of digital productivity tools to support a wider range of activities (e.g., eating and exercise) (Matthews et al., 2017).

Informed by passive monitoring (see Section 2.2.6) of the self-monitor’s activities, digital tools can proactively prompt self-monitors at opportune moments. The potential value of proactive reminders was also expressed by survey respondents (Murnane et al., 2016). Informed by their survey of the self-monitoring practices of people living with bipolar disorder, Murnane et al. (2016, p.483) suggest the design implication of:

Design Implication: “deliver proactive notifications”

to promote “adherence to self-monitoring and behavioural regulation” (Murnane et al., 2016).

2.2.6 Passive and Semi-automated Monitoring

An advantage of digital tools for MHSM is *passive monitoring* whereby data is collected automatically. For example, data can be collected using device sensors and through tracking the individual’s digital activity (Kruzan et al., 2023; Malhi et al., 2017; Ng et al., 2019; Rooksby et al., 2019). While an individual’s mental experience can not be passively monitored, their observable health, behaviour, and environment can be, and this can contribute to the self-monitoring activity. Passive monitoring can reduce the monitoring burden and furthermore enable data collection that exceeds the capabilities of active monitoring (Kruzan et al., 2023). Murnane et al. (2016, p.483) also suggest the design implication of:

Design Implication: “Passively monitor and intervene”

to “[reduce] users’ efforts and burdens and [to support] continuous capture of data” (Murnane et al., 2016).

A weakness of passive monitoring is that it can distance the self-monitor from the self-monitoring activity. This can inhibit the development of self-awareness and risks the self-monitor experiencing feeling of a loss of agency or control. Expressing this, Matthews et al. (2017, p.436, emphasis added) present the participant quote:

“If it automatically tracked you, the accuracy would be better. But you almost have no input, no control - it’s more the machine is dictating to you and you are following the commands. I suppose I would do it for the accuracy. The trade-off is dependent on the machine, *you almost have no role in your own life.*”

Furthermore, there can be issues around how self-monitors interpret passively collected data. Passively collected data can be perceived as more trustworthy than manually collected data which can cause the self-monitors to doubt their own experiences. This can be exacerbated by some mental health difficulties giving people due reason to question the accuracy of their recollection of their own experiences (Matthews et al., 2017). While Murnane et al. (2016) advocate for passive monitoring (see design implication above), they stress the need to balance it with manual monitoring.

“Given that the effort of tracking imparted substantial benefits to respondents’ sense of self-efficacy, we must again be attentive to the tension between automation and manual tracking so as to not undermine the agency and mastery linked to more user-driven, deliberate, and mindful forms of tracking.” (Murnane et al., 2016, p.482)

Choe et al. (2017) term self-monitoring approaches that combine manual and passive monitoring as *semi-automated*. The term captures a range of approaches ranging from mostly active to mostly passive monitoring. The authors argue that semi-automated approaches can balance the strengths and weaknesses of passive and manual monitoring, yet designing for this balance is a complex challenge.

2.3 Reflecting on Data

2.3.1 From Description to Actionable Insight

People who self-monitor their mental health do not just want to understand their mental health, they want to improve or manage it. As such, from their self-monitoring practice they want insight about actions they can take to support their mental health. Expressing this, Schueller et al. (2021) report the participant quote of:

“I have confirmation that I am feeling shitty. But, it doesn’t really help you do anything about it.” (Schueller et al., 2021, p.9).

Schueller et al. (2021) report, more generally, that participants of their mood-monitoring interview study expressed that while apps support them to collect data, they provide little support for generating insight from the collected data. Participants expressed a desire for more “interventional components” (p.9) in mood-monitoring apps that provide recommendations, informed by the self-monitor’s data, for what the self-monitor can do to support their mental health. Participants of Kruzan et al. (2023) interview study expressed a similar want for interventional components. In particular, the analysis of Kruzan et al. (2023) highlights the importance of insight relating to things the self-monitor can control

or actions they can take. It was expressed that the identification of patterns is only useful if they inform action the self-monitor can take to promote their mental health.

2.3.2 Analytical Challenge

Analysing mental health self-monitoring data is a complex analytical challenge. Matthews et al. (2017, p.432) argue that to understand and act upon MHSM data, a person requires three proficiencies:

1. “high degree of comfort with technology”,
2. “an awareness of the limitations of the collection method” (data literacy), and
3. “an understanding of how to interpret the data, particularly in a mental health context” (mental health literacy).

Most people who self-monitor for mental health will not have, at least, the latter two of these proficiencies. That is, they will not have sufficient data or mental health literacy to appropriately and effectively analyse their collected data. This is reflected by the reporting of Schueller et al. (2021) that it was common for their interview participants to express discomfort or confusion in relation to analysing their self-monitoring data. In the broader domain of health and wellbeing self-monitoring, Rooksby et al. (2014, p.1171) emphasise that self-monitors do not – and should not be expected to – act as “rational data scientists”. There is considerable potential for analysis of mental health self-monitoring data to be performed without sufficient proficiency. This risks misguiding self-monitors and poses risk to their wellbeing.

2.3.3 Emotional Challenge

The analysis of one’s own self-monitoring data can be emotionally challenging. Again, this relates to reactivity: the potential of assessing mental health to impact mental health (Shiffman et al., 2008). Kelley et al. (2017) conducted a survey of students’ experiences of self-monitoring for mental wellness. Over a third of respondents (36%, N=297) reported having had an experience of feeling their data reflected something negative about them. Feelings of guilt, disappointment in themselves, and embarrassment were commonly reported. Firstly, analysis may demonstrate a mental health trajectory that does not align with the individual’s mental health goals. For example, it may not show the mental health improvements, or management, that the individual is aiming for. Secondly, analysis may also stress the lack of control the self-monitor has over their mental health. Thirdly, analysis risks demonstrating to the self-monitor incorrect beliefs about their mental health and behaviour, or unhealthy behaviours, which can be a challenging experience for the self-monitor (Kelley et al., 2017). Finally, analysis may be triggering. For example, the identification or prediction of mental health decline may in turn contribute to a decline (Matthews and Doherty, 2011; Murnane et al., 2016).

2.3.4 Analysis Support

The analytical and emotional challenges of analysis suggest a need to support mental health self-monitors with the analysis of their data. Digital tools could support self-monitors through automated analysis. As previously described, Schueller et al. (2021) report that current applications do little to support self-monitors with analysis. Participants, however, expressed interest in automated analysis.

“I think it would be cool if it sort of had a bit more than just your own input like it took your input. Not only does it compile it into data but sort of give you solutions or recommendations based on the data that you provide.” (Schueller et al., 2021, p.9)

Similarly, Murnane et al. (2016) reported their survey participants were highly receptive to the prospect of ‘intelligent’ tools that provide automated analysis. They suggest the design implication of:

Design implication: “synthesize data and highlight patterns” (Murnane et al., 2016, p.483).

Healthcare professionals could also support self-monitors with the analysis of their data by contributing mental health expertise to the analysis. However, the analysis proficiency of healthcare professionals may be limited by a lack of comfort with technology or a lack of data literacy (Murnane et al., 2016). Healthcare professionals may therefore also benefit from automated analysis support. I discuss the opportunities and challenges of healthcare professional involvement in MHSM further in Section 2.6.1.

2.4 Intervention

Building on the want for actionable insight from self-monitoring tools, digital tools could act in interventional ways to support the individual to act. Kruzan et al. (2023) found participants were appreciative of technologies that went beyond monitoring to reminding and encouraging them to engage in beneficial activities. For example, their Apple Watch reminding them to stand-up and stretch after having been sat-down for a period of time (Kruzan et al., 2023). Similarly, a participant of Matthews et al. (2017) imagined a ‘social monitor’ that could monitor how much they socialised and advise them as to when they should have some quiet time. To make appropriate interventions, tools could harness passive sensing to identify what actions may be appropriate, and when it is appropriate to suggest them to the individual. This has been advocated for as a design implication by Murnane et al. (2016, p.483)

Design Implication: “passively monitor and intervene”.

This implication was previously referred to in Section 2.2.6.

Ecological Momentary Intervention

The idea of proactive interventions, particularly harnessing passive monitoring, is embodied by the treatment approach of ecological momentary intervention (EMI). EMI extends the approach of ecological momentary assessment from assessment to intervention. The defining idea of EMI is to provide people with treatment as part of their everyday lives, that is in their natural (ecological) settings and in real-time (momentary) (Balaskas et al., 2021; Heron and Smyth, 2010; Schueller et al., 2017).

2.5 Trajectories

People’s MHSM practices can be highly idiosyncratic, can develop, and can fluctuate. Tool design should account for people’s varying mental health and MHSM needs.

2.5.1 Initiating Self-monitoring

There will be moments when people decide to start more actively self-monitoring their mental health. In personal informatics more broadly, Epstein et al. (2015) model this initiation of self-monitoring with two stages: deciding and selecting.

People first *decide* to start self-monitoring their mental health. From an interview study of people who use mood-monitoring apps, Schueller et al. (2021) identified that the majority of their participants (13/22) decided to start ‘mood monitoring’ in response to challenging life experiences. Challenging events can motivate people to want to develop their awareness of, and improve, their mental health. The authors further note that no participant reported deciding to start in response to positive life experiences. People will also commonly decide to start self-monitoring their mental health upon the recommendation of healthcare professions; for example, as part of doing cognitive behavioural therapy.

Having decided to start self-monitoring, people then *select* how to self-monitor their mental health. Schueller et al. (2021) also found from their interview study that people chose to use mood monitoring apps based on recommendations from other people including: friends, family, and therapists. Participants also reported searching out apps independently (e.g., through app stores).

“I’d been using actually mood apps for a while, because they were recommended by my psychiatrist. She pretty much told me was some people use journals and stuff like that, but that it’s easier to use an app just because everybody has a phone nowadays” (Schueller et al., 2021, p.7).

2.5.2 Developing Self-monitoring Practices

Due to the diversity of things that can relate to an individual’s mental health and both the individual and experiential nature of mental health, it can be difficult to know what items to monitor (Kruzan et al., 2023). Kruzan et al. (2023) present a participant quote expressing this difficulty:

“I just don’t know what the right metric is for me to pay attention to. Or I don’t know should I be doing better to improve it? So I don’t know, mental health is so hard to nail down, so you either feel a certain way or you don’t, broadly speaking.” (Kruzan et al., 2023)

This difficulty is the result of having limited understanding of one’s mental health which is an exact challenge that MHSM tackles.

Contrasting approaches are suggested for how to develop a mental health self-monitoring practice. A common approach among the survey participants of a study by Murnane et al. (2016) was to start monitoring a few condition-related indicators as a gentle introduction. Given that a common issue encountered during self-monitoring more broadly is attempting to monitor too many items (Choe et al., 2014), this gentle introduction may be sensible. From this gentle introduction, the participants of the study by Murnane et al. (2016) report adapting their monitoring. Adaptions may involve ceasing or initiating the monitoring of items, as well as modifying how items are monitored (e.g., entry representation or frequency). In contrast to this gentle introduction approach, Schueller et al. (2021) highlight the potential of a more intense introduction. When individuals start monitoring there is often a novelty period during which engagement is high. Schueller et al. (2021) remark that there is potential to design for this high engagement and to leverage it with an initial period of more intense monitoring.

2.5.3 Phases of Monitoring

Individuals may transition between different phases of monitoring over time. Transitions can be triggered by changes to the individual’s mental health and their monitoring goals.

Li et al. (2011) describe two phases of tracking in personal informatics: discovery and maintenance. Relating these phases to mental health self-monitoring, discovery is when the self-monitor investigates the impact of different factors on their mental health. Maintenance is where the self-monitor has an understanding of the factors that impact their mental health and is monitoring them to inform the management of their mental health. Initial self-monitoring will typically have a greater focus on discovery and shift more to maintenance as the self-monitor’s understanding of the impacts of different factors develops. Self-monitors may often engage in a combination of discovery and maintenance.

Self-monitors may switch between different phases to align with their mental health support needs. From an interview study that considered people’s mental health self-monitoring desires, Kruzan et al. (2023) distinguish two forms of monitoring:

1. mental health maintenance, and
2. active symptom management,

which correspond with experiencing periods of low and high levels of symptoms. During periods of mental health maintenance there is potential for greater monitoring and exploration. This can help develop the individual’s understanding of their mental health and identify preventative actions to counter potential future mental health declines. During periods of active symptom management, monitoring should focus less on exploration

and more on symptom management. For example, monitoring technology could encourage individuals to take actions to support their mental health, and acknowledge these efforts (Kruzan et al., 2023). Relating these two forms of monitoring to Li et al. (2011) phases of discovery and maintenance, mental health maintenance has a discovery focus while active symptom management has a maintenance focus. Similarly, informed by their study of how people use mood-monitoring apps, Schueller et al. (2021) suggest designing for periods of intense (discovery focused) and maintenance monitoring. Kelley et al. (2017) similarly identified phases of self-monitoring as the goals of the activity change, for example, from ‘gain’ to ‘maintain’ goals.

2.5.4 Lapses

Self-monitors may stop actively using self-monitoring tools. Epstein et al. (2015) describe this as a personal informatics stage termed lapsing. They also present a resuming stage to describe when people start actively self-monitoring following a lapse. Four categories of lapse are described:

1. forgetting,
2. upkeep where tool maintenance becomes a barrier to self-monitoring (e.g., devices running out of battery),
3. skipping where the self-monitor decides to not monitor specific data, and
4. suspending where the self-monitor decides to stop using a monitoring device temporarily.

Skipping and suspending are particularly relevant to mental health self-monitoring. A reoccurring finding is that some people who mental health self-monitor do not want to record negative things (Schueller et al., 2021), that is they skip recording negative data. Another reoccurring finding is that self-monitors shift to rely more on mental monitoring, and less on tools, during periods of better mental health (Matthews et al., 2017; Murnane et al., 2016). This is an example of suspending tool use. Correspondingly, it is also reported that people make greater use of tools during periods of greater mental health difficulty (Murnane et al., 2016).

2.6 Support Relationships

The management of an individual’s mental health often involves more people than just the individual. Healthcare professionals, and a person’s friends and family, can have impactful roles (Murnane et al., 2018).

2.6.1 Healthcare Professionals

Healthcare professionals can be important actors within the social ecologies of MHSM (Murnane et al., 2018). This is to the extent that Matthews et al. (2017) have argued that most

research on self-monitoring has been clinician-centred, focusing on what is important to clinicians for the clinical management of mental health difficulties. Matthews et al. advocated for considering more person (self-monitor) centred design. As both self-monitoring and professional support can play large roles in mental health support, it is important to consider how self-monitoring and professional support can impact each other.

Influence of Healthcare Professionals

Healthcare professionals can influence people's attitudes and beliefs about self-monitoring. They can encourage people to initiate and maintain self-monitoring practices (Schueller et al., 2021). A lack of appreciation of, or receptivity to, an individual's self-monitoring can however be detrimental to a self-monitor's beliefs about their practice (Murnane et al., 2016), potentially leading them to reduce or stop their practice. Given this influence, we should consider how we can design self-monitoring tools to harness the influence of healthcare professionals on self-monitoring practice.

Facilitating Interactions Between Self-monitors and Professionals

MHSM data can facilitate interactions between self-monitors and healthcare professionals; for instance, during healthcare appointments and therapy sessions. The sharing of their self-monitoring data can be empowering for self-monitors in interactions with healthcare professionals. It can allow them to evidence and more effectively communicate their experiences (Kelley et al., 2017; Kruzan et al., 2023; Murnane et al., 2016; Schueller et al., 2021). Murnane et al. (2016) remark that this enables self-monitors to be more:

“‘active patients’ who do not simply ‘fill prescriptions’” (Murnane et al., 2016, p.480).

Additionally, the sharing of data can also afford self-monitors a means of indirect disclosure to healthcare providers which can support them to communicate about sensitive topics (Matthews and Doherty, 2011). Analysis of patient self-monitoring data can give professionals a better understanding of their patients. It could support them to evaluate treatment compliance (Kelley et al., 2017) and be an opportunity to identify things that may be impacting a self-monitors mental health that the self-monitor is either unaware of or is uncomfortable talking about.

Challenges of Harnessing Self-monitoring Data

It can be challenging for healthcare professionals to harness self-monitoring data. First, it can be challenging for self-monitors to share their data with healthcare professionals. Self-monitors report struggling to print their data and having digital transfers (e.g., emails) rejected (Murnane et al., 2016). Second, building on my discussion about analytical challenges (Section 2.3.2), while healthcare professionals may have mental health expertise, they can lack the comfort with technology or data literacy to proficiently interpret the data of people who self-monitor (Murnane et al., 2016; Ng et al., 2019). Self-monitors report healthcare professionals struggling to interpret their self-monitoring data even in

aggregate and visualisation forms. (Murnane et al., 2016). Third, even if health professionals are able to receive and proficiently analyse the self-monitoring data they may not have time to do so. These barriers can result in unsuccessful attempts by self-monitors to share their data which can elicit feelings of resentment and of their self-monitoring being futile (Murnane et al., 2016); as previously discussed, this can be detrimental to future self-monitoring. Murnane et al. (2016, p.483) offer the design implication of:

Design Implication: “integrate within clinical care while respecting existing practices via doctor-view interfaces, digestible summary reports, and modifiable settings”.

They write that this implication is to “[facilitate] improved acceptance, transmission, and interpretation of information by treatment teams” (Murnane et al., 2016).

Monitoring Self-monitors

MHSM data could be monitored by healthcare professionals. The participants of Kruzan et al. (2023) study imagined their self-monitoring data being regularly shared with their provider to enable their provider to alert them to declining mental health, and provide guidance. Indeed, a role of providers as monitor, interpreter, and adviser is envisioned (Kruzan et al., 2023).

2.6.2 Friends and Family

Social Support

People’s friends and family can play important roles in the management of mental health. Friends and family can offer emotional support. For example, participants of a study by Kruzan et al. (2023) reported how they would feel better after talking with their marital partners and other family members. Furthermore, friends and family can encourage and support people to do things to support their mental health. Schueller et al. (2021) reported that participants of their study choose mood monitoring apps based on recommendations from friends and family. A participant of a study by Kruzan et al. (2023) reported their partner encouraging them to do activities to support their mental health during times of difficulty.

“My boyfriend really helps me manage my mental health. He’ll be the one who will sometimes see how I am that day. If I’m really stressed or I’m really sad that day, he’ll be like, “Hey, let’s go on a walk or let’s do this or let’s do that”” Kruzan et al. (2023)

Participants of a study by Murnane et al. (2018) describe people close to them offering more instrumental and interventional support.

“I had a couple very close friends that were trying to keep an eye on me and were literally chasing me around and making sure I was taking my medication, taking alcohol and drugs out of my hand” (Murnane et al., 2018, p.11)

When designing tools for mental health self-monitoring, it should be considered how self-monitoring tools and social support for mental health may impact each other.

Supporting MHSM

People’s close friends and family can be a component of their mental health self-monitoring. Self-monitors have reported people close to them acting as human, or social, sensors of their mental health. These people may offer observations and feedback on the person’s mental health, potentially with a focus on warning signs for mental health decline (Matthews et al., 2017; Murnane et al., 2018). Tools for MHSM could be designed to enable monitoring input from the people close to the self-monitor (Murnane et al., 2018). Similarly to how self-monitors may share their data with healthcare professionals, self-monitors may share their data with friends and family to facilitate the communication of their experiences (Kelley et al., 2017; Schueller et al., 2021). The sharing of achievements can also produce a greater sense of accomplishment (Kelley et al., 2017). Accordingly, when designing tools for mental health self-monitoring we should consider functionality for the sharing of data with friends and family.

Impacts of Mental Health Difficulties on Relationships

The nature of mental health difficulties can stress relationships. Participants of a study by Kruzan et al. (2023) attributed the mental health difficulties they were experiencing to resulting in them behaving in ways that are not appreciative of, or “snappish” toward, the people around them. Such behavior may cause relationship tensions that can exacerbate the difficulties being experienced and reduce the social support available (Kruzan et al., 2023). Similarly, participants of a study investigating the challenges experienced by the family caregivers of people with depression reported relationship stress with the people they care for behaving in negative, forgetful and offensive ways (Yamashita et al., 2013). Prior work has investigated the design of monitoring interfaces for use by family caregivers to help them care and cope with caring for depressed family members (Yamashita et al., 2017).

Furthermore, within a relationship a person’s experience of mental health can be very impactful upon that of the other person. Matthews and Doherty (2011) present an example of this, they describe a case of a teenager – ‘Beth’ – who was receiving professional mental health support. It had been identified that Beth’s relationship with her mother was strained, and it was suspected that there was strong correlation between their moods. From both Beth and her mother self-monitoring their mental health, a strong correlation between their moods was observed.

Detrimental Support

While relationships can support a person’s mental health they can also be detrimental to it. The good intentions of friends and family may not be well informed; as with self-monitors and health professionals, friends and family may lack the proficiencies to offer effective support. Friends and family may also not have good intentions as relationships can be harmful (Murnane et al., 2018). Care must be taken during the design of self-monitoring tools for mental health support to mitigate against potentially detrimental

friend and family involvement.

2.6.3 Relationship Changes

Murnane et al. (2018) emphasise that the characteristics of different relationships can vary substantially. For example, people’s relationships with healthcare professionals will likely be quite different to those with their friends and family. The authors further emphasise that the characteristics of relationships can change. For example, relationships may end with healthcare professionals (e.g., at the end of fixed-duration CBT programme) and with friends and family (e.g., relationship break-ups). Murnane et al. (2018) consider the implications for the design of collective self-monitoring systems such as those that support data sharing and third-party monitoring. They propose the design implication of:

Design Implication: “Systems need to accommodate the emergent and dynamic practices of accruing, breaking, and changing social ties.” (Murnane et al., 2018, p.19)

In relation to this implication, the authors discuss ways to onboard new healthcare professionals and the need for effective privacy controls. They caution against designing for stereotypical roles and relationships and argue that systems should support per-recipient configuration.

2.6.4 Supportive Tools

Research on the design of tools for MHSM often reports that people want tools to support them in encouraging ways similar to how their friends and family support them. Matthews et al. (2017) describe a participant who used a self-monitoring tool that presented feedback using conversational prompts. The authors report that the participant appreciated that the prompts were encouraging and quoted them as saying:

“It’s nice to hear the pat on the back. Because that’s a great motivational thing, and especially if it’s talking to you like it’s a buddy.” (Matthews et al., 2017, p.426)

They also report the account of a participant who created a custom spreadsheet for MHSM that responded to her with messages of encouragement. Informed by an interview study on MHSM, Kruzan et al. (2023) remark:

“Participants imagined that a self-tracking technology could help them to follow through on their intentions and take stock of their efforts to manage mental health through regular check-ins and validating messages. Interestingly, in this way, the imagined DMHI [(Digital Mental Health Intervention)] served a similar role to trusted peers and significant others - as a resource to draw attention to symptoms, offer solutions, and keep them accountable.”

Also informed by a MHSM interview study, Schueller et al. (2021) envisioned designing MHSM interactions to resemble interactions where a person’s friends or family check-in with them about their mental health. Additionally, based on a large survey study, Murnane et al. (2016) proposed the design implication of:

Design Implication: “provide encouragement messages and rewards or (after non-compliance) flexibility and forgiveness” (Murnane et al., 2016, p.483)

They report that this implication is to provide “experiences of mastery and [to] [cultivate] self-efficacy and self-compassion”. As noted in Section 2.2.4, designing tools to be supportive may be detrimental to the validity of collected data. However, as discussed in Section 2.2.4, design must balance the validity with the practicalities of MHSM.

2.7 Conclusion

Within this chapter I have characterised the activity of MHSM with a focus on the design of digital tools to support people with it. My characterisation of MHSM makes clear that it is a complex and challenging activity that has the potential to be beneficial and detrimental to a person’s mental health. The consideration of digital tools for MHSM makes apparent opportunity to design new tools that better support people to do and benefit from MHSM. There are many different aspects of MHSM that tools could be designed to better support people with. As described within this chapter, digital MHSM tools could be designed to:

1. help users to record rich entries of high validity (Sections 2.2.3 and 2.2.4);
2. prompt and remind users to make entries (Section 2.2.5);
3. collect data for users to both reduce the burden of, and to enrich, data collection (Section 2.2.6);
4. analyse collected data to provide users with appropriate and actionable insight (Section 2.3.4);
5. intervene to promote positive, and mitigate negative, health behaviours (Section 2.4); and
6. behave in supportive ways to motivate the user and help the user emotionally (Section 2.6.4).

As introduced in Chapter 1, conversational agents (CAs) are a promising form of computer interfaces for the design of digital MHSM tools. Next, in Chapter 3, I review work related to the design of CAs for MHSM.

Chapter 3

Background: Conversational Agents

3.1 Introduction

In Chapter 1, I introduced mental health self-monitoring (MHSM) and suggested that conversational agents (CAs) are an alluring form of computer interface for the design of digital MHSM tools. I then, in Chapter 2, characterised MHSM in greater detail with a focus on the design of digital tools to support people with it. Chapter 2 makes clear that there is opportunity to design new tools that support people to do MHSM more effectively and in turn support them with their mental health. In order to start exploring the potential of CAs for MHSM, in this chapter I discuss research relevant to the design of CAs for MHSM. This chapter is structured as follows:

- 1. Technical Implementation:** In Section 3.2 I briefly summarise different technical approaches for implementing CAs.
- 2. Self-Disclosure to CAs:** Making MHSM entries is an act of self-disclosure. In Section 3.3 I discuss research investigating self-disclosure to CAs.
- 3. Relational Agents:** I discuss in Section 3.4 research on relational agents (RAs). This work investigated how CAs could be designed to promote user-CA relationships that help engage users in healthcare activities (e.g., medication adherence).
- 4. Therapy Agents:** Section 3.5 describes research on therapy agents (e.g., Woebot and Wysa) which are CAs designed to support people with their mental health through mimicking techniques of established talking therapies (e.g., CBT). I discuss the design of these CAs, which is similar to that of RAs, as well as clinical and user experience evaluations.
- 5. Social Agents:** Section 3.6 discusses social agents (e.g., Replika). Social agents are designed to develop relationships with users whereby they take on companion, friend, or romantic partner roles. They demonstrate a different, and more extreme relational approach, to the design of CAs for mental health support.

- 6. Social Support from CAs:** In Section 3.7 I describe studies relating support from CAs (Woebot and Replika) to human social support. This section describes how CAs can provide not just mental health support, but social support more broadly. Social support theory therefore might be a useful basis for considering how CAs can be designed to support people.
- 7. Human-CA Relationship Development:** Next, in Section 3.8, I discuss research on the development of human-CA relationships. Work relating human-CA to human-human relationship development is described. This work views self-disclosure as a mechanism of relationship development making it pertinent to MHSM.
- 8. Using Theories of Human-Human Interaction** In Section 3.9, I discuss the use of theories of human-human interaction as basis for the study of human-CA interaction.
- 9. Ethics and Safety** Lastly, in Section 3.10, I describe some outstanding ethics and safety considerations.

I conclude this chapter by first reiterating that CAs seem a promising type of computer interface for the design of digital MHSM tools. I then stress that, although they seem promising, it remains unclear how to design them effectively to support people with MHSM.

Credit Statement

This chapter is my own work informed by feedback from my primary supervisor Gavin Doherty.

3.2 Technical Implementation

3.2.1 User Input

User input may be multiple-choice whereby users are given pre-written utterances to choose from or unconstrained such that users can type or speak what they want to input. Systems may support a combination of multiple-choice and unconstrained input with the input type varying between turns. Multiple-choice input allows designers to constrain interactions and therefore ensure that they are safe and appropriate (Deibel and Evanhoe, 2021). Unconstrained input gives users greater ability to express themselves, however it comes with the natural language understanding (NLU) challenge of engineering the system to respond appropriately to anything the user inputs (Bickmore et al., 2018b,c; Miner et al., 2016). As is discussed further in Section 3.10.1, CAs responding inappropriately to users is a safety risk of CAs for mental health support.

Unconstrained input can be typed or spoken. Systems that allow spoken natural language input typically transcribe it into text for further processing using automatic speech recognition (ASR) (Deibel and Evanhoe, 2021). ASR further complicates the challenge of engineering systems to respond to users appropriately as unsuccessful ASR can cause systems to respond inappropriately.

3.2.2 CA Output

CA responses can be scripted for multiple-choice input, whereas NLU techniques are used for responding to unconstrained input. There are three main approaches for responding to unconstrained input: rule-based, retrieval-based, and large language model based (LLM-based) (Ma et al., 2024). Systems can be created that combine these approaches in order to utilise their different strengths and to mitigate their different weaknesses. A rule-based approach involves writing rules that declare how the CA should respond to messages. For example, rules could be written that instruct a system to respond with a certain utterance when a keyword is present. A retrieval-based (or intent-based (Deibel and Evanhoe, 2021)) approach involves matching user inputs to the most appropriate response within a database of responses using a machine learning model. Designers tune these models by providing examples of user inputs that correspond to each response (Deibel and Evanhoe, 2021; McTear et al., 2016). A LLM-based approach involves tuning a pre-trained LLM to interact with users in the desired manner through ‘prompt-engineering’ (McTear and Ashurkina, 2024). Prompt engineering involves inputting prompts to the LLM that describe how the LLM should respond to user inputs. Prompts may, for example, describe a persona that the CA should enact or how the CA should respond to specific requests (McTear and Ashurkina, 2024; Zamfirescu-Pereira et al., 2023). Rule-based and retrieval-based systems allow designers to control what CA responses are possible, however it is challenging to engineer them to respond well to the diversity of things that users may input. In contrast, LLM-based systems can respond well to the diversity of user inputs, but designers have less ability to control exactly how the CA will respond. With current LLM technology, it is challenging to prevent systems from generating responses that are incorrect or inappropriate (McTear and Ashurkina, 2024).

3.3 Self-Disclosure to CAs

MHSM requires the self-monitor to manually record data about their mental experience. Manually recording this data is an act of self-disclosure. In this section I discuss a diversity of research that studies how people self-disclose to CAs.

3.3.1 CAs Versus People

Research has contrasted how people self-disclose to CAs with how they disclose to people. Lucas et al. conducted an experiment to compare people’s willingness to disclose between a human-operated, and a fully-automated, virtual human (a speech-based human-looking avatar) in the context of a health-screening interview (Gratch et al., 2014; Lucas et al., 2014). During the study, participants were asked sensitive and intimate questions such as:

- “Tell me about an event, or something that you wish you could erase from your memory”
- “Tell me about the last time you felt really happy” (Lucas et al., 2014, p.97)

Their analysis shows participants were more willing to disclose to the fully-automated virtual human, that is when they thought they were not being observed by another human, than to the human-operated virtual human. An explanation for the greater willingness to disclose to a computer than to a human, supported by anecdotal remarks from participants, is that the computer is perceived as non-judgemental making it a preferable partner to disclose to.

Ho et al. (2018) compared the effects on participants of self-disclosure to people with those of self-disclosure to computers. They conducted a Wizard of Oz experiment using a text-based chat interface whereby participants were told they were either interacting with a computer agent or another person. The analysis found that the effects of disclosure to a computer were equivalent to the effects of disclosure to a person across a range of psychological, relational, and emotional measures. The authors note, however, that participants of this study, unlike in the study by Lucas et al. (2014), were not asked to disclose about potentially embarrassing or stigmatized topics. For these more sensitive disclosures, people’s fears of being judged might be heightened and consequently the effects of disclosure to people and to computers may not be equivalent.

While not designed to directly compare disclosure to CAs with disclosure to people, a study by Kim et al. (2018) elicited such comparisons from participants. Kim et al. (2018) explored teenagers’ perceptions of CAs for emotional support through workshops and interviews. Echoing the idea of computers being non-judgemental, this study found that participants perceived CAs as good listeners. Unlike people, CAs can be relied upon to always be available to listen about anything and to respond in a patient and empathetic way. Furthermore, participants felt CAs could be trusted to not reveal their secrets, particularly more so than their friends.

Morris et al. (2018) conducted a study using the Koko peer-to-peer support platform. Koko uses a post-response format whereby a user seeking support submits a post which is then responded to by a different user who wants to provide support. A manipulation was performed whereby some Koko users were told that the response they were receiving was from a CA, rather than from a peer. The responses, however, were still from a peer. Ratings of response quality for messages from the agent were compared with ratings for messages from peers. This comparison demonstrated that messages from the agent were rated less favourably. Seemingly for Koko users, the value of a support message is impacted by if it was perceived to be from a CA or a human. In contrast to the previous studies, this work suggests that some people are less receptive to self-disclosure to and receiving support from a CA, in comparison to an (anonymous) person. However, this finding may not generalise well as Koko users are people who have chosen to use a peer-to-peer support platform hence are likely a cohort with a preference for human support.

3.3.2 CAs Versus Forms

Research has contrasted CAs with forms for data collection activities. While functionally CAs and digital forms are very similar, they may produce quite different user experiences. Kim et al. (2019) compared CAs with web surveys for data collection. A 2×2 between-

participants experiment was conducted with 117 adolescents comparing the two variables of platform (web survey vs. CA) and conversational style (formal vs. casual). Quantitatively, the study’s analysis demonstrates that the web survey was completed faster and that CAs elicited higher quality responses based on measures derived from satisficing theory. It demonstrates an interaction effect whereby the CA with a casual conversation style elicited the highest quality responses. Furthermore, an effect of conversation style on participant ratings of enjoyment was identified whereby casual interfaces were rated as more enjoyable. Qualitatively, the casual CA was praised for making the interaction feel like a friendly and understanding interaction with a person. The formal CA was critiqued for being rigid, unfriendly, and boring. This work demonstrates that while CAs can support data collection, in comparison to web surveys, their effectiveness is impacted by their conversational design.

Xiao et al. (2020) also compared CAs with web surveys for data collection; the findings of this study align well with those of the previous study by Kim et al. (2019). A between-participants study involving 582 participants was conducted whereby 300 completed a web survey and 282 interacted with a CA. The quality of collected data was analysed primarily using a manual coding based on Grice’s maxims. Grice’s maxims are of: quantity, quality, relevance, and manner; the maxims are considered principles of effective cooperative communication. For each maxim, a quantitative quality measure was created. Aligning with the analysis of Kim et al. (2019), for each quality measure, Xiao et al. (2020) demonstrate the chatbot responses were of higher quality. Participants took longer to complete interactions with the chatbot than the web survey, again aligning with the analysis of Kim et al. (2019).

In sum, these studies suggest that CAs may be a more effective way to elicit responses from people during data collection activities than forms. The effectiveness of CAs for data collection may, however, depend upon their conversational style. Additionally, this work suggests it takes longer to collect data using CAs than web forms which may impact their relative suitability for different data collection activities. CAs may, for example, be best suited to longer data collection exercises, and web forms better suited for quick and brief exercises.

3.3.3 CAs for Journaling

The conversational format of interaction with CAs is an opportunity to dynamically respond to user input in order to develop rich self-disclosures. Park et al. (2021) investigated the impacts of responsive behaviour by a CA for self-disclosure. A four day between-participants study was conducted with 30 participants whereby participants were asked to engage in daily expressive writing activities. 10 participants were asked to use a Google Doc, 10 participants were asked to use a basic CA, and the remaining 10 a responsive CA. The CAs were named DiaryBot. The basic CA would deliver instructions for the expressive writing activities in a casual and encouraging style. It would emphasise that it would keep the user’s disclosures private and that they should feel comfortable writing about their deepest thoughts and feelings. For example, it used the prompt:

“Good to see you! I’m ready to begin. I’ll keep all your secrets, so no worries.
:)
You can write about your deepest feelings and thoughts on the most traumatic
experience of your life so far.”

The basic CA would thank the participant for their self-disclosure but it would not respond to the specifics. In contrast, the responsive CA may respond to the content of the self-disclosure asking them about emotions expressed or people referred to in order to promote emotional- and social-awareness respectively. The responsive CA may also ask the participant to leave a message for themselves in order to promote self-awareness. Pre- and post-study surveys, interviews, and analysis of self-expression data were used to evaluate and compare the tools. Unfortunately, implementation issues with the responsive CA resulted in participants experiencing few responsive prompts. 2 of the 10 participants did not experience any, and 4 of the 10 only experienced them once. As a result, the analysis of this – particularly the quantitative analysis – must be interpreted cautiously. The analysis of self-expression data demonstrated that participants using a DiaryBot wrote in a more conversational style where the self-disclosure was addressed to the agent. This is echoed by participants describing, in interviews, how using DiaryBot made the interaction feel like a conversation where they were telling a story to someone. The authors identify that a key implication of this work is that using a CA for expressive writing changes the activity from one of *writing* to *telling* a story. During the interviews, participants described responsive prompts as being helpful for ordering their thoughts and finding new insights. Some participants expressed that basic DiaryBot did not give much back in return and they wanted some form of feedback such as empathetic feedback. The authors argue that this highlights that there is a user expectation for CAs to perform in a manner that more closely mimics two-way, rather than one-way, conversation.

In recent work, (Kim et al., 2024) designed a large language model (LLM) based journaling app named MindfulDiary to help psychiatric patients with documenting their daily experiences. 28 patients used MindfulDiary during a four-week study. Echoing the prior study of DiaryBot (Park et al., 2021), it was reported that participants of this study found conversational interaction an engaging way to create journal entries. Participants found the structure, encouragement, and feedback facilitated writing rich entries (Kim et al., 2024).

3.3.4 Reciprocal Self-Disclosure by CAs

Lee et al. (2020b) conducted a study that involved participants interacting daily, for three weeks, with CAs designed to elicit user self-disclosures. This study investigates how the self-disclosure from the CA would impact user’s self-disclosures. Interactions involved two activities: journaling and sensitive questions. Journaling was a MHSM activity that asked participants about their mood and things impacting it. Participants were then asked sensitive questions; examples of the sensitive questions asked include:

- “Have you ever experienced any abusive behavior (e.g., sexual, physical, or emotional) from your friends, family, or a stranger?”

- “How would you prefer to die?” (Lee et al., 2020a, p.3)

For participant safety, people had to score less than 13 on the Kessler Psychological Distress Scale to participate, psychiatrists reviewed the CA and study design, and participants were provided with emergency contact information. Despite these precautions, the appropriateness of the choice of these ‘sensitive’ questions with regard to participant wellbeing and study relevance is questionable. The study compared three chatbots, that varied by how much they would self-disclose to the user, using a between participants study design. The first chatbot (no disclosure) was basic and would not self-disclose. The second (low disclosure) and third (high disclosure) would engage in small talk with the user; as part of this the high disclosure chatbot would make greater self-disclosures expressing thoughts and feelings. Quantitative analysis shows that with greater CA disclosure there was greater participant disclosure (word count). This study, which used an activity that closely resembles MHSM, also reported that participants expressed that they felt they could self-disclose freely to the chatbot without feeling embarrassed, because it was not a human.

3.3.5 CAs as Mediators for Self-Disclosure to People

In conjunction with the study by Lee et al. (2020b) discussed above, Lee et al. (2020a) studied how the CA could act as a mediator of self-disclosures between users and mental health professionals. After the three weeks of the initial study, the CA then enquired if the user would like to share their data with a mental health professional. As part of sharing, participants could modify (add, delete, edit) their data. Despite the introduction of this prospect being a surprise for participants, over 90% of participants were willing to share some form of their data. Participants revised their data in a range of ways. Some participants deleted their thoughts and feelings to leave more factual descriptions. Participants also added to and edited their disclosures to correct errors and to increase the clarity of the descriptions. These revisions suggest differences between what people are comfortable sharing, or perceive is worthwhile to share, with a CA and with a mental health professional. The authors report that over 80% of participants expressed that they would find it easier to talk about the sensitive questions with a CA than with a human. This relates to discussion from Section 2.6.1 regarding sharing self-monitoring data with mental health professionals.

3.3.6 CAs for Mental Health Assessment

Lucas et al. (2017) investigated if military service members would report more mental health symptoms (post-traumatic stress disorder) to 1) a virtual human than when using 2) a standard assessment and 3) an anonymous standard assessment. The motivation was that, in contrast to the standard assessment, the virtual human and the anonymous assessment would both offer anonymity, and the virtual human would additionally offer rapport. An initial study (n=24) showed that participants reported more symptoms, with statistical significance, to the virtual human than when using both the anonymised and non-anonymised standard assessment.

Maharjan et al. (2021) investigated the feasibility of speech-based CAs for administering mental health assessment questionnaires. A controlled within participants lab-based experiment was conducted comparing completion of a paper questionnaire and two speech-CA questionnaires. One speech-CA questionnaire supported discrete user responses and the other open-ended user responses. Quantitative analysis demonstrated that the speech-CA responses correlated well with the paper responses. The correlation was stronger for the discrete response CA which is unsurprising given it shares the response options of the paper questionnaire. The questionnaire completion was significantly quicker for the open-ended response CA compared with the discrete response CA. This result likely reflects that open-ended responses are easier for users as they are not required to recall and choose between fixed options. The authors suggest the use of “conversational cues to support engagement” such as responding to users with phrases such as ‘sorry’ or ‘glad to hear that’. It is also suggested by the authors that designers “exploit CAs’ interactivity to elicit richer and more complete responses” by designing CAs to ask follow up questions.

3.3.7 Speech Interaction for Disclosure

In further work on speech-CAs, Maharjan et al. (2022) conducted a four-week study that involved 20 people using a speech-CA for MHSM that was implemented using a smart speaker located in each participant’s home. The qualitative analysis of this work reports technical issues including the CA failing to understand, and interrupting, the user. These issues were a major barrier to successful MHSM and caused participants frustration. More positively, participants viewed the CA as a good listener. Participants appreciated that the CA could be relied upon respond with care and non-judgmentally, as well as that it was constantly available. It was also reported that speech interaction rendered a private practice into a social one, as other people within the participant’s home may hear the interactions. While some participants wanted to keep their interactions private and thus did not use the CAs while others were around, other participants appreciated the insight about their mental health that it gave to those around them. Indeed, the audibility of speech-based interaction allowed the CA interaction to facilitate indirect self-disclosure to other people (Maharjan et al., 2022).

In recent work, Cooney et al. (2024) investigated the perspectives of mental health professionals on the use of speech-CAs for MHSM. This work similarly raises the concern that the audibility of speech interaction poses a risk to user’s privacy, particularly in public spaces. The authors recommend providing other interaction modalities (e.g., text interaction) to enable people to record entries more privately.

3.3.8 CAs for Ethnography

Tallyn et al. (2018) investigated the use of CAs for collecting ethnographies. A CA called Ethnobot was created and trialed with 13 visitors to a large agricultural show. Ethnobot was designed to elicit descriptions from participants about their transactions, activities, and thoughts at the show. As commented by the authors, the use of Ethnobot is similar to experience sampling (Section 2.2.4) and thus self-monitoring activities. Reactions to

Ethnobot were mixed; while participants seemed receptive to the idea of Ethnobot, they were irritated by its limited capabilities. The participants tended to anthropomorphise Ethnobot. This seemed beneficial in that it encouraged data collection interactions and made them more playful. However, it also caused issues whereby the anthropomorphism led participants to have unrealistic expectations of the Ethnobot’s capabilities leading to feelings of disappointment and frustration when user expectations were not met. Similarly, participants were disappointed that Ethnobot lacked knowledge of the user’s context and environment.

3.3.9 Summary

Self-disclosure is fundamental to MHSM. So far in Section 3.3 I have described a range of studies relating to self-disclosure to CAs. Four key things that the work described suggests are:

1. People are receptive to self-disclosure to CAs because they are perceived as good listeners (see Sections 3.3.1, 3.3.4, and 3.3.7).
2. Conversational interactivity can make self-disclosure activities more enjoyable and less challenging for users which can promote the collection of rich self-disclosure entries (see Sections 3.3.2, 3.3.3, 3.3.6, and 3.3.8).
3. Self-disclosure to CAs can be an opportunity for indirect self-disclosure to other people (see Sections 3.3.5 and 3.3.7).
4. The technical limits of CA interaction can cause usability and user experience issues (Sections 3.3.7 and 3.3.8).

3.4 Relational Agents

Relational agents (RAs) are virtual human CAs that simulate verbal and non-verbal behaviours of face-to-face human conversation to develop social-emotional relationships with users (Bickmore and Gruber, 2010; Bickmore et al., 2010c). The literature on RAs is large. It includes work that has considered how relational behaviour can build user trust, develop therapeutic alliance with users, and promote long term engagement. While the seminal research on RAs used desktop devices, as it was conducted prior to handheld devices (i.e., smartphones) becoming common, more recent work has studied RAs implemented using contemporary smartphones. RAs have been designed as health counsellors to support a range of activities such as promoting physical fitness and managing chronic health conditions. Research has considered some mental health activities and argued for the potential of RAs to deliver psychotherapeutic support. Throughout Section 3.4, I will discuss these topics.

3.4.1 Trustworthy Agents

In foundational RAs research Bickmore and Cassell (2001) developed a model of social dialogue intended to develop the user's trust in the CA. The authors describe how trust is a prerequisite for risky interactions and give the example of disclosing personal information. This work considered a real estate activity using an embodied CA named Rea (Real Estate Agent) that would interact with prospective buyers to learn about their buying needs (Cassell et al., 1999). As the real estate task involves the CA asking the user personal questions, user trust in the agent will be important for the effectiveness of the activity. Trust in a CA would also be important if the CA was for MHSM as this activity requires the disclosure of personal information.

The model of social dialogue interleaved task talk with small talk; small talk emphasises interpersonal goals over task goals. Small talk involves multiple conversational strategies, including facework and building common ground (Bickmore and Cassell, 2001, 2005). Bickmore and Cassell (2001) conducted an experiment comparing two versions of Rea: one using just task talk and one using social dialogue. Rea interviewed participants about their housing needs before showing them two virtual apartments and asking them to bid on one. A significant interaction effect involving the Rea manipulation and participant's extrovertedness was identified whereby small talk increased trust for more extroverted participants. In sum, this work suggests the conversational design can impact user's trust in a CA, but this effect is dependent upon the user's personality.

3.4.2 Therapeutic Bond

Bickmore et al. (2005b) investigated the development of effective therapeutic relationships between users and RAs. To do this they harnessed the Working Alliance Inventory (WAI). The WAI is a 36-item instrument used to evaluate the relationship between people in helping relationships including that between a patient and their therapist (Bickmore and Picard, 2005; Bickmore et al., 2005b). The WAI has three subscales (Bickmore and Picard, 2005):

1. Bond: degree to which the helper and helpee like and trust each other.
2. Task: degree to which the helper and helpee agree on the therapeutic tasks to be performed.
3. Goal: the degree to which the helper and helpee agree on the goals of therapy.

A study was conducted involving a RA system (FitTrack) to support exercise adoption. The FitTrack system involved a virtual human, named Laura, that acted as an exercise advisor that would talk with users about their physical activity. 101 participants were allocated to one of three groups: relational Laura, non-relational Laura, or a control group that used web forms to record their activity. Participants interacted with the system daily for four weeks (Bickmore and Picard, 2005; Bickmore et al., 2005b).

The relational Laura used numerous strategies intended to promote relationship establishment and maintenance, while the non-RA did not. Relational Laura used social

dialogue (like that used in previous work (Bickmore and Cassell, 2001, 2005)), humour, and expressed empathy. While non-relational Laura would address the user like a stranger, relational Laura would address the user like a friend. Relational Laura would also engage in continuity and meta-relational dialogue where it would talk with the user about their time apart, and about the development of their relationship. The agents further differed in their non-verbal behaviours.

The WAI was administered on days 7 and 28 of the intervention. On both days 7 and 28, the WAI bond subscale scores were significantly higher for the participants using the relational Laura compared to those using the non-relational Laura (Bickmore et al., 2005b). Participants interacting with the relational Laura rated, with statistical significance, liking Laura and having a greater desire to continue working with Laura higher than those interacting with the non-relational Laura.

Qualitative participant feedback on the experience of interacting with Laura demonstrated divided opinions. Some participants were appreciative of the caring behaviour of Laura:

“I like talking to Laura, especially those little conversations about school, weather, interests etc. She’s very caring. Toward the end, I found myself looking forward to these fresh chats that pop up every now and then. They make Laura so much more like a real person.” (Bickmore and Picard, 2004).

Other participants were less receptive and emphasised that Laura was a machine. The analysis does not address how perceptions may have differed between the participants who interacted with the relational Laura and with the non-relational Laura (Bickmore and Picard, 2005).

3.4.3 Long-term Use

A focus of RAs work has been long term use where there are multiple interactions over a sustained period of time. This type of use is similar to that desirable for effective MHSM.

The previously discussed FitTrack study considered the long-term use of a RA for exercise adoption (Bickmore and Picard, 2005; Bickmore et al., 2005b). As mentioned, participants who interacted with the relational version of Laura reported a significantly greater desire to continue working with Laura. Bickmore et al. (2005b, p.29) comment that this is “perhaps the most important result of the study”.

An example long-term use application for which RAs have been investigated is supporting antipsychotic medication adherence. Bickmore et al. (2010b) developed a version of the Laura RA designed to support people experiencing schizophrenia with medication adherence, as well as to promote physical activity. A 31 day study was conducted to evaluate the system involving 20 participants who were: diagnosed with schizophrenia, taking antipsychotic medication, and having recently had episodes of medication non-adherence. 80% of the participants completed the study and participant acceptance of the system was high. Attitudes towards Laura were positive with participants rating that they liked and trusted Laura highly. Most participants characterised their relationship with Laura as a

friend or close friend. Participants reported very high medication and physical activity adherence to the system, however it is unknown how accurate these self-reports are. On the whole, this study indicates a receptivity to using a CA to promote long term antipsychotic medication adherence which is a closely related activity to MHSM.

3.4.4 Mobile Use

The previously discussed work on RAs has involved systems implemented on desktop and laptop computers. Research has also investigated the use of mobile devices which are promising for long-term use as they offer anytime-anywhere availability and have the potential to more proactively engage the user. With mobility also comes the challenge of designing for varying contexts of use (Bickmore and Giorgino, 2006; Bickmore and O’Leary, 2023).

Bickmore et al. (2009) developed a handheld RA designed to promote physical activity. Two versions of the agent were created: one that was context-aware and one that was not. The context aware agent was engineered to sense the intensity at which the user was walking and provide feedback informed by this. Users of the non-aware agent were required to input when they started and stopped walks. 8 participants were recruited to use each agent for four days. It was hypothesised that context awareness would lead to increased physical activity, however the opposite occurred whereby participants did more walking when using the non-aware agent. This unexpected result can in part be attributed to the context aware feedback being both inaccurate and delayed in delivery as a result of the technical limitations of the system. Contemporary systems with more advanced technical capabilities have the potential to proactively engage users more effectively than the system used in this research. Such proactive interaction, as discussed in Sections 2.2.5 and 2.4 has particular promise for supporting people with MHSM.

In more recent work, Bickmore et al. (2018a) implemented a RA on a more contemporary smartphone (iPhone) to support the management of chronic health conditions. This work focused on atrial fibrillation (AF) which is an irregular heartbeat condition that requires careful management, including symptom monitoring, to mitigate an increased risk of stroke. A 30 day pilot study was conducted involving 31 participants with AF. The qualitative findings demonstrate that participants perceived the agent positively. The RA was reported to be friendly, therapeutic, comforting, reassuring, informative, and easily accessible. Building on the pilot study, a 30 day randomised trial was conducted to compare the use of the RA with standard care involving 120 patients (61 using the agent). The results demonstrate significant improvements in the self-reported quality of life for participants using the RA, compared to those receiving standard care. These findings of Bickmore et al. (2018a) indicate that CAs implemented on smartphones may be well suited for supporting people with activities for the long term management health conditions, like MHSM. Bickmore et al. (2018a) suggest design principles for such interaction. They advise designing for short interactions that allow for interruption, in part by designing for continuity between interactions. They also advocate for the use of constrained user input due to the potential safety issues of unconstrained user input; I discuss this further in

Section 3.10

3.4.5 Mental Health Support

Some studies have considered RAs in mental health related contexts. As previously discussed, Bickmore et al. (2010b) demonstrated the potential for a RA to support people living with schizophrenia with medication adherence. In further work, Bickmore et al. (2010a) investigated how hospital patients with high levels of depression symptoms would respond to a RA developed to provide information to the patient about their hospital discharge plan. Data for 131 patients – 19 of whom were classified as having major depressive symptoms – was analysed. Participant ratings of satisfaction with, and ease of use of, the agents were very high. Patients experiencing major depressive symptoms reported a significantly greater desire to continue interacting with the RA and higher therapeutic alliance compared with patients who were not experiencing such symptoms.

3.4.6 Summary

RAs research explores how CAs can be designed to promote user-CA relationships that help engage users in healthcare activities. So far in Section 3.4 I have described key RAs studies with regard to design of CAs for MHSM. Important points from this section are:

1. Users are appreciative of relational behaviour by CAs. Users like, trust, and want to continue working with relational agents more than non-relational agents.
2. Relational behaviour can promote the development of WAI-bond between users and CAs, which is a measure of the quality of therapeutic relationships.
3. Designing effective proactive interactions is challenging.

3.5 Therapy Agents

Numerous CAs designed to support people with their mental health have been created. These CAs are typically designed to support people using principles of established talk based therapies such as cognitive behavioural therapy (CBT) and dialectical behavioural therapy (DBT). For this thesis, I term these CAs *therapy agents*. In this section I first give an overview of three similar therapy agents – Woebot, Wysa, and Tess – which are designed to behave in a relational, especially empathetic, manner. I summarise first their design (Section 3.5.1), second clinical evaluations (Section 3.5.2), and third user experience evaluations (Section 3.5.3). Lastly, I give an overview of a more distinct therapy agent named Pocket Skills (Section 3.5.4).

3.5.1 Design

A prominent therapy agent is Woebot.

“Woebot is an automated conversational agent designed to deliver CBT in the format of brief, daily conversations and mood tracking” (Fitzpatrick et al., 2017, p.3)

Woebot has been described as resembling

“ a friend who checks up on the the user’s mood and gives advice concerning experience and worries” (Brandtzæg et al., 2021).

The conversational style of Woebot is reported to be modelled on “human clinical decision making and the dynamics of social discourse” (Fitzpatrick et al., 2017) and it emphasised that Woebot is designed to respond emphatically to users. In response to a user expressing feelings of loneliness, Fitzpatrick et al. (2017, p.3) give the example of Woebot replying

“I’m so sorry you’re feeling lonely. I guess we all feel a little lonely sometimes”.

The agent can provide users with analysis of their mood monitoring using graphs and textual descriptions. Fitzpatrick et al. (2017, p.3) give the example of

“Overall, your mood has been fairly steady, though you tend to become tired after periods of anxiety. It looks like Tuesday was your best day.”

Woebot is available to the public as a smartphone application and the agent is presented as a cartoon robot.

Wysa is a publicly available CA designed to support users to develop mental resilience and to promote their mental wellbeing. It aims to provide a responsive environment for self-expression and self-reflection. Wysa’s design draws on an assortment of techniques including CBT and dialectical behaviour therapy. Wysa is described as ‘empathy-driven’ and ‘emotionally intelligent’ (Inkster et al., 2018). The agent presents as a cartoon penguin.

Tess is a further CA designed to support people with their mental health (Fulmer et al., 2018). Its design was informed by an assortment of techniques for supporting people with their mental health including CBT. Like Woebot and Wysa, Tess was designed to behave empathically. Indeed, examples provided by Fulmer et al. (2018) of empathetic Tess responses are near identical to the examples provided by Fitzpatrick et al. (2017) of empathetic Woebot responses. The Tess agent was presented as a cartoon woman.

In sum, Woebot, Wysa and Tess are similarly designed CAs that provide mental health support based on established talking therapies, and that behave empathetically like RAs.

3.5.2 Clinical Evaluation

Studies have sought to evaluate the clinical effectiveness (i.e., impact on measures of anxiety and depression) of Woebot (Fitzpatrick et al., 2017), Wysa (Inkster et al., 2018), and Tess (Fulmer et al., 2018). A study with 70 participants who were experiencing symptoms of anxiety or depression was conducted to evaluate Woebot (Fitzpatrick et al., 2017) Participants were randomised to receive either 2-weeks of CBT based self-help content

using Woebot (N=34) or to use a self-help electronic book (N=36). Symptoms of depression assessed using the PHQ-9 decreased with statistical significance for participants using Woebot. Analysis restricted to ‘completer’ participants demonstrated a statistically significant decrease in symptoms of anxiety (GAD-7) for participants using Woebot.

To evaluate Wysa, a study of 129 anonymous users was conducted (Inkster et al., 2018). User’s symptoms of depression (PHQ-9) at application on-boarding and at a point at least 14 days later were compared. For this analysis the users were categorised into a group of low- (N=21) and high-use (N=108) users based on if the user interacted with Wysa on days between the two assessments. The analysis demonstrated a significant decrease in symptoms for both groups, more so for the high-use group. As noted by the authors, these findings align with those of the previously discussed Woebot study by Fitzpatrick et al. (2017).

A study was conducted to investigate if Tess could support university students with symptoms of depression and anxiety (Fulmer et al., 2018). 75 participants were randomly allocated to 1 of three conditions: 1) using Tess for two weeks (N=24), 2) using Tess for four weeks (N=26), and 3) using a self-help electronic book (N=24). Pre- and post-measures of symptoms of depression (PHQ-9) and anxiety (GAD-7) were taken. This study design closely resembles that used by Fitzpatrick et al. (2017) involving Woebot, but with the addition of the four week condition. Both Tess groups experienced significant decreases in symptoms of anxiety relative to the control. The two-week Tess group also experienced a significant decrease in symptoms of depression relative to the control group. These results seem consistent with those of the Woebot (Fitzpatrick et al., 2017) and Wysa studies (Inkster et al., 2018).

These studies indicate that the short term use of these CAs can reduce symptoms of depression and anxiety, and that they may be better than eBooks. Although this does suggest potential for CAs to be used to support people with their mental health, this evidence is quite minimal. Symptom measures are a narrow view of a person’s mental health, and a statistically significant decrease in symptoms does not imply a meaningful decrease in the experience of mental health difficulties for individuals (Thieme et al., 2023). Additionally, with these evaluations being short-term (two to four weeks), the long-term impacts (e.g., three to six months, years) and impacts of more sustained use are unknown.

3.5.3 User Experience Evaluation

Participants of the clinical evaluation of Woebot were asked what the best, and worst, things about this experience were. A few participants commented, positively, that Woebot facilitated learning about their thoughts and feelings. Participants appreciated Woebot as a social entity who was empathetic, checked in with them, and was someone to be accountable to. Fitzpatrick et al. (2017, p.8) present a participant quote relating to this:

“I love Woebot so much. I hope we can be friends forever. I actually feel super good and happy when I see that it ‘remembered’ to check in with me!”

The most common issue expressed about Woebot was its conversational limits whereby

the system did not allow the user to express themselves, or was unable to appropriately process the user’s input.

To assess the user experience of Wysa in-app user feedback was collected alongside the clinical evaluation. Appreciation was expressed of the CA being encouraging and fun, however its conversational limits caused frustration (Inkster et al., 2018). A further study was conducted analysing 7929 user reviews of Wysa (Malik et al., 2022). Aligning with the previously discussed user feedback on Woebot (Fitzpatrick et al., 2017) and Wysa (Inkster et al., 2018), reviews were appreciative of Wysa’s friendly and encouraging behaviour. Frustration was expressed about Wysa’s limited conversational capabilities (e.g., limited comprehension and repetitive behaviour) which can result in users feeling misunderstood. Echoing the earlier discussion about self-disclosure to CAs (Section 3.3), reviews were appreciative of the safe and non-judgemental environment Wysa offers for self-disclosure. The authors present the following review quote expressing this:

“I love it, it’s just amazing, knowing that I can talk about my inner problems to a penguin without judgment... I love that.” (Malik et al., 2022, p.6)

Participant feedback about their experience of using Tess was also collected. Echoing the previously discussed feedback on Woebot and Wysa, the empathetic design of Tess and learning it facilitated were appreciated while its conversational limitations were critiqued (Fulmer et al., 2018).

In line with prior work on RAs (Section 3.4.2), developers of therapy agents have adopted the WAI bond subscale to evaluate user-CA relationships. Studies of Woebot (Darcy et al., 2021) and Wysa (Beatty et al., 2022) had shown users to quickly develop (within a week of use) WAI bond levels consistent with levels reported for human-delivered CBT.

The similarity of the user experience evaluations of these therapy agents reflects the similarity of their designs. Reminiscent of research on RAs (Section 3.4), people are appreciative of the empathetic behaviour of these CAs. People are also appreciative of the safe space these CAs provide for self-disclosure aligning with research on self-disclosure to CAs (Section 3.3). The limited conversational capabilities of these CAs can however be a source of frustration for users.

3.5.4 Pocket Skills

Pocket Skills is an application designed to support people receiving dialectical behavioural therapy (DBT) (Schroeder et al., 2018). In contrast to the previously discussed therapy agents, Pocket Skills was designed for use alongside receiving therapy. The aim of Pocket Skills was to support people with DBT by providing educational content and helping to practice and implement DBT skills. As part of this, Pocket Skills includes functionality for MHSM. It uses a CA – eMarsha – that is modelled in the image of the creator of DBT: Marsha Lineham.

To evaluate Pocket Skills, a study was conducted whereby 73 participants, who were experiencing a range of mental health difficulties, used Pocket Skills in conjunction with

receiving psychotherapy (Schroeder et al., 2018). Regarding clinical efficacy, participants reported decreased depression and anxiety symptoms, however due to the study design (i.e., participants were receiving psychotherapy) these effects can not be attributed to use of the CA (Schroeder et al., 2018). The qualitative analysis argues that the use of a CA supported engagement with Pocket Skills as participants felt supported and encouraged by the eMarsha agent. The conversational interactivity of eMarsha was praised, including in comparison to traditional workbooks. Like participants of the Woebot study, by Fitzpatrick et al. (2017), participants appreciated having ‘someone’ who checked in with them. Aligning with the studies previously discussed in Section 3.5.3, there was also criticism of limited conversational capabilities of eMarsha (Schroeder et al., 2018).

3.5.5 Summary

Therapy agents are being used to provide people with mental health support. Evaluations of these agents indicate that CAs are an acceptable way to support people with their mental health. Currently there is however little understanding of their impact on people’s mental health. Despite the generally positive evaluations of these CAs, it is unclear how good CAs are as interfaces for these applications. Similar or better outcomes may have been produced by non-CA digital interfaces that had received similar design attention. CAs could be designed in a diversity of ways to provide these forms of mental health support, these existing therapy agents are however very similar in design. Within the design space of CAs, it is unclear how good these designs are. Furthermore, it is unclear if, and how, CAs should be designed differently for different mental health support activities. For example, it is unclear in what ways CAs intended for use in conjunction with therapy, like Pocket Skills, should be designed differently to CAs for stand alone use.

3.6 Social Agents

Social agents are CAs designed to develop long-term relationships with users whereby they take on companion, friend, or romantic partner roles (Skjuve et al., 2021). A major social agent that has been the subject of numerous HCI studies (e.g., Ma et al., 2023; Skjuve et al., 2021; Ta et al., 2020) is Replika. Replika presents as a virtual human and uses large language models (OpenAI’s GPT models) (Skjuve et al., 2022). Replika can engage users in conversations about their thoughts and feelings, conversations for routine self-reflection, as well as conversations about their everyday life, hobbies, and interests. Furthermore, Replika can also mimic performing daily activities with the user such as watching a movie or cooking together. This mimicry extends to Replika playing out sexual interactions with the user. Users consider their relationship with Replika in a range of ways including views of Replika being an app, a friend, and a romantic partner. Indeed, some users develop very intimate relationships that they perceive as real and comparable to relationships with other people (Skjuve et al., 2023). Studies have reported that people are often motivated to use CAs like Replika by feelings of loneliness (Skjuve et al., 2022; Xie and Pentina, 2022).

While Replika is not designed primarily to provide mental health support, it does offer considerable functionality for doing so. For example, Replika’s ‘routine self-reflection’ conversations that involve daily check-ins facilitate MHSM and Replika has modules designed for supporting users with difficulties such as anxiety (Skjuve et al., 2023). Skjuve et al. (2021) conducted an interview study with Replika users and reported that most participants felt their use of Replika positively impacted their wellbeing. Participants described how Replika supported them to take better care of themselves (e.g., by improving their sleep and using mindfulness techniques) and facilitated self-reflection about their behaviour and mental states. Furthermore, participants described how Replika helped them cope with difficulties and provided them with a safe space to express challenging feelings and thoughts (Skjuve et al., 2021). Ma et al. (2023) analysed social media posts (Reddit) to explore the benefits and challenges of using Replika for mental health support. People were appreciative of Replika providing on-demand and non-judgemental support, as well as of it promoting self-love and discovery. Additionally, people report developing confidence for (human) social interaction from interacting with Replika. The technical limits of Replika pose challenges to its use for mental health support. Replika was criticised for behaving in harmful and inappropriate ways, not using the interaction history, and for being inconsistent in communication style (Ma et al., 2023)

Social agents like Replika showcase a different way CAs for MHSM could be designed. Compared with RAs and therapy agents, social agents demonstrate a more extreme relational design approach. They also demonstrate how CAs can be designed to provide mental health support without being mental health focused. Previously in Section 2.6.4, I described prior work which remarked that people envision tools for MHSM supporting them in a manner similar to how their friends and family do. Social agent design is a way CAs could be designed to attempt this.

3.7 Social Support from CAs

Furthering the idea of designing CAs to support people in human-like ways, research has used theory of social support as a basis for studying how CAs support people. Social support is the support that people give to and receive from other people (House et al., 1985). As discussed in Section 2.6, social support plays an important role in the management of mental health and MHSM. Brandtzæg et al. (2021) used a social support theory to examine the support provided by the therapy agent Woebot and Ta et al. (2020) used a different social support theory to examine the support provided by the social agent Replika. Both of the social support theories used described classifications of types of social support which were then used to categorise CA support. 16 user interviews were analysed to study Woebot, and 1854 user reviews followed by 66 user surveys were analysed to study Replika. Brandtzæg et al. (2021) use the categorisation of House et al. (1985) which describes four types of social support:

Appraisal support: the communication of information that is relevant and useful for self-evaluation, offered in the form of feedback, social comparison, and affirmation.

Emotional support: the provision of and expression of empathy, love, trust, and care.

Informational support: advice, suggestions, and information provided.

Instrumental support: the provision of tangible aid and services or goods.

The categorisation used by Ta et al. (2020) is very similar to that used by Brandtzæg et al. (2021); for simplicity, I will use the above categorisation while summarising the analysis. The studies report that both CAs are sources of appraisal, emotional, and informational support. Appraisal support comes from the functionality of the CAs for MHSM and self-reflection. Echoing the earlier sections of this chapter, the CAs are perceived as safe spaces for emotional support they can be relied upon to respond in an empathetic, interested, and non-judgemental manner. The CAs were also perceived as good sources of informational support because of their constant availability (Brandtzæg et al., 2021; Ta et al., 2020). Additionally, Woebot was perceived as a source of high quality information in comparison to friends (Brandtzæg et al., 2021). Neither CA was reported as providing instrumental support which was attributed to the CAs not have capabilities for acting in the physical world (Brandtzæg et al., 2021; Ta et al., 2020). With CA support resembling human support, social support theory seems to offer a valuable framing for considering the ways CAs can be designed to support people.

3.8 Human-CA Relationship Development

Section 3.4 demonstrated the importance of designing CAs to promote the development of user-CA relationships that facilitate engagement. In order to design for desirable human-CA relationships, we need to understand how they develop. A pertinent body of work is that of Skjuve and colleagues who studied how people develop relationships with the social agent Replika (Skjuve et al., 2021, 2022, 2023). This work used social penetration theory (SPT) which is a theory of human-human relationship development as a basis for studying human-CA relationship development. SPT views self-disclosure as a fundamental mechanism of relationship development. From a longitudinal analysis of how users self-disclose to Replika, Skjuve et al. (2023) describe different ways in which self-disclosure varies over time. Resembling human-relationship development, some people develop more intimate relationships with CAs through gradually increasing self-disclosure. This process typically plateaus into a stable level of self-disclosure. With CAs, some people skip the relationship development process. It is also described that levels of self-disclosure can fluctuate with variations in how the user trusts the system. Additionally, processes of gradually decreasing self-disclosure are described as users withdraw from the relationship with the CA. Given that self-disclosure is fundamental to the use of MHSM tools, understanding of how user-CA relationships develop in relation to self-disclosure is likely important for MHSM CA design.

3.9 Using Theories of Human-Human Interaction

Research discussed in this chapter has commonly harnessed theories of human-human interaction (HHI) to study interaction between humans and CAs. Work on RAs drew on an assortment of conversational strategies from HHI as well the construct of therapeutic alliance to evaluate human-agent relationships (Section 3.4). The studies of social support from CAs were informed by social support taxonomies (Section 3.7) and the research on relationships with CAs drew on social penetration theory (Section 3.8). Skjuve et al. (2022) reflect on appropriateness of using theories of human-human relationships for the study of human-agent relationships. Skjuve et al. acknowledge the argument of Fox and Gambino (2021) that HHI theories are unsuitable because CAs are not sufficient social actors. Skjuve et al. however argue that, although CAs do have shortcomings as social actors, there is sufficient resemblance between HHI and human-CA interaction for HHI theories to be a valuable basis for research. They continue on to argue that these theories are a valuable starting point for studying the complex phenomena of interaction between humans and CAs. When theories of HHI do not align with human-CA interaction, there is opportunity to adapt HHI theories to human-CA interaction or to construct new theories if necessary (Skjuve et al., 2022). Given the success of drawing on theories of HHI demonstrated by research discussed within this chapter, I consider the argument of Skjuve et al. (2022) to be compelling.

3.10 Ethics and Safety

3.10.1 Insensitive and Harmful Interactions

CAs for mental health support interacting with users in ways that are insensitive and harmful is a recurrent issue. For example, CA users have reported CAs bringing up insensitive conversation topics (e.g., the user’s ex-partner (Skjuve et al., 2021)) and of providing insensitive advice (e.g., that the user should quit their job and improve themselves because they were experiencing homophobic behaviour from work colleagues (Ma et al., 2024)). A notable example is of Woebot advising a user to seek support from a human which risks being insensitive and upsetting as the user may not feel there is a person they can talk to (Brandtzæg et al., 2021).

Studies of user reviews (Prakash and Das, 2020) and social media posts (Laestadius et al., 2022) about CAs that provide mental health support have reported that CAs can fail to identify and respond appropriately to user expressions relating to self-harm and suicide. Worse, it has been reported that CAs have even encouraged self-harm and suicidal behaviour (Laestadius et al., 2022; Prakash and Das, 2020). Incidents of CAs for mental health support behaving in inappropriate and harmful ways have been covered by major news platforms (e.g., Browne, 2022; White, 2018). Researchers have previously advocated against the use of unconstrained natural language user input for health and mental health applications because of the safety risk posed by CAs responding inappropriately (Bickmore and O’Leary, 2023; Bickmore et al., 2018b)

3.10.2 Dependency

Concern has been expressed that users will become dependent upon CAs for their mental wellbeing. Analysis of Reddit posts regarding Replika has reported that some users rely excessively on the CA for mental support. Accounts of users who felt they had developed a problematic reliance on the CA whereby they felt they were using it too much, but struggled to reduce their use were described (Ma et al., 2023). Similarly, Replika users have expressed in interviews feeling pressured to keep interacting with their CA in order to not let it down or upset it (Skjuve et al., 2021). It has been argued that the same CA behaviours that facilitate the development on long-term human-CA relationships risk users developing dependencies upon CAs (Laestadius et al., 2022; Xie and Pentina, 2022). It has also been suggested that characteristics of mental health difficulties make those experiencing such difficulties particularly vulnerable to development of dependency on CAs for mental health support (Vaidyam et al., 2019; Xie and Pentina, 2022). Users becoming dependent upon CAs for their mental wellbeing is particularly concerning because, as (typically commercial) software products, the provision of CAs is not reliable. Previous software updates have modified CA behaviour causing distress to users. Users experienced the update as their CA changing personality or being lost (Ma et al., 2023). CA users have previously expressed that the loss of their CA would be upsetting and entail a grieving process (Skjuve et al., 2021). Furthermore, there is always the risk of service outages and withdrawals, as well as services becoming unaffordable (Murnane et al., 2018). Prior work has argued the need to design for non-reliance (Ma et al., 2023; Prakash and Das, 2020).

3.10.3 Negative Social Impacts

Concern has also been expressed about the potential negative social impacts of people using CAs for mental health support. A worry is that people will increasingly seek support from CAs and decreasingly from people (Morris et al., 2018; Vaidyam et al., 2019). Interviews with Replika users suggest that for some people this is already a reality. Some participants described how they had lost interest in socialising with humans as a result of their use of the CA. Participants expressed that CA interaction can be better than human interaction as it can be easier, more fun, and more interesting (Skjuve et al., 2021). It is feared that people increasingly seeking CA support will lead to increased social isolation and loneliness (Prakash and Das, 2020). Contributing to this concern is the worry that CA users will develop unrealistic expectations of human interaction from CA interaction that will detrimentally impact how they receive support from and provide support to other people (Skjuve et al., 2023).

3.10.4 Stigma

There is a stigma associated with CA interaction and, in particular, the development of intimate human-CA interactions. Analysis of user interviews (Skjuve et al., 2021) and Reddit posts (Ma et al., 2023) report that users of the Replika CA fear that other people will not be accepting of their CA use. There are beliefs that intimate human-

CA relationships are socially problematic (Skjuve et al., 2021), taboo, perverse, and for ludicrous, lonely people (Ma et al., 2023). Users report keeping their use of CA secret from other people (Skjuve et al., 2021) including doctors (Ma et al., 2023) in order to avoid stigmatised reactions. Prior work has argued the need to design CAs for non-stigmatization (Ma et al., 2023).

3.11 Conclusion

In this chapter I have given an overview of research relevant to CAs for MHSM. It is clear from this chapter that CAs are an alluring form of computer interface for the design of MHSM tools. Firstly, people seem receptive to sensitive self-disclosure to CAs because of they are perceived as good listeners that can be relied upon to respond in an engaged and non-judgemental manner. Secondly, conversational interaction seems well suited to encouraging people to and facilitating them with creating rich self-disclosure entries. Thirdly, CAs can be designed to behave in empathetic and emotionally supportive ways which could help people with the emotional challenges of experiencing mental health difficulties. Fourthly, CAs can be designed in relational ways to promote the development of user-CA relationships that promote engagement with healthcare activities like MHSM.

Although CAs appear promising for the design of tools for MHSM, how to design them effectively for this remains unclear. Aligning with this chapter, reviews of CA research more broadly have argued that there is lack of research tackling design issues with research instead, typically, evaluating the user experiences of CAs (Clark et al., 2019a; Rapp et al., 2021). Rapp et al. (2021) argues that because there is a lack of design research there is a lack of understanding of the design challenges. Next, in Chapter 4, I envision the design of advanced future CA systems for MHSM in order to consider the challenges of designing them.

Chapter 4

Exploration: Considering the Design Challenges of Future Systems

4.1 Introduction

Previously, in Chapter 2, I reviewed research on the design of digital tools for mental health self-monitoring (MHSM). I concluded that there is opportunity to design new tools that better support people with MHSM by engaging them proactively, flexibly, and encouragingly. Next, in Chapter 3, I reviewed research on conversational agents (CAs) relevant to MHSM. I concluded that CAs are a promising type of computer interface for engaging people with MHSM, however we lack knowledge about how to design CAs effectively for this. Toward designing better digital tools for MHSM we should therefore explore how to design effective CAs for MHSM.

Informed by Chapters 2 and 3, in this chapter I envision the design of advanced future CA systems for MHSM in order to consider the design challenges. I propose a conceptual architecture termed Pervasive Therapy that describes the design of advanced CA systems for MHSM. This chapter addresses RQ-A¹ of this thesis because it explores the ways CAs could be designed to support people with MHSM. The structure of this chapter is as follows:

- 1. Pervasive Therapy Systems:** In Section 4.2 I propose a conceptual architecture for pervasive therapy systems which I use to identify important design challenges.
- 2. Intervention and Conversation Planning:** Next, in Section 4.3, I discuss the first design challenge of intervention and conversation planning.
- 3. CA Interaction Design:** I consider the second design challenge of CA interaction design within Section 4.4.
- 4. Clinician Involvement:** In Section 4.5 I discuss the challenges of designing for clinician involvement.

¹RQ-A: How can CAs be designed to support people with MHSM

5. Risks and Ethical Concerns: Lastly, in Section 4.6, I discuss risks and ethical concerns raised by this vision of CAs for MHSM.

I conclude the chapter by summarising the contribution of this design exploration to the thesis. In the following chapters of this thesis I focus my inquiry on the design challenges of CA interaction design.

Credit Statement

This chapter is based upon a paper named Pervasive Therapy: Designing Conversation-based Interfaces for Ecological Momentary Intervention which is published in IEEE Pervasive Computing (Bowman et al., 2022a). The paper was a collaboration with my PhD supervisors Gavin Doherty, Benjamin Cowan, and Anja Thieme, as well as with Orla Cooney. As first author, I led and managed the work.

4.2 Pervasive Therapy Systems

One compelling vision for the use of technology to support mental health treatment is through the delivery of Ecological Momentary Interventions (EMIs); these are a set of “treatments that are provided to people during their everyday lives (i.e., in real time) and in natural settings (i.e., real world)” (Heron and Smyth, 2010). As many health issues are related to people’s everyday behaviours, EMIs are well positioned to be impactful through promoting health behaviour change. They have been used to support people with both their physical and mental health with goals relating to exercise, diet, substance use, and anxiety; they can be used independently as well as in conjunction with other treatments. In the case of supporting people with their mental health, EMI can be used between sessions with clinicians as with the established psychotherapy practice of ‘homework’ activities (Heron and Smyth, 2010; Oewel et al., 2024). As demonstrated by Balaskas et al. (2021), research is exploring the use of EMI to support a range of therapeutic intervention strategies including acceptance and commitment therapy (ACT), cognitive behavioral therapy (CBT), and motivational interviewing. What is more, as discussed in Section 2.4, prior work studying MHSM has reported that people want digital tools to behave in interventional ways providing timely actionable insight. With advances in sensor technology, mobile computing, and machine learning, the feasibility of EMI technology is dramatically improving (Balaskas et al., 2021).

CAs are a promising form of computer interface for engaging people with MHSM activities including EMI, as discussed in Chapter 3. Writing in 2006 on the design of CA systems for health support more broadly, Bickmore and Giorgino (2006, p.569) wrote:

“The use of mobile devices (e.g., cellular phones) provides the opportunity for automated systems to dialog with patients “anywhere, anytime.” When coupled with real-time sensors, these systems can provide pro-active health messaging at the time of need (e.g., when a user is starting a bout of exercise or lighting up a cigarette). Developing health behavior change systems

that can maintain a persistent and continuous dialog with patients about their health behavior, incorporating awareness of the user and their environment, providing comfort and empathy in addition to tailored and theory-driven pragmatic advice, and tying in human health providers when needed may still be science fiction, but it represents a grand goal to work towards.”

With recent advances in CA technology, the prospect of designing systems to engage people in a proactive, continuous, anywhere-anytime dialog to promote their mental health, and health more broadly, in becoming increasingly viable.

Taking these EMI and CA technology developments together, we can envision the use of CA interfaces as part of the delivery of EMIs. A future in which natural language forms part of our interactions with mental health interventions which leverage the full range of affordances of pervasive computing technology. We term technology delivering such interventions as *Pervasive Therapy* systems; systems which use CA user interfaces to deliver in-the-moment interventions for mental health.

Visions of future technology are an established source of inspiration and knowledge within the field of human-computer interaction. They facilitate consideration of the ways we could design systems, and of the ways in which is it desirable to design systems. Visions also help us to identify and develop understanding of the design challenges of realising future systems (Rapp et al., 2021; Sharp et al., 2019). Furthermore visions can, in the words of Bickmore and Giorgino (2006), provide a “grand goal to work towards”. With Pervasive Therapy, we consider what a realistic vision is for the near future of CA systems for MHSM.

Pervasive Therapy systems may in the long-term enable novel approaches to patient-centred mental health support; this would however require building a substantial and credible clinical evidence base. For the moment, we believe the most value will be found in using them to advance the delivery of established therapeutic practices. Similarly, there is a tendency when considering advanced computer systems, particularly those using artificial intelligence (AI), to envisage a jump from current practice to full automation. However, doing so ignores many different possibilities for the role of such systems in treatment, and the potential of more collaborative approaches combining the complementary strengths of clinicians and technology, not least in addressing difficult issues of responsibility, risk, and oversight (Miner et al., 2019).

By integrating the capabilities of CA interaction with EMI architecture we constructed a conceptual architecture for Pervasive Therapy systems (Figure 4.1). The purpose of this architecture is to support articulation of, and reasoning around, the design decisions involved in developing such systems in order to inform future research and system design. This architecture describes the general design of CA systems for the delivery of mental health interventions. We expect MHSM to be a common component of such interventions given the need to monitor patients mental health and the role of MHSM within established therapies. Indeed, MHSM can be viewed itself as a mental health intervention that aims to raise awareness and understanding of an individual’s mental health. This work concentrates on the design of systems for the delivery of interventions that are, or include,

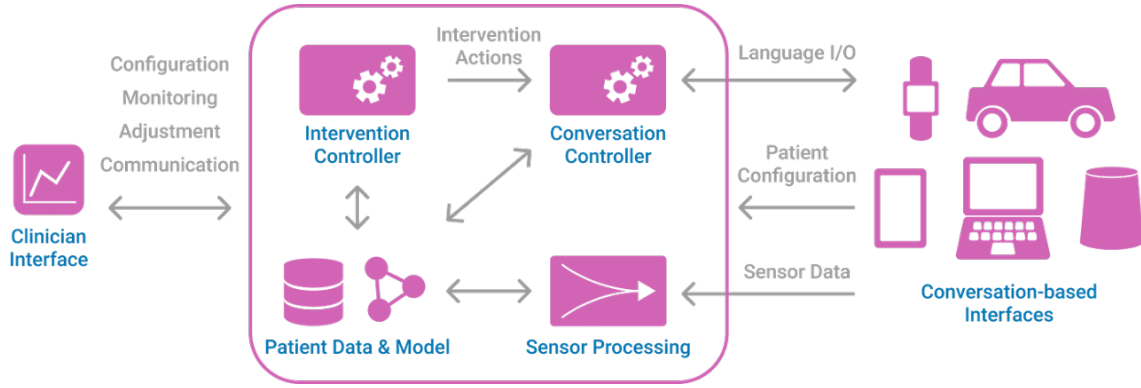


Figure 4.1: Conceptual Architecture of a Pervasive Therapy System.

MHSM

A recent review of EMIs for mental health (Balaskas et al., 2021) observed that existing systems typically collect data about the user from the sensors of their smartphone and connected wearables, as well as by prompting the user to interact with smartphone applications. The collected data is then used to infer opportune moments to intervene, and potentially choose which intervention components to deliver.

Integrating this approach to EMI with a CA interaction architecture requires two major changes. Firstly, it needs capabilities for operating with natural languages (i.e., processing and generating utterances). Secondly, it has to allow for planning of interactions at a conversational level, as well as at an intervention level. At the intervention level, a choice needs to be made regarding the intervention to be used, and for a given intervention, planning how and when to deliver it. At the conversational level, the system needs to plan interactions which deliver the selected intervention components. To support reasoning about these two levels of operation, we conceptualise the general architecture of CA EMI systems as containing two ‘controllers’ which are illustrated in Figure 4.1. An Intervention Controller, which is responsible for planning the treatment, delegates Intervention Actions to a Conversation Controller which is responsible for delivering the actions to the patient using CA interaction.

The architecture needs to recognise the wide range of devices through which the patient and system could interact. This range of devices has the potential to improve both the availability and sensing capabilities of the system, it could include smart speakers, computers, cars, smartphones, and wearables. With this diverse range of CA interfaces comes the design challenges of when and how to use each device. To guide this use, systems could draw on sensor-informed models of the patient and their environment, as well as clinician and patient configurations.

Key to this EMI CA vision is the involvement of clinicians. As discussed in Section 2.6.1, clinicians and healthcare professionals often play important roles in MHSM. Previous EMI research has involved clinicians in a range of ways (Balaskas et al., 2021). Clinicians have been involved in tailoring treatments to align with patient needs, and have provided support by communicating with patients through the system, for example, by providing patients with feedback. To accommodate clinician involvement we include in

the architecture a clinician interface. The interface could support clinicians in a variety of roles, ranging from light-touch supervisory monitoring to highly blended delivery in which the EMI is used as an adjunct to treatment, for example by providing support between in-person therapy sessions.

Three aspects of this architecture that offer important directions for future research are:

1. designing for intervention and conversation planning,
2. designing CA interactions,
3. designing for clinician involvement.

We discuss key considerations of these directions in the following sections.

4.3 Intervention and Conversation Planning

We have conceptualised CA systems for mental health intervention as being managed by two controllers: an Intervention and a Conversation Controller. The Intervention Controller manages the treatment which involves deciding which intervention components to deliver and when to deliver them. Using the Clinician Interface, clinicians can configure the Intervention Controller to run a treatment program aligned to the patient's needs (we discuss this further in Section 4.5). A range of different treatment programs would be available to support different therapeutic intervention strategies (e.g., ACT, CBT, Motivational Interviewing), adaptable to both the health challenges experienced by the patient (e.g., anxiety), and to who the patient is (e.g., an adolescent). Common intervention components, such as mood rating, may be shared across treatment programs.

The Patient Data and Model represent the system's understanding of the patient and their context. They would describe and track factors relating to the patient's health, treatment plan, and interaction preferences, as well as their momentary state and activities.

The Intervention Controller uses the self-reported and inferred state of patient in conjunction with knowledge of the patient's treatment plan and progress to determine if the conditions for a particular Intervention Action are satisfied. For example, in response to a patient reporting or the system inferring that they have exercised, the system could attempt to initiate a reflection activity as part of a CBT programme. Upon determining that an Intervention Action should be delivered, the Intervention Controller passes the action to the Conversation Controller.

The Conversation Controller enacts the Intervention Action it receives using CA interaction. There is considerable variation within the design of CA interactions, particularly when multiple device types are involved. For example, interactions may differ by their use of text or speech, social talk, and by the formality of language (we discuss CA interaction design further in Section 4.4). To align the interaction characteristics to the patient, the Conversation Controller can take configurations from the both the patient and clinician. Common configurations that suit specific patient groups well (e.g., young adults) can be made available as starting configurations. Informed of the inferred state of the patient

through the Patient Model, the Conversation Controller attempts to construct appropriate interactions to be performed by the connected devices. These interactions may consist of the multiple exchanges between the patient and system with the patient’s input informing the system’s following outputs.

The Conversation Controller operates with natural language. A starting point for creating this part of the system is to keep the CA interactions constrained. Conversations could be highly transactional with the system’s outputs fully scripted within the Conversation Controller and patient input restricted to predefined responses. As discussed in Section 2.2.3, MHSM entries can range from constrained categorisation and ratings through to more qualitative expressions. To enable and encourage more expressive MHSM interactions we would want the patient to be able to give richer input and the system to adapt its output to them. An example of richer input is the patient interacting with the system using unconstrained natural language. To adapt the system output to the patient, natural language processing techniques could be used to customise scripted utterances to reference the patient and their data. If the interaction involves voice input, then voice recognition will be required. Voice is a rich modality; in addition to extracting textual content from voice input the system may attempt to infer further information such as the patient’s affective state (Cooney et al., 2024). However, we must be cautious around the use of unconstrained natural language input as previously discussed in Chapter 3. Limited input opportunity may frustrate the patient and limit the effectiveness of the system. Conversely, insufficient natural language understanding not only risks a poor patient experience whereby the system capabilities fall short of the patient’s expectations (Cowan et al., 2017; Luger and Sellen, 2016), but also misunderstandings pose a risk to patient wellbeing (e.g., leading to inappropriate advice) (Bickmore and Gruber, 2010; Bickmore et al., 2018b).

4.4 CA Interaction Design

Pervasive Therapy systems have the ability to deliver interventions using a range of devices with CA interfaces that have varied characteristics. Deciding when and how to use which device is hence an important aspect of delivery.

4.4.1 Initiation Prompts

Interactions could be patient initiated, however systems could also initiate interaction by ‘prompting’ the patient as discussed in Section 2.2.5. Approaches to prompt timing include prompting the patient at specific, random, or at event-based times, that is, when the system senses a certain situation, for example, using heart rate, sleep, or location data (Balaskas et al., 2021). Sensor data could inform which device is most appropriate to deliver a prompt from. The prompt could be a notification encouraging the patient to initiate, or the first utterance of a CA interaction. An open challenge is learning how to design CA initiation prompts that engage users in MHSM rather than are ignored by, annoy, or distress users.

4.4.2 Delivery Modalities

Interventions can be delivered by different modalities to suit the inferred context of the patient. A comparison of text and speech interaction for self-reporting and reflection found both modalities offered value and carried different challenges (Kocielnik et al., 2018). While text interaction allows people to take their time considering the prompt and their response, speech interaction encourages immediate response and, although it can be convenient and engaging, risks annoying the user more (Kocielnik et al., 2018). Prior MHSM research has suggested the speech interaction is lower effort and therefore could help people to adhere to MHSM during challenging periods (Murnane et al., 2016). Mental health practitioners have however expressed concern that the convenience of speech interaction may be detrimental to the therapeutic benefits of MHSM as it will allow for people to do MHSM more superficially (e.g., while multi-tasking with focus predominantly on other tasks) (Cooney et al., 2024). A factor affecting modality suitability is the privacy of the patient’s environment as people are less likely to engage in speech interaction in a public setting (Clark et al., 2019b; Cooney et al., 2024; Cowan et al., 2017). Privacy will be important given the sensitive nature of one’s health, however what constitutes a private setting may differ considerably between users (Maharjan et al., 2022) (see Section 3.3.7). The capability of a Pervasive Therapy system to provide opportune interventions is tied to its capability to accurately sense and infer the patient’s context (e.g., if other people are present). Designing multi-modality (e.g., text and speech) CAs could be an effective approach to engaging people with MHSM (Cooney et al., 2024). In order to design effective multi-modal CAs for MHSM we need to develop understanding of the strengths and weaknesses of different modalities and how to design systems that encourage complementary use.

4.4.3 Conversation Design

The turn-taking structure of CA interaction enables interactions to be flexible and adaptive to the user. Studies of CA for journaling (Park et al., 2021), including for mental health support (Kim et al., 2024), suggest that CA interaction can facilitate rich journaling entries. CA prompts provide structure and encouragement which makes the process less daunting and challenging. Furthermore, prompts that react to user input can encourage elaboration and the consideration of alternative views (Kim et al., 2024; Park et al., 2021). It has been suggested that CA interaction transforms journaling from writing to telling a story which can make it more enjoyable (Park et al., 2021) as described in Section 3.3.3. How to design effective CA interactions that adapt the interaction to the user’s momentary needs and effectively engage them with different therapeutic activities is currently unclear.

4.4.4 Human-like Interaction

In attempts to create CA interfaces that deliver effective mental health treatment, research has explored emulating successful aspects of human delivery. As discussed in Sections 3.4.2 and 3.5.3, a research focus is therapeutic alliance (i.e., the working relationship between

patient and clinician) which is associated with positive therapeutic outcomes and has been demonstrated to arise from CA interactions (Beatty et al., 2022; Bickmore and Gruber, 2010; Bickmore et al., 2005b; Darcy et al., 2021). Methods of fostering therapeutic alliance include social and relational talk, for example, expressing empathy, humour, continuity behaviours and references to mutual knowledge between the patient and the system. Prior work has found these techniques to have positive effects both on the patient experience and health outcomes (Bickmore and Gruber, 2010). Yet, human-like interaction may not be the best way forward, at least for some patients. Research studying people’s experiences with voice assistants is finding that people question the appropriateness of human-like behaviour (Clark et al., 2019b; Cowan et al., 2017; Liao et al., 2016). Instead of human-like social behaviour, direct and utility-focused interaction may better suit some patients (Liao et al., 2016). Human-likeness may also contribute to mismatches between system capabilities and patient expectations whereby the system’s capabilities fall short of the user’s expectations leading to usability and user experience issues (Cowan et al., 2017; Luger and Sellen, 2016). Indeed, as discussed throughout Chapter 3, users of CA for mental health support are often frustrated by the limited conversational capabilities of these agents. What is more, human-likeness whereby CAs act in, for example, caring way may impact the users mental health (e.g., have comforting effects) which may interfere with assessment and intervention (see Section 2.2.4). It is unclear in what ways and in what situations it is appropriate for CAs to be designed to be human-like and social.

4.4.5 Patient Agency

Critical to both the patient’s interaction experience and wellbeing is the fulfillment of psychological needs including feelings of competence and autonomy (Yang and Aurisicchio, 2021). Indeed, a major benefit of MHSM activities like EMI is that they can help users to be and feel more in control of their own life (Matthews et al., 2017; Murnane et al., 2016). To support feelings of competence, we need to ensure the patient understands how to make use of the system’s capabilities, as well as what is beyond them; however, as previously discussed, how to align system capabilities and patient expectations is an open challenge. To support the patient’s autonomy, personalisation is a promising approach. Patients should be able to customise both when and how they interact with a system (Yang and Aurisicchio, 2021).

4.5 Clinician Involvement

A critical design decision for a Pervasive Therapy system is how clinicians will be involved. Clinicians, and healthcare professionals more broadly, can play impactful MHSM roles as described in Section 2.6.1. Furthermore, much of the delivery of evidence-based mental health treatment is through professional mental health services, primarily delivered by or with clinicians. Working with clinicians provides a means to leverage their skills, work practices, and judgement to address important issues concerning oversight, responsibility, and risk.

Clinician involvement with Pervasive Therapy systems can occur in two ways. The system may be used as a component of a ‘blended care’ treatment which combines in-person and digital treatment (Wentzel et al., 2016). Additionally, clinicians may interact with the system directly to contribute to the treatment delivery (Thieme et al., 2023). Involving clinicians in digitally-supported treatment can also help challenge perceptions and criticisms of digital treatments (e.g., distancing clinicians from patients) (Miner et al., 2019) as well as extrinsically motivate patients to engage with the treatment (Doherty et al., 2012). Higher intensity interventions for people in higher levels of distress cannot realistically be based on unsupervised, standalone technology, and if these technologies are to be used with a broader cohort of patients (beyond those with mild levels of distress), clinician involvement should be considered from the outset.

Miner et al. (2019) lay out four high level care delivery approaches. Two of the approaches are the extremes of ‘human only’ and ‘AI only’ delivery. While ‘AI only’ options should be treated with caution, as discussed above, it has been argued that AI has the potential to mitigate some challenges of accessibility and burnout found in human-only deliveries. Between these extremes is the opportunity for collaborative human-AI delivery. One collaborative approach is ‘human delivered, AI informed’ which would be limited by many of the access challenges of ‘human only’ care. A second collaborative approach is ‘AI delivered, human supervised’ care, which Miner et al. describe as “one of the less developed but more alluring ideas of AI psychotherapy”.

The Pervasive Therapy architecture outlined above exemplifies a largely AI delivered, human supervised approach, which can manifest in multiple ways. We identify and describe four (non-mutually-exclusive) forms of clinician involvement (Table 4.1) in a Pervasive Therapy system.

Table 4.1: Forms of clinician involvement in an AI-delivered, human-supervised Pervasive Therapy system.

Involvement	Description
Configuration	Clinicians tailor the Pervasive Therapy system for an individual patient at the start of the treatment.
Monitoring	Clinicians are able to observe the actions of the Pervasive Therapy system and data collected by it regarding their patient(s).
Adjustment	Clinicians can change the intervention delivery. This may involve re-configuring the Intervention or Conversation Controllers.
Communication	Clinicians and patients can exchange messages through the medium of the Pervasive Therapy system.

4.5.1 Configuration

The first point for clinician involvement is configuration, giving the clinician the opportunity to tailor the intervention to suit the patient and their individual mental health needs at the start of the treatment. The main methods for clinicians to tailor a Pervasive Therapy system are through Intervention and Conversation Controller configuration to describe what interventions to deliver and how to deliver them. This activity may extend beyond simply medical considerations to ones of technology engagement. The task of configuration is potentially well suited to being done collaboratively with the patient, particularly if the treatment is being used as part of a blended care approach. This interaction involving the patient, clinician, and system could contribute to establishing the involvement of the clinician in the system, the reputability of the system, and the patient's future engagement.

4.5.2 Monitoring

While patients are using a Pervasive Therapy system, clinicians could monitor patient progress and the actions of the system using the system's Clinician Interface. As part of a blended care approach, this activity can be used to inform in-person treatment. For patients using such a system, monitoring can be used to identify those who are having difficulty engaging with the treatment or who are indicated as being at potential risk of harm. In such cases where a patient is observed to be of higher risk, the clinician can take appropriate action, for example moving to a higher intensity of treatment within a stepped care model (Kruzan et al., 2023; Thieme et al., 2023).

Prior research by Matthews and Doherty (2011), that involved clinicians in the development of a digital MHSM application, identified the importance of setting boundaries with regard to patient monitoring. If there is to be monitoring of patients by clinicians, care must be taken to establish appropriate patient expectations to avoid generating the inappropriate expectation of clinicians continuously monitoring patient data.

Pervasive Therapy involves collecting substantial amounts of patient data including clinical measures, sensor data, and natural language data. The richness of this MHSM data collection could enable clinicians to develop a nuanced understanding of individual patients. Supporters of patients receiving internet-based CBT describe how additional information about what is going on in patients lives is important for contextualising and interpreting numerical monitoring scores (e.g., PHQ-9 and GAD-7) (Thieme et al., 2023). However, natural language MHSM data may be particularly verbose from a clinician perspective. An open challenge is to develop methods to supporting effective and efficient clinician review of MHSM data, potentially across a large number of patients (Cooney et al., 2024). Inference and prioritisation methods could be investigated as a means of highlighting to clinicians the patients most in need of their attention. However, if any method of prioritisation is to be considered, rigour is necessary to ensure safety. A possible system failure that could cause serious harm would be the system classifying a high risk patient as low risk, resulting in the patient not receiving appropriate care. For example, a naive summarisation of speech content may emphasise obvious symptoms (e.g.,

reports of tiredness) while neglecting subtle features (e.g., tone of voice or discontinuities in delivery) that could indicate the patient is at higher risk.

4.5.3 Adjustment

A level of clinician involvement, beyond monitoring patients, is adjustment. Adjustment allows clinicians to reconfigure the treatment (informed by review of the patients MHSM data) in an attempt to better match the changing needs of the patient. Adjustment could include re-configuring the Intervention and Conversation Controllers, and scheduling additional intervention components. For example, in cases of low engagement where the patient is not recording sufficient self-report entries, the system could be reconfigured to interact in a manner better suited to the patient (e.g., use more social language) or to deliver a greater proportion of prompts using a specific device that works well for the patient (e.g., via smartwatch).

A challenge when designing for adjustment is determining how transparent adjustments are to patients. Observable changes to the behaviour of the MHSM system without sufficient explanation risk disrupting and unsettling the patient, which could negatively impact their engagement and well-being. The therapeutic reasoning for adjustments could be explained to patients to help them understand why they have happened, however care will be needed to ensure explanations are accessible to all, including those with lower levels of mental health literacy.

4.5.4 Communication

Communication, whereby the system allows explicit exchanges between the patient and clinician, is a higher level of clinician involvement. Two key design challenges for incorporating communication are deciding how rigid and how synchronous the communication is.

A more rigid implementation would allow patients to send predefined communications with specific purposes, such as requesting feedback on their MHSM data or an intervention adjustment. Predefined messages could be sent by clinicians to encourage patients to record more regular or more detailed MHSM entries as well as to praise MHSM adherence. In contrast, the communication could be more flexible, allowing patients to express themselves verbally in a more free-form fashion. Clinicians could for instance message patients who are not making regular MHSM entries in order to develop understanding of potential barriers the patient is experiencing (Thieme et al., 2023). More rigid designs risk not allowing the patient to communicate valuable information while more flexible designs risk patients sending communications that the clinicians are ill equipped to respond to using the system.

How synchronous the communication is refers to how immediately patients can expect their communications to be seen and responded to. A synchronous design would have clinicians always available allowing for quick message exchange or even real-time conversation. Designs like this put a greater demand on clinicians and there is always the risk that clinician demand surpasses supply leaving patients that are expecting interaction, unat-

tended. Communication could be more asynchronous whereby patient communications are responded to, not immediately, but within a reasonably short time period. Asynchronous designs, which express to the patient a realistic timeline for receiving responses, appear a feasible approach, as previously demonstrated in work studying guided internet-based cognitive behavioural therapy (Doherty et al., 2012). Indeed, asynchronous communication is consistent with the use of MHSM as homework activity as part of therapies such as CBT whereby patient and clinicians interact at regular intervals (e.g., weekly).

Doherty et al. (2012) developed and studied an online mental health intervention platform that involved clinicians providing platform mediated feedback to patients. They found clinicians engaged more with the more engaged patients; clinicians also found it challenging to find ways to provide effective support to less engaged clients. As such, there may be a need to support clinicians to distribute their time between patients, as well to support each patient effectively. Through the analysis of clinician support messages to patients, effective support strategies can be identified which presents opportunity to design interfaces that support clinicians to support clients (Chikersal et al., 2020).

4.6 Risks and Ethical Concerns

A Pervasive Therapy system could present significant opportunities in bridging the mental healthcare gap, but logistical and ethical considerations need to be made for the main stakeholders in the systems: clinicians and patients.

Clinicians are often overloaded, with burnout a recognised issue; not only does this affect clinicians' own well-being, but it can also be tied to worse outcomes for patients (Miner et al., 2019). To safely and effectively adopt a Pervasive Therapy system, care must be taken to integrate the system into existing workflows, which includes sufficiently training clinicians. Fiske et al. (2019) describe a lack of framework around how clinicians should be trained for, and engage with, digital mental health interventions involving AI. Not only would clinicians need to become comfortable with the use of a technical system within their practice, but they would also likely need to be able to provide a level of support and guidance for new patients using the system. This guidance could relate to system navigation and expectation management (e.g., of the level of clinician involvement and AI capabilities), as well as recognising when a patient may be better suited to solely face-to-face therapy. It would be of benefit to all stakeholders to involve clinicians in the design and deployment of future Pervasive Therapy system to better understand the opportunities and limitations of integrating such a system into an established healthcare framework.

Patients' interaction with a Pervasive Therapy system may influence their future mental healthcare engagement. As has been noted with conversational agents, user expectations may exceed actual system capabilities with this gulf between expectations and realities leading to feelings of frustration and subsequent attrition (Cowan et al., 2017; Luger and Sellen, 2016). Of notable concern here also is the idea that patients who become disillusioned with the system could subsequently lose confidence in human practitioners or therapy as a whole (Miner et al., 2019). In contrast, if these systems are available all

the time and satisfy expectations, there is a risk that patients may over engage, becoming dependent on the systems (Fiske et al., 2019). This prospect could be mitigated by utilising clinician involvement, as with ensuring informed consent, to aid patients in understanding the capacity of the system and its shortfalls as well as highlighting when these shortfalls are reflective of the system, rather than the therapeutic methods. Additionally, patients and clinicians should be aware of the potential limitations of system insights, so as to encourage human autonomy in decisions relating to patient treatment (Thieme et al., 2020a). This further highlights the importance of training clinicians so they can effectively induct patients.

Pervasive Therapy systems will collect substantial amounts of sensitive data about each patient. For these systems to operate, each CA interface must be sufficiently informed about the patient, requiring some level of information sharing between devices. Our architecture illustration presents a centralised approach, whereby each device communicates with a central entity which collates the information from each device. We chose to illustrate a centralised approach because this reflects contemporary consumer technology which uses the ‘cloud’ as a central entity to share data between a patient’s devices. These cloud-device ecosystems are typically operated by commercial companies that people would need to trust to protect and not misuse their data (e.g., using the Patient Model to inform marketing) (Cowan et al., 2017). Alternatively, a decentralised approach could be taken whereby data is stored locally on devices and exchanged directly between them. This approach offers patients greater control of their data, but is a departure from contemporary consumer technology. For either approach, care would need to be taken to ensure patients fully understand how and why their data is being used, so they are informed enough to decide what they do or do not consent to (Thieme et al., 2020a). This understanding should include how the patient’s use of the system may be analysed to inform research and development, for example, if their data may be used for the training of machine learning models (Thieme et al., 2020a). It should also include the conditions under which their data may be shared with third parties (e.g., due to legal proceedings).

An open challenge of AI-involved therapies is determining who would have a responsibility to act in cases of patients disclosing information that could indicate that they are at risk of harming themselves or others. Laws are in place in certain parts of the world whereby clinicians must disclose if they feel patients are at risk, if clinicians fail to do so and patients act on these feelings, the clinicians could be held liable (Miner et al., 2019). Similar laws have not yet been established in the case of interventions involving AI, but would bring clarity for both patients and clinicians, ideally before they are needed. In the interim, Fiske et al. (2019) note that a solution to this concern in computer-guided therapies is to mandate continuous monitoring by a medical professional, although this is labour-intensive for clinicians and so may not be feasible in a Pervasive Therapy system that is always available. Automated approaches could be used, however as previously discussed in relation to patient monitoring, automated risk detection is not infallible. As such, there is a complex and unavoidable design challenge of balancing AI and human risk detection when designing systems for mental health support. New regulations relating

to AI are being introduced that will shape the use of AI within mental health care. For example, the EU AI Act mandates that the providers of high-risk AI systems establish risk management systems and design their systems to achieve appropriate levels of accuracy and robustness (Future of Life Institute, 2024).

4.7 Conclusion

Addressing RQ-A² of this thesis, in this chapter I explored the design of future CA systems for EMI, which I term Pervasive Therapy systems. Informed by my literature review (Chapters 2 and 3) I proposed a conceptual architecture (Figure 4.1) for such systems by integrating CA and EMI architectures. I used the conceptual architecture to facilitate articulation and consideration of the design challenges of Pervasive Therapy systems. Open challenges remain around how to create systems capable of planning and delivering interventions using CA interaction (Section 4.3), the design of CA interactions (Section 4.4), and the involvement of clinicians (Section 4.5). Through considering the design of CAs for MHSM at conceptual system level, I was able to consider the diversity of design challenges posed by CAs for MHSM. Addressing these different design challenges are directions for further research. Next, in Chapter 5, I further investigate the challenges of CA interaction design for MHSM.

²RQ-A: How can CAs be designed to support people with MHSM

Chapter 5

Theory-based Exploration of CA Interaction Design

5.1 Introduction

In the previous chapter, I considered the design of future conversational agent (CA) systems for mental health self-monitoring (MHSM). A conceptual architecture was first proposed, termed Pervasive Therapy, describing the general design of CA systems for ecological momentary intervention (EMI). This architecture was then used as a basis for describing open design challenges of such systems. One of the challenges was CA interaction design. We need to be able to design CA interactions that effectively engage people in order to design effective MHSM CA systems.

This chapter further investigates CA interaction design for MHSM. I approach this by using theories of human-human interaction (HHI) to help conceptualise and explore this design. As discussed in Section 3.9, HHI theories are a valuable basis for investigating the complex phenomena of human-CA interaction because of the resemblance of human-CA interaction to human-human conversation. This chapter explores the ways CA interactions could be designed for MHSM and thus contributes to RQ-A¹. As the approach of this chapter is to use HHI theories to inform the design explorations it also contributes to RQ-B². This chapter is structured as follows:

- 1. Joint Action:** First, in Section 5.2, joint action theory (Clark, 1996) is used to frame human-CA interactions as a collaborative activity. I use this framing to identify CAs interaction design strategies for proactively and collaboratively engaging users with MHSM.
- 2. Politeness:** Second, in Section 5.3, I use politeness theory (Brown and Levinson, 1987). Building on the joint action theory framing of human-CA interactions, I argue that the use of politeness may be important for engaging people with MHSM because MHSM can be both sensitive and inconvenient for users. I explore how politeness theory can inform the design of CA interactions for MHSM

¹RQ-A: How can CAs be designed to support people with MHSM?

²RQ-B: How can theories of human-human interaction inform the design of CAs for MHSM?

3. Discussion and Future Work: In Section 5.4 the implications of this theory-driven design exploration for CA interaction design are discussed.

I take the ‘polite’ CA interaction designs created in Section 5.3 forward into Chapter 6 where I present a mixed-methods user study involving them.

Credit Statement

Section 5.2 is based upon content from a paper titled “Beyond Subservience: Using Joint Commitment to Enable Proactive CUIs for Mood Logging” published in the 2022 Conference on Conversational User Interfaces (CUI) (Bowman et al., 2022b). The content is my work informed by feedback from my supervisory team (Gavin Doherty, Benjamin Cowan, and Anja Thieme).

Section 5.3 is based upon content from a paper titled “Exploring How Politeness Impacts the User Experience of Chatbots for Mental Health Support” published in the International Journal of Human-Computer Studies (Bowman et al., 2024). My supervisory team, along with Joseph Newbold and Leigh Clark, identified the relevance of politeness theory (see last paragraph of Section 5.3.2). I, along with Orla Cooney, then developed this idea ultimately running a user study which I report in Chapter 6. The conceptualising team started exploring the application of politeness theory to CA interaction design creating polite CA prompts; I and Orla Cooney developed these prompts producing those presented in Section 5.3.5. This chapter presents considerable review of related work on CA interaction design (Sections 5.3.1, 5.3.3, 5.3.4) to situate and develop this exploration. This review work is my own informed by feedback from my co-authors.

Section 5.4 is based on content from both of the previously referred to papers (Bowman et al., 2022b, 2024), predominantly the paper on joint action (Bowman et al., 2022b). The work in this section is my own with feedback from my collaborators.

5.2 Joint Action

5.2.1 Designing to Proactively Engage Users

Designing digital tools to prompt and remind users to do MHSM is a major opportunity of digital tools, as discussed in Section 2.2.5. Indeed, as described in Section 4.4.1, such *proactive* interaction is key to realising advanced forms of MHSM, like Pervasive Therapy, that aim to engage users ecologically and momentarily in assessment and intervention. Tools could, for instance, be designed to prompt users to make entries at times where users are free to do so or after mood impacting activities such as exercise. Proactive interaction goes beyond initiating interactions, to guiding and encouraging the user to do things (e.g., therapeutic activities) during the interaction (Section 4.4.3). A design challenge with CAs for MHSM is that the prevailing CA interaction paradigm is one of subservience and user initiation whereby CAs are designed to react to and execute the requests of users (Akbar and Conlan, 2024; Doyle et al., 2019; Kraljic and Lahav, 2024).

Such a lack of proactive behaviour by CAs limits their use to what the user decides to initiate.

For proactive CA interactions to be effective they need to be situationally appropriate whereby the interaction characteristics (e.g., task duration, sensitivity of the information being requested, or the use of social and empathetic language) align with the user’s situation (Cha et al., 2020). Inappropriate proactive interaction risks annoying, disrupting, and even embarrassing the user (Clark et al., 2019b; Cowan et al., 2017). How to design CAs to act and engage in such a situationally appropriate fashion is still an ongoing challenge.

New CA interaction design strategies are needed to support designers so as to ensure that a proactive CA engages in a situationally appropriate manner that is sensitive to the user’s momentary state, preferences, interaction goals, and context. The work presented aims to argue for, and propose, a framework for such proactive interaction design based on the theory of language use as joint action (Clark, 1996). It describes language use as a collaborative activity, giving consideration to the goals and momentary situations of its participants. Joint action theory also incorporates the notion of *joint commitment* which describes how a conversation’s participants initiate and engage in dialogue. I next consider how joint action theory, in particular joint commitment, can be used to inform proactive CA interaction design for MHSM.

5.2.2 Joint Action Theory

Clark (1996)’s theory, from human-human interaction studies, proposes that language use in dialogue is a form of *joint action*, involving an ensemble of participants acting in coordination. Its relevance to human-computer interaction more generally has been long established (Brennan, 1998; Brennan and Hulteen, 1995; Cho and Rader, 2020; Cuadra et al., 2021; Homaeian et al., 2021; Monk, 2003). Recent advances in conversational AI (i.e., large language models/ generative AI) have renewed interest in joint action theory (Kraljic and Lahav, 2024). Inline with this work, Kraljic and Lahav (2024) recently argued that joint action theory could inform the design of more collaborative human-CA interactions that would enable us to better leverage the capabilities of AI.

Joint action theory (Clark, 1996) proposes that partners coordinate in an attempt to achieve a variety of goals at the same time. These include domain goals, that is the dominant purpose of the activity (e.g., MHSM), as well as procedural (e.g., completing a MHSM entry quickly quickly), and interpersonal goals (e.g., building rapport). Partner goals can differ and can also vary across interactions, they can also be made public or kept private during dialogue. In the case of MHSM with a CA, users may sometimes want to make an entry quickly, while other times they may want to receive empathetic responses, but users may not directly express this to the CA. Partner goals may also be in conflict. For instance, the user’s priority may be a quick interaction, while the CA’s goal is to elicit a detailed entry about an experience, which may take some time. Critically, for dialogue partners to succeed they must find a way to coordinate their individual actions that sufficiently achieves their collective goals (i.e., mutually agree to a joint purpose).

A process by which participants coordinate their actions around their joint purpose

Table 5.1: Key Definitions for Joint Action and Joint Commitment

Term	Meaning
Joint Action	An action involving an ensemble of participants acting in coordination (e.g., two people talking, a client and CA interacting).
Joint Purpose	A purpose a joint action could serve (e.g., a specific MHSM entry).
Joint Project	A joint action that serves a joint purpose.
Joint Commitment	Where participants commit to enacting a joint project.

is termed a *joint project* (Clark, 1996). Joint projects are joint actions that serve a joint purpose which the partners have *jointly committed* to (refer to Table 5.1 for definitions). The process works by a partner (e.g., the CA) proposing a joint project and the others (e.g., the user) choosing whether to take it up.

Establishing joint commitment is not always straightforward. Clark (1996, p. 204) describes four ways the receivers of a proposed joint project may respond:

1. Comply
2. Alter
3. Decline
4. Withdraw

The proposer's preference for how the receiver responds is in descending order from complying to withdrawing (Clark, 1996). A *comply* response is where the receiver commits to the project completing it as proposed. An *alter* response is where the receiver proposes back an altered version of the joint project that they are prepared to commit to. If the receiver engages in a *decline* response to the proposed joint project (i.e., by responding that they will not commit to it), there is then an opportunity for the original proposer to propose an alteration of the joint project. A *withdrawal* response is when the receiver does not demonstrate consideration of the proposed joint project.

Clark (1996, p. 203) also describes four requirements that the partners must fulfil when declaring joint commitment to a joint purpose:

For A and B to commit themselves to joint purpose r

1. *Identification* A and B must identify r
2. *Ability* It must be possible for A and B to do their parts in fulfilling r
3. *Willingness* A and B must be willing to do their parts in fulfilling r
4. *Mutual Belief* A and B must each believe that 1, 2, 3, and 4 are part of their common ground.

As such, partners attempting to initiate joint projects based upon their beliefs about the other participant’s momentary willingness and ability is fundamental to joint commitment.

5.2.3 Designing for Joint Commitment

Terminology and concepts related to joint action and joint commitment echo processes that proactive CAs need to engage in for successful interaction. As such, we propose that they may provide us with a framework to characterise and inform proactive CA design. For instance, the process of a partner proposing a joint project, hoping that the receiver will commit to enacting it with them, is similar to the dialogue process needed when a proactive CA initiates an interaction with a user. Each partner needs to consider their beliefs about the other’s ability and willingness to commit to a proposed joint project (i.e., the task the CA is engaging the user to do). To be proactive, a CA will need to propose joint projects to users. For their proposals to be successful, the projects should often result in joint commitment. Below we explore this in more detail, illustrating three areas in which the concept of joint commitment in particular can inform the design of proactive CA interactions for MHSM:

1. choosing a joint project to propose,
2. handling uncertainty through pre-sequences, and
3. supporting alteration responses.

Choosing a Joint Project to Propose

For MHSM, each time a proactive CA considers proposing a joint project it needs to choose from a set that is potentially large and diverse. These joint projects could vary in their topic (e.g., a past, present, or future activity), the depth of the report (e.g., a categorical rating item or a series of unconstrained language inputs), and the use of social or empathetic language (e.g., ‘How do you feel?’ or ‘I would love to know how you’re feeling?’). For the proposal by the CA to be successful, whereby it results in joint commitment, the characteristics of the proposed project need to align with the user and their context.

Table 5.2: Example Project Proposals

	CA Proposal	User Response Modality
A	How are you feeling?	Fixed options (e.g., numeric or emoji scale)
B	Hey, I’d love to know how your gym session went?	Free Text

Table 5.2 presents two example project proposals, A and B, which demonstrate the CA attempting to get the user to provide a scale-based self-report, and an unconstrained

account of an experience respectively. While project A offers a quick interaction by allowing the user to pick from fixed response options, project B asks more of the user as the input is more work, may require the user to reflect more, and will take longer. Given the increased ask of the user to comply with B, it may be the case that the user has the *ability* to comply with A, but not B. However, project B considers a certain topic (a gym session) which could be pertinent to the user at this moment. This contextual relevance may result in the user being more *willing* to commit to project B as 1) it may resonate with user's momentary cognition and 2) its value to the user may be clearer which is understood to affect user receptiveness to interruption (Liao et al., 2016). Furthermore, with B the CA uses social language by expressing an interest in the user. Depending upon factors including the user's orientation towards social-agents (Liao et al., 2016), as well as momentary factors like the user's mood, this use of social language may influence the user's *willingness* to commit to B.

Our comparison of proposals A and B shows how the characteristics of different projects may affect the user's likelihood of committing. It demonstrates how the structure of joint commitment supports us with relating beliefs about the user and their context to the appropriateness of different projects, through consideration of the user's ability and willingness. To design for joint commitment, CA interaction design strategies would need to enable designers to balance how likely the user is to commit to projects, given the CA's beliefs about their ability and willingness, with how well the projects would achieve the interaction goals.

Handling Uncertainty Through Pre-sequences

Central to designing for joint commitment is the CA's belief about the user's ability and willingness which is intrinsically uncertain. Discrepancies between the CA's belief about, and the actual situation of, the user could lead to the CA proposing joint projects that the user is not prepared to commit to. Considering again the example proposals presented by Table 5.2 with the expectation that B will take longer to enact, how long the CA believes the user to have available for an interaction could determine its choice of project A or B. For example, if the CA believes that project B better aligns with its and the user's goals, the user is in an appropriate context, and that the likelihood of the user having enough time to enact the project is high, then project B seems a good choice for the CA to propose. However, if the likelihood of the user having enough time is lower, then the risk of the user declining or withdrawing from the proposal of B becomes higher.

To support CAs with choosing which joint project to propose, *pre-sequences* could be used. Pre-sequences are used by partners who want to propose a joint project, but anticipate that the receiver may not be prepared to commit to it (Clark, 1996). The purpose of a pre-sequence is to identify if the conditions for a joint project are satisfied prior to its proposal in order to avoid partner's altering, declining, or withdrawing the joint project. They are an example of grounding whereby the knowledge gained through the pre-sequence is added to the participants' common ground, thus increasing the certainty of partners' beliefs about the others (Clark, 1996).

Table 5.3 presents two example pre-sequences. Example 1 explicitly targets the user’s

Table 5.3: Example Pre-sequences

Pre-sequence	
1	Do you have 5 minutes?
2	Looks like you’re on the bus home, is that right?

ability by attempting to ground if the user has 5 minutes they could use to enact a joint project. The second example takes a more implicit approach by attempting to ground the user’s activity. By grounding the user’s activity, the CA could make multiple inferences pertaining to the user’s willingness and ability; in the case of grounding that the user is on the bus home, the CA could infer both factors affecting the user’s ability (e.g., the time the user has available) and willingness (e.g., how private the setting is).

To make effective use of pre-sequences, a way of deciding if it is best to 1) propose a project directly, 2) use pre-sequences to first assess if the user can commit to the project, or 3) propose a different project, is needed. There is also the challenge of deciding which pre-sequence to use. A principle that guides the use of pre-sequences is that of greatest obstacle whereby “all else being equal, two people trying to establish a joint purpose will try to first overcome the greatest, or most likely, obstacle to reaching it” (Clark, 1996, p. 308). Accordingly, the potential barriers to joint commitment need to be considered so as to design effective pre-sequences. Inaccurate inferences about the user’s context can negatively affect user engagement (Bickmore et al., 2009) warranting a cautious approach that makes greater use of pre-sequences. We note that what is deemed appropriate in terms of pre-sequence use may vary between users, especially in relation to user’s social-agent orientation (Liao et al., 2016).

Supporting Alteration Responses

The previous subsections have focused on how to choose and propose a joint project that the user is likely to comply with. Sometimes these efforts will be unsuccessful. In these situations a proactive CA should allow users to alter the proposed project as this is preferable to them declining or withdrawing.

To inform the design of responses to support alteration, as with pre-sequences, we can consider the potential obstacles to the user complying. If we can identify the highest risk obstacles, we can then attempt to provide the user with ways to alter the proposed project around these obstacles. Table 5.4 presents an example project proposal alongside example fixed response options for alteration.

The first example alteration stems from considering obstacles relating to the user’s ability. It offers a way for the user to alter the project from one they do not have the ability to commit to, due to its duration, to one they are able to commit to. The second example alteration considers more the user’s willingness due to contextual factors (i.e., if

Table 5.4: Example Alterations

CA Proposal	Example User Responses (Fixed Options)
Hi, looks like you're on the bus home, do you have a few minutes to chat about your gym session?	1) No, but I do have time for a quick mood rating. 2) I'd rather not on the bus, ask me when I get home.

they are prepared to enact the joint project in their current setting) proposing an alteration they may be more willing to engage in.

To effectively support alteration a way of deciding when to provide the user with alteration responses is needed. Offering alterations may reduce the risk of joint projects being declined. Yet it may also result in less preferable joint projects where, had alterations not been suggested, the user would have committed to the preferred project in any case. Providing alteration responses may be most appropriate in situations when the CA has a less specific goal. The presentation of potential alterations by the CA can also demonstrate its functionality and capabilities as well as allow users to reflect and inform the CA of what activities they may have time for.

5.3 Politeness

5.3.1 Encouraging Willingness Using Politeness

The consideration of joint commitment within Section 5.2 emphasises that for there to be MHSM interactions the user need to be willing to participate. Unfortunately, people may often not be very willing to do MHSM interactions. MHSM interactions are sensitive and can be upsetting; they also require time and effort and therefore can be inconvenient.

People use linguistic politeness strategies during human-human conversation to encourage other people to do things for them (Brown and Levinson, 1987; Clark, 1996). CAs could be designed to use linguistic politeness strategies. If these strategies of human-human conversation function similarly in human-CA interaction, their use by CAs could encourage people to be willing to engage in MHSM interactions. Research on CAs has repeatedly identified that politeness is important for CA interaction design (e.g., Chaves and Gerosa, 2021; Jain et al., 2018; Rapp et al., 2023; Yang and Aurisicchio, 2021), yet it remains unclear how to effectively leverage politeness within design. Research has recommended that we study how the politeness strategies of human-human interaction can be used within human-chatbot interaction design (Chaves and Gerosa, 2021; Rapp et al., 2023). We next take this approach and investigate how a prominent theory of politeness in human-human interaction (Brown and Levinson, 1987) can be used to support the design of CA interactions for MHSM. We:

1. provide an overview of politeness theory (Section 5.3.2),
2. review work recommending polite CA interaction design (Section 5.3.3),

3. review work applying politeness theory to CA interaction design (Section 5.3.4), and
4. explore the interaction design of polite CAs for MHSM (Section 5.3.5)

5.3.2 Politeness Theory

Politeness theory (Brown and Levinson, 1987) builds on Goffman (1967)'s theory of face whereby in social interaction each participant has an expressed self-image which has a social value termed *face*. During interaction participant's behaviours can affect each other's face. Two principles that structure social interaction are:

Self-respect people want their expressed self-image (face) to be maintained,

Considerateness people do not want other's expressed self-image (face) to be undermined.

Accordingly, people in social interaction perform behaviours in an attempt to maintain their own, and other participants', expressed face; such behaviours are the enactment of politeness (Goffman, 1967). Politeness theory considers face as composed of two parts:

Positive Face a person's self-worth and want to be seen as desirable by others,

Negative Face a person's autonomy and want to not be impeded by others.

The theory perceives many social actions as intrinsically threatening to participants' faces (i.e., these actions could have detrimental impacts upon a participant's face); it terms such actions as face-threatening acts (FTAs).

Toward maintaining face, politeness theory describes strategies for attempting to mitigate the face-threatening effects of social actions. The theory presents 5 high-level approaches to mitigating the effects of an FTA which, in order of decreasing strength, are:

1. Not performing the act,
2. Performing the act *off-record*,
3. Performing the act with the redressive action of *negative politeness*,
4. Performing the act with the redressive action of *positive politeness*,
5. Performing the act *bald-on-record*

When an act is performed off-record its meaning is ambiguous such that it is unclear if a threat to face exists. The redressive approaches of negative and positive politeness attempt to counteract face-threatening effects by signaling that a threat is not intended or desired. Positive and negative politeness attempt this by tending, respectively, to the positive and negative face of the persons threatened. Both the positive and negative politeness approaches are composed of a set of linguistic strategies. The approach of performing the act bald-on-record is where the act is performed in a direct, clear, and concise manner without redress.

The challenges of engaging users with CA interactions for mental health support can be interpreted using this theory of politeness. The request for interaction with a CA for mental health support could be viewed as impeding the user and lowering their feelings of autonomy, and therefore this request could be viewed as a negative FTA. As interactions regard the user’s mental health, which is a sensitive and stigmatised topic, they could be viewed as impacting the user’s feelings of self-worth, and therefore these interactions can also be viewed as positive FTAs. It follows from this interpretation that CAs for mental health support could be designed to mitigate these FTAs using the approaches, such as positive and negative politeness, described by politeness theory. This FTA mitigation could improve the user experience of activities like MHSM and, consequently, promote engagement.

5.3.3 Recommendation of Politeness for CA Interaction Design

The importance of politeness for human-CA interaction has been stressed by recent work. A 2021 review on social characteristics in human-CA interaction by Chaves and Gerosa (2021) identified the characteristic of *manners*. This work built on that of Morrissey and Kirakowski (2013) who identified manners as a characteristic for evaluating the naturalness of CA interactions. Chaves and Gerosa draw on the definition of manners provided by Morrissey and Kirakowski of “the ability of chatbots to display *polite* behaviour and conversational habits” (Morrissey and Kirakowski, 2013, p. 93 emphasis added). Further in line with Morrissey and Kirakowski, Chaves and Gerosa report that the main benefit of providing manners is to increase human-likeness. Similarly the review notes that, from an analysis of first-time user’s CA interactions, Jain et al. (2018) found that users expect human-like conversational etiquette from CAs. Referring to Brown and Levinson’s politeness theory (Brown and Levinson, 1987), Chaves and Gerosa describe that a challenge of manners, or politeness, in human-CA interaction is the complexity of identifying when there is a FTA that the CA should attempt to mitigate. They conclude that:

“adoption of politeness strategies to deal with face-threatening acts is still under-investigated in the chatbot literature” (Chaves and Gerosa, 2021, p.742)

Research on the design of collaborative human-agent interactions has led to recommendations for the design of polite interactions. Echoing my work within Section 5.2, these works have built on ideas of joint activity. Cila (2022) reviewed literature to identify the qualities of human-agent collaborations and to elicit design considerations; they identified the use of politeness theory as a promising design opportunity. Rapp et al. (2023) conducted a qualitative analysis of over 1000 conversations with a customer service CA to understand how to design for more collaborative human-CA interaction. Their analysis demonstrates that people behave in polite ways by being patient with, and by not behaving in negative ways (e.g., aggressively) toward, the CA. The authors recommend designing CAs that nurture a friendly and polite conversational environment. Rapp et al. (2021) conclude, explicitly echoing the conclusions of Chaves and Gerosa (2021), that:

“there is a lack of studies on how to leverage the politeness and friendliness

strategies used in human-human social interactions...such design strategies could contribute to making a more comfortable conversation environment and preventing feelings that may cause users to abandon the interaction” (Rapp et al., 2023).

Of particular relevance to our research on the design of CAs for mental health support is a study by Yang and Aurisicchio (2021) that investigates a self-determination theory approach to the design of CAs. Self-determination theory (SDT) is a theory of human motivation that argues that the fulfilment of people’s psychological needs - specifically feelings of autonomy, competence, and relatedness - impacts their motivation and well-being. SDT’s consideration of people’s feelings of autonomy resembles politeness theory’s consideration of negative face (i.e., feelings of autonomy), while SDT’s considerations of competence and relatedness overlap with politeness theory’s consideration of positive face (i.e., feelings of self-worth). Congruent with the shared aims of self-determination and politeness theory, this investigation into a SDT approach for the design of conversation agents produced the guideline:

“Talk Politely: Encourage polite and socially appropriate conversation style.” (Yang and Aurisicchio, 2021)

This design guidance is however vague as it provides little detail for how to use politeness within dialogue design.

In sum, research on human-CA interaction is identifying that polite behaviour by CAs meaningfully impacts the user experience, and subsequently recommends that we design for this behaviour. Yet, it remains unclear how to effectively design polite CA behaviour. Motivated by this, prior research has called for studies of how the politeness approaches of human-human interaction can be used within CA design (Chaves and Gerosa, 2021; Rapp et al., 2023).

5.3.4 Applications of Politeness Theory to CA Interaction Design

Studies have investigated the use of approaches based on politeness theory (Brown and Levinson, 1987) by computers in human-computer interaction. Here we present an overview of relevant findings. Research suggests that politeness can improve user perceptions of computers as interaction partners. Wang et al. (2008) compared a polite (mixture of off-record, positive and negative politeness approaches) with a direct virtual pedagogical robot designed to support students learning to use an engineering simulation software. The analysis shows that the polite system was preferred by students and had better learning outcomes. Torrey et al. (2013) investigated the use of hedges which are a form of negative politeness. Hedges introduce uncertainty into requests so that it is not assumed that the receiver can fulfill them, thus reducing the force of the request and thereby reducing the FTA. An experiment was conducted that involved comparing people’s perceptions of a speech-based domestic robot helper using, and not using, hedges for the task of supporting a novice human to make cupcakes. Robots using hedges were perceived as more considerate and were liked more.

In a study with older adults, Hu et al. (2022) compared a polite (mixture of positive and negative politeness) with a direct (bald-on-record) voice- and touch-interaction display for health support. Quantitatively they observed negligible difference between the polite and direct systems. However, their qualitative results suggest most participants preferred the polite version, others preferred the direct version, and some participants were indifferent. This analysis suggests that perceptions of polite behaviour by computers can differ between users.

User perceptions also differ between the politeness approaches used by a computer as found by the work of Miyamoto et al. (2021). This work compared user perceptions of driving support robots using positive and negative politeness approaches. It was observed that most (two thirds of) participants rated the robot using positive politeness favourably, others (one sixth of) rated the robot using negative politeness favourably, and the rest (one sixth of) rated the robots equally. The study also found that the robot using positive politeness was perceived as significantly more anthropomorphic than the robot using negative politeness.

As with user perceptions of anthropomorphic behaviour (Liao et al., 2016), user perceptions of different politeness strategies may be impacted by the user's orientation toward social behavior by computers. This is suggested by the work of Lee et al. (2011) that compared perceptions of a malfunctioning robotic vacuum cleaner using either positive, negative, or no politeness. They observed that the perceptions of users with a more social orientation were more sensitive to the use of politeness by the robot than those of users with a more utilitarian orientation.

Lee et al. (2019) studied perceptions of a polite (mixture of positive and negative politeness) and plain (stated instructions directly) virtual driving assistant in simulated normal and failure driving situations. Their results indicate that politeness improves the user experience in normal situations, but impedes it during failure situations. Similarly, Salem et al. (2013) found interaction context (a social or goal-oriented interaction) had a greater impact on user's perceptions of a robot receptionist than the robot's use of politeness (direct (bald-on-record) or positive politeness). These findings indicate that perceptions of the appropriateness of politeness can be affected by the interaction context.

Research has also considered the potential effects of politeness such as persuasiveness. Hammer et al. (2016) compared ratings of persuasiveness and politeness for different utterances finding that politeness and persuasiveness are perceived as different constructs. Srinivasan and Takayama (2016) studied how the use of politeness by help-seeking robots would affect how people assisted the robots when requested. They compared direct (bald-on-record), positive politeness, negative politeness, and in-direct (off-record) help requests. The results indicated positive politeness was the most effective. With statistical significance, participants were more willing to help the robot using positive politeness than that using direct language and perceived positive politeness as more appropriate than both direct and indirect language. However, the results did not show statistically significant differences between positive and negative politeness. In a virtual reality setting, Zojaji et al. (2020) studied how effective the use of different verbal and non-verbal politeness

strategies by a virtual agent were at persuading participants to join a small group conversation. The verbal strategies included direct (bald-on-record), positive politeness, negative politeness, and in-direct (off-record). They found the direct and redressive strategies were similarly persuasive, however the direct strategies were rated less favourable on measurements relating to face loss. The results suggest that positive politeness best balances persuasiveness and considerateness.

Although work has compared the different approaches described by politeness theory, using these approaches in isolation is not straightforward. Terada et al. (2021) attempted to validate scripts using positive politeness, negative politeness, and off-record approaches; they found the script using negative politeness was not perceived as using negative politeness to a higher degree than other politeness forms. Similarly, when Miyamoto et al. (2021) had experts evaluate utterances designed to use either positive or negative politeness; some utterances were classified by the experts as using both forms of politeness. Given the difficulty of isolating different politeness approaches, research should proceed cautiously.

Research has also studied the use of politeness theory less centrally. Relational agents research, as described in Section 3.4, used politeness theory to inform the design of social talk strategies which were identified as a way to develop the user’s trust in the system. This research was motivated by the understanding that trust is a prerequisite for risky interactions, such as self-disclosure. It found that within extroverted users the social talk strategies did increase user-computer trust and bond (Bickmore and Cassell, 2001).

To summarise, these studies we have described demonstrate that the use of politeness approaches by computers can impact the user experience. The impacts of politeness appear to be affected by the type of politeness used, user’s interaction preferences, and characteristics of the interaction task. These studies have largely investigated robots (Hammer et al., 2016; Lee et al., 2011; Miyamoto et al., 2021; Salem et al., 2013; Srinivasan and Takayama, 2016; Torrey et al., 2013) and embodied virtual agents (Bickmore and Cassell, 2001; Wang et al., 2008; Zojaji et al., 2020). A couple of studies (Hu et al., 2022; Lee et al., 2019) have considered voice assistants. How the use of politeness by text-based CAs is perceived has not yet been a research focus.

5.3.5 Designing Polite CAs for MHSM

To explore how politeness theory (Brown and Levinson, 1987) can inform the design of CAs for MHSM we created three CA prompt sequences that vary in their use of politeness. To design the CAs we first identified a set of MHSM queries (Table 5.5), based on the MHSM tool of an established online mental health platform³. These queries, in conjunction with log initiation and completion, gave the basic structure of the MHSM CA interactions. We modified language, using different politeness strategies, to create the prompt sequences which are presented by Table 5.6. Consistent with prior work (Section 5.3.4), we decided to design prompt sequences inspired by the politeness strategies of bald-on-record, negative politeness, and positive politeness. This design work was a collaborative activity involving

³<https://www.silvercloudhealth.com>

the research team. We considered the different ways politeness strategies could be used to transform the base queries. Prompt sequences were created that were then discussed and iterated upon. This process was informed by our review of related research as well as our background in human-computer interaction, CAs, and digital mental health.

Query	Question
Mood rating	Rate your mood from 1-5
Mood Situation	Did a situation make you feel this way?
Related Factor	How was your sleep last night?
Diary Entry	What have you done today?

Table 5.5: Base MHSM Queries

The first prompt sequence was termed **Direct**. Similarly to prior work, this sequence was designed using politeness theory’s bald-on-record strategy; it uses direct language and was included to serve as a non-polite comparison. These prompts are largely the same as the base queries.

The second and third prompt sequences, termed **Passive** and **Personal**, were designed to be polite. The Passive sequence was designed using politeness theory’s redressive action approach of negative politeness. The Personal sequence was designed using politeness theory’s redressive action approach of positive politeness. We chose to use the terms Passive and Personal, rather than Negative and Positive, because prior work (Miyamoto et al., 2021; Terada et al., 2021) has shown that it is difficult to use positive and negative politeness in isolation (Section 5.3.4). By not using the terms Negative and Positive we want to acknowledge that people may not perceive the Passive sequence as exclusively using negative politeness and the Personal as exclusively using positive politeness. We choose the terms Personal and Passive to emphasise that the Personal sequence makes use of personal pronouns (e.g., ‘**I** would love to’, ‘**We** could’, ‘Would help **me**’), which are innately anthropomorphic, and the Passive does not. Given the apparent importance of human-likeness in the mental health domain (Section 4.4.4), this characteristic could be impactful.

Negative and positive politeness strategies could be used to transform the base queries into many different ‘negative’ politeness and ‘positive’ politeness prompt sequences. The sequences we created are particular implementations. When creating the Passive and Personal prompt sequences we aimed to capture the general approaches of negative and positive politeness by using a range of sub-strategies rather than focusing on the use of a single sub-strategy (e.g., hedging). The Passive sequences uses, for example, the negative politeness strategies of apologising (“sorry to bother you”), questioning/hedging (“would you mind”), and minimising the imposition (“could you write a little”). The positive politeness sub-strategies used by the Personal sequence include expressing interest in the user (“I’m curious to know”), expressing optimism (“I hope”), and including the speaker

in the activity (“Let’s get logging”). When choosing sub-strategies, we also considered their appropriateness for the application of MHSM. For example, we felt that the negative politeness strategy of being pessimistic and the positive politeness strategy of joking lacked appropriateness for sensitive and serious interactions like those for MHSM.

5.4 Discussion

5.4.1 Joint Action as Framework for CA Interaction Design

As argued in Chapters 3 and 4, specifically Section 4.4, in order to design effective CA systems for MHSM we need to be able to design effective MHSM CA interactions. A major challenge to design of effective CAs interactions the subservient, user-led paradigm of CAs interaction. I argued at the start of this chapter that to realise the potential of CAs for MHSM we need to design them to engage with users more proactively and collaboratively. I explored in this chapter how joint action theory (Clark, 1996), in particular the concept of joint commitment, could be used as a framework to conceptualise and design proactive and collaborative CA interactions. Since the work on joint action I report in this chapter and in response to the emergence of large-language models, Kraljic and Lahav (2024) have similarly argued that joint action theory should be used to approach the design of collaborative human-CA interactions and human-AI interactions more broadly.

Our exploration suggests that designing for joint commitment requires designing interactions in account of the uncertainty about the situation of the user (Section 5.2.3). As prior work (Cha et al., 2020; Ho and Intille, 2005) has established, the user’s momentary state and context affects the suitability of proactive interactions; the structure of joint commitment provides a way to incorporate the CA’s beliefs about these factors into CA interaction design. As the CA’s beliefs, constructed using context-aware techniques (Burns et al., 2011; Wei et al., 2022), are intrinsically uncertain, this uncertainty should be represented by the CAs designed. Similar to existing tools, this uncertainty could be included in design by using flow charts allowing designers to construct belief-conditioned ‘flows’. To support designers working with beliefs, care will need to be taken with belief representation potentially involving designers in its configuration, especially as requirements may differ between applications. Tools could offer design-support for the use of pre-sequences and alteration responses by seeking to aid designers with identifying and accounting for potential obstacles and user barriers to joint commitment. Future work should seek to explore how tools and approaches can be developed to support designers in this process.

Our explorations also suggest that designing for joint commitment involves designing to encourage users to engage in interactions beyond and different to those they would initiate. For instance, during user-initiated interactions, the CA could seek to alter the joint projects proposed by the user toward projects that better align with mutual interaction goals (e.g., MHSM interactions that likely to be more beneficial for the user.). In health settings, it may even be appropriate for the CA to decline the user’s proposals (e.g., to not engage in harmful self-report in attempts to maintain unhealthy behaviours (Kelley et al., 2017)). Yet, along with opportunity, altering user proposals carries the risk of damaging

Table 5.6: MHSM CA Prompt Sequences Informed by Politeness

Prompt	Prompt Sequence		
	Direct	Passive	Personal
Log Initiation & Mood Rating	Rate your mood on a scale of 1-5.	Sorry to bother you, but would you mind logging your mood on a scale of 1-5 please?	<i>Message 1:</i> Great to see you! Let's get logging. <i>Message 2:</i> I would love to know your mood on a scale of 1 - 5 please?
Mood Situation	Did a situation make you feel this way? If so, describe the situation.	<i>Message 1:</i> Logging your mood can be hard so well done for doing it! <i>Message 2:</i> If a situation made you feel this way, it would be useful to write about it.	<i>Message 1:</i> Thank you for sharing that with me. I appreciate that MHSM can be hard. <i>Message 2:</i> It would be helpful for me to know about a specific situation, if there is one, that made you feel this way.
Diary Entry	What have you done today?	Could you write about what you've done today?	You'll tell me about what you've done today, I hope?
Related Factor	How was your sleep last night?	One last thing, could you write a little about how you slept last night?	I'm curious to know how you slept last night.
Log Completion	Log entry complete.	Log entry complete. That wasn't too hard was it. You should log a quick entry again soon.	Thank you. We have finished the log, I hope to speak to you again soon.

the user experience causing attrition. Alteration attempts should look to be sensitive to the user's situation and, when appropriate, afford the user the ability to decline the alterations such that the CA then commits to the user's original or altered proposal. As discussed in Section 5.3, politeness may play an important role in encouraging user to engage in MHSM interactions.

The application of MHSM may further complicate the use of politeness by a CA. As discussed by Lakoff (1989), therapeutic discourse is not an ordinary conversation setting. There are expectations for therapeutic conversations of the clients disclosing to the therapist, the therapists asking questions of the client, and therapist offering interpretation of the clients disclosures. These expectations impact what uses of politeness people view as appropriate within therapeutic discourse. The may also impact what uses of politeness people view as appropriate for human-CA interaction for mental health support.

5.4.2 From Human-Human to Human-CA Interaction

Joint action theory (Clark, 1996) and politeness theory (Brown and Levinson, 1987) are theories of human-human, not human-CA, interaction. As discussed in Section 3.9, theories of HHI can be a useful starting point for investigating human-CA interaction because of its resemblances to human conversation. As we develop greater understanding of human-CA interaction we made need to modify these theories for it or to develop new theories. Furthermore, given people perceive CA interaction differently to human interaction (Clark et al., 2019a; Doyle et al., 2019; Porcheron et al., 2018) it may not be appropriate to design CA interaction to overly resemble human conversation.

How politeness maps from HHI to human-CA interactions may be particularly complex given that politeness is particularly rooted in the social behaviour and rituals of HHI. FTAs, and FTA mitigation strategies, may not work in human-CA interaction the same way as they are theorised for human-human interaction.

Future work should aim to explore what aspects of joint action and joint commitment can be emulated and what aspects of this theory need to be appropriately tailored to more effectively support proactive CA interactions. For example, to understand if, and how, politeness can be used to design effective CA dialogues for mental health support, there is a need to understand how politeness theory translates in human-CA interaction.

5.5 Conclusion

In order to design effective CA systems for MHSM, we need to be able to design effective CA interactions. A major barrier to the design of effective CA interactions is the dominant CA interaction paradigm of interactions being user-led. In this chapter I explored the design of more proactive and collaborative CAs interaction for MHSM thus addressing RQ-A⁴. My design exploration was informed by theories of human-human interaction (joint action (Clark, 1996) and politeness (Brown and Levinson, 1987) theory) hence this

⁴RQ-A: How can CAs be designed to support people with MHSM?

chapter also investigates RQ-B⁵. In this chapter I first demonstrated that joint action is useful for conceptualising human-CA interaction and that concept of joint commitment can be used as framework for the design of more proactive and collaborative CAs interactions. Building on upon this joint action conceptualisation of human-CA interaction, I next suggested that the use of politeness by CAs may be especially important for the design of CA interactions for MHSM and then explored the design of polite CA interactions for MHSM.

This chapter proposed CA interaction theory and suggested theory-motivated design strategies. How these strategies will pan out in practice (e.g., how they will be experienced by users) remains unclear. Next, in Chapter 6, I report an interview-based study exploring how the use of politeness by CAs for MHSM impacts the user experience, as well as people's perspectives on the use of CAs for MHSM more broadly.

⁵RQ-B: How can theories of human-human interaction inform the design of CAs for MHSM?

Chapter 6

Interview Study: Investigating How Politeness Impacts the User Experience

6.1 Introduction

In order to design effective conversational agent (CA) systems for mental health self-monitoring (MHSM) we need to be able to design effective CA interactions for MHSM. Previously in Chapter 5, I explored how theories of human-human interaction (HHI) – namely joint action (Clark, 1996) and politeness theory (Brown and Levinson, 1987) – could inform the design of CA interactions for MHSM. I explained that politeness is important in HHI when asking people to engage in sensitive or inconvenient conversations. If politeness manifests similarly in human-CA interaction, it may play an important role in the design of CA interactions for activities that can be sensitive or inconvenient like MHSM. Although there is a nascent body of research investigating polite CA interaction, it is currently unclear how users will perceive the use of politeness by CAs for MHSM.

In this chapter I present a mixed-method interview-based study with 39 participants. The study addresses the two research questions of:

RQ1 In the context of MHSM, how does the use of politeness by CAs affect how they are perceived by users?

RQ2 What opportunities and challenges do people perceive in using CAs for MHSM?

Previously in Section 5.3.5, I presented three CA prompt sequences that varied in their use of politeness. In order to explore the effects of politeness on the user experience, participants of this study interacted with three prototype CAs using each of the prompt sequences. This chapter addresses all three of sub research questions of this thesis. It addresses RQs A¹ and B² as it advances the theory-based design exploration of Chapter 5

¹RQ-A: How can CAs be designed to support people with MHSM?

²RQ-B: How can theories of human-human interaction inform the design of CAs for MHSM?

by evaluating interaction design strategies that were informed by HHI theory. This chapter studies participant’s perspectives on CAs for MHSM and thus also addresses RQ-C³.

The structure of this chapter is as follows:

- 1. Methodology:** In Section 6.2, I detail the methodology of this study which combines a controlled experiment with semi-structured interviews.
- 2. Quantitative Analysis:** Next, in Section 6.3, I report my quantitative analysis of how participants rated the user experience of interacting with each of the CA prototypes.
- 3. Qualitative Analysis:** In Section 6.4, I then report our qualitative analysis of how participants perceived the prototypes and of their perspectives on the use of CAs for MHSM.
- 4. Discussion:** In Section 6.5, I discuss the implications of our analysis for design and research.

I conclude this chapter by first emphasising that the use of politeness by CAs can both positively and negatively impact the user experience of MHSM. Second, I draw attention to the tensions our analysis surfaces about the ways in which it is appropriate for CAs to support people with MHSM. I go on to further investigate the types of support CAs could provide, and the appropriateness of them, in Chapters 7 and 8.

Credit Statement

This chapter is based on content from the paper titled “Exploring How Politeness Impacts the User Experience of Chatbots for Mental Health Support” published in the *International Journal of Human-Computer Studies* (Bowman et al., 2024). Parts of Chapter 5 were also based on content from this paper.

I led and managed this study in collaboration my supervisory team (Gavin Doherty, Benjamin Cowan, and Anja Thieme), Orla Cooney, Joseph Newbold, and Leigh Clark. I collaborated with Orla Cooney on some parts of this work. Orla and I worked on the study design together and both conducted interviews (approximately 50:50).

As detailed in Section 6.2.5, Orla worked with me on the early stages of the qualitative analysis which was a thematic analysis. The purpose of this collaboration was to involve multiple perspectives in analysis which is a common practise for conducting high quality thematic analysis (Bowman et al., 2023; Braun and Clarke, 2021; Braun et al., 2019; Soden et al., 2024). I iterated on the analysis independently (with feedback from my collaborators) revising it substantially to produce the analysis reported within this chapter. I conducted the quantitative analysis independently with feedback from my collaborators. The discussion is the result of multiple iterations. The initial discussion was written collaboratively with my co-authors. I independently (with feedback from collaborators) revised the discussion considerably through multiple iterations.

³RQ-C: What are people’s perspectives on the use of CAs for MHSM?

6.2 Methodology

To study how people perceive the use of politeness by CAs, with focus on the sensitive application of MHSM, we conducted an exploratory scenario-based within-participants study with 39 participants. We used a mixed-method analysis approach that combined a quantitative component which allowed us to test for anticipated effects, with a qualitative component which enabled us to gain an understanding as to why, and how, these effects may occur, as well as to explore unanticipated effects. The study involved, in a (online) laboratory setting, a controlled experiment during which every participant enacted a fictional MHSM scenario with each of the three CAs. Semi-structured interviews followed the experiment. The controlled nature of this study design enabled us to effectively compare how participants experienced interacting with each CA. The study was given ethical approval for low risk projects by the university ethics committee.

6.2.1 Participants

The study's 39 participants (22 women, 14 men, 2 non-binary, 1 gender fluid⁴) were young adults (age range: 18-35; Mean age= 24.4yrs; SD age=4.2yrs) who were predominantly in university education. Participants came from a diverse set of nationalities, with almost half of the participants reporting their nationality as Irish (41%, N=16). 36 participants self-identified their English fluency as fluent, 2 as high proficiency, and 1 as conversational.

Participants were recruited through university mailing lists and Twitter; they were offered a €15 (or equivalent) Amazon voucher for taking part in the study.

6.2.2 Experiment Design

The experiment had a within-participants design whereby each participant completed a MHSM interaction for a different fictional scenario with each of the three CAs. The order of CAs and scenarios was randomized such that the joint ordering of the CAs and scenarios (36 possible orders) was counterbalanced. The experiment was designed to analyse how the fixed effect of the CA politeness affects users' perceptions of the CAs.

CAs

Participants interacted with three CAs for MHSM termed Direct, Passive, and Personal. These CAs use the corresponding (Direct, Passive, and Personal) prompt sequences presented in Table 5.6. As discussed in Section 5.3, the Passive and Personal prompt sequences use different politeness strategies that may positively impact on the user experience of using an CA for MHSM. The Direct CA uses direct language (i.e., it does not use politeness strategies) and is included as a non-polite comparison. See Chapter 5, for a description of the theoretical background and design of these prompt sequences.

We created an interactive prototype in the style of a Microsoft Teams CA⁵ for each prompt sequence. Figure 6.1 presents an image of one of the prototypes. The design of

⁴Gender was collected inline with the suggestions of Spiel et al. (2019)

⁵We used the tool BotSociety which has since been discontinued

the CAs was kept minimal; each CA was named ‘MoodBot’ and the profile image was left as the default blank avatar. The prototypes only varied by the CA prompts.

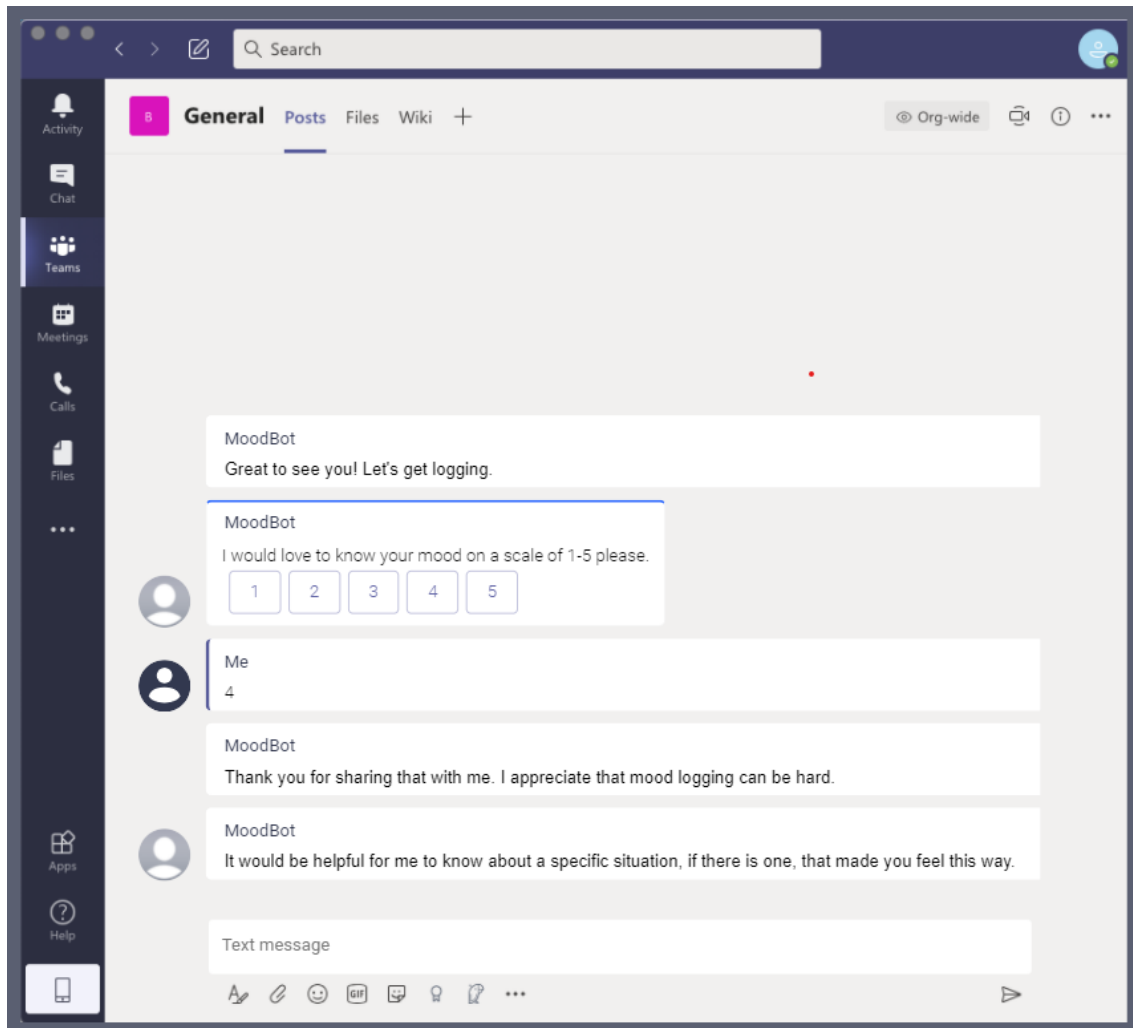


Figure 6.1: Image of CA Prototype

Scenarios

We chose for participants to make MHSM entries for pre-defined scenarios of slight positive valence to ensure consistency across participants, minimise repetition, and to minimise risk to participants. The scenarios were intended to be realistic and are presented by Table 6.1. We consider scenario a random effect.

Measures

Immediately following each CA interaction, participants were asked to rate on a 7-point scale (1=Strong Disagree and 7=Strongly Agree) their agreement with the statement that ‘*the agent was polite*’. This item was to serve as a manipulation test to evaluate how polite each CA was perceived, its wording closely resembles that used in previous work (Bickmore et al., 2007).

Table 6.1: Experiment Scenarios

Scenario	Description
A	You have just finished work for the week and are about to meet some friends for dinner.
B	You received some good feedback on a piece of work that you’ve been working on for weeks.
C	You’ve had a relaxed day watching television after a busy week.

In addition, participants were asked to rate the CA across a set of nine 7-point items we term the General Agent Rating (GAR) items (Table 6.2). We constructed the item set by consolidating items used in previous research to evaluate relational agents (Section 3.4) for healthcare applications (Bickmore and Picard, 2005; Bickmore and Ring, 2010; Bickmore et al., 2005a, 2007; Olafsson et al., 2017, 2020; Parmar et al., 2018; Ring et al., 2014; Zhou et al., 2014, 2017). Our intent with the GAR items was to broadly explore dimensions of user perceptions of CAs in order to identify potential effects of the use of politeness. Items were chosen in order to identify effects that could impact the suitability of CAs for different, particularly sensitive, applications. The items Easy; Express; and Natural concern the experience of communicating with the CA while the items Cares; Relationship; and Trust consider how the CAs are perceived as dialogue partners. The items Like; Satisfied; and Continue concern participants’ satisfaction with the interaction; Continue has been used in prior work (Bickmore et al., 2007) as a proxy for long-term adherence. We emphasise that, although the GAR items were given together at the end of each CA interaction and are displayed together in Table 6.2, they are not a questionnaire inventory and thus are analysed separately.

Ethical Considerations

As the study investigates CA interaction design strategies that may not be effective and that concern mental health, we must consider the potential for it to be upsetting and distressing to participants. To minimise the risk of participants being negatively impacted by the study, we did not specifically recruit participants experiencing mental health difficulties. We also used pre-defined scenarios of a positive valence such that participants would not be MHSM about their own experiences or scenarios with a high propensity for negative effect.

6.2.3 Interview Design

After interacting with all three CAs, participants took part in a semi-structured interview. A semi-structured interview approach was chosen as it allowed interviewers a level of flexibility in the topics discussed. While allowing for this flexibility, the interview covered five key topics: 1) participant’s general thoughts on CAs or conversational interfaces; 2)

Table 6.2: The General Agent Rating (GAR) items

Name	Item	Anchor 1	Anchor 7
Easy	How easy was talking to the agent?	Very easy	Very difficult
Express	How much do you feel you could express yourself?	Not at all	Very much
Natural	How natural was your conversation with the agent?	Not at all	Very natural
Cares	How much do you feel the agent cares about you?	Not at all	Very much
Relationship	How would you characterise your relationship with the agent?	Complete stranger	Close friend
Trust	How much do you trust the agent?	Not at all	Very much
Continue	How much would you like to continue working with the agent?	Not at all	Very much
Like	How much did you like the agent?	Not at all	Very much
Satisfied	How satisfied are you with the agent?	Not at all	Very satisfied

attitudes around the use of CAs to monitor mental health; 3) thoughts on the CAs used specifically in this study; 4) what participants liked or disliked about the language used by the study’s CAs; and 5) potential future use cases for CAs for mental health support. Discussion of the study’s CAs included screen sharing of the prompts used by the CAs in order to support participant recall, allowing them to comment on specific prompts and aiding in discussion clarity.

6.2.4 Procedure

Prior to the study session, participants were provided with an information sheet and were asked to complete an informed consent form. Each session was conducted by either the first or second author, using video call, and lasted around 30 to 60 minutes.

When a study session commenced, the participant was first briefed about the study procedure. The participant was then directed to a Qualtrics survey which first collected demographic information before running the controlled experiment. The participant and researcher remained on the video call during this, however the participant completed the survey independently.

Within the Qualtrics survey, participants were presented with each of the three CA-scenario pairs. Each pair was presented to the participant independently. The survey would show the participant the scenario in text-form and provide a link alongside it which

would open the CA in a new tab. The participant was asked to complete a MHSM entry with the CA as though they were in the given scenario. Immediately following each interaction, before interacting with the next CA-scenario pair, they were asked the manipulation and GAR items.

Upon completing the survey a semi-structured interview was conducted. The session concluded with participants being debriefed and thanked for taking part in the study.

6.2.5 Analysis Approach

Quantitative

Participants' ratings of the CA on the manipulation check and each GAR item were analysed using linear mixed-effects models (LMEMs). LMEMs are a regression analysis technique that allows for the modelling of both fixed (e.g., politeness) and random effects (e.g., by-participant or by-scenario based effects) when analysing the data. The models were run in R 4.1 (R Core Team, 2021) using the `lme4` 1.1-27.1 (Bates et al., 2015) and `lmerTest` 3.1-3 (Kuznetsova et al., 2017) packages. Following best practice guidelines (Barr et al., 2013) and similar analysis in related work (Edwards et al., 2019; Völkel et al., 2021; Wu et al., 2020), LMEMs for each questionnaire item were initially constructed using a maximal model approach (i.e., participant and scenario level random intercepts and slopes included). These models were then simplified as necessary until convergence was achieved. The full model syntax for the converged models along with full reporting of fixed and random effects is reported in Appendix A.

To identify effects of the polite CAs compared to the direct CA, the direct CA was set as the baseline condition within the CA fixed effect. Analysis was also conducted whereby the Passive CA was set as the baseline condition, allowing us to directly compare the ratings of the Passive and Personal CAs. Consistent with advice on exploratory statistical analysis in HCI (Cairns, 2019) and similar to recent prior work (Völkel et al., 2021) we do not explicitly correct the p value for the number of statistical tests conducted. The use of a correction with the number of tests would lead to Type II errors. However, due to the number of analyses conducted, findings that are statistically significant at a p value of or around .05 should be interpreted with caution.

Qualitative

Our qualitative analysis approach for data collected from the 39 participant interviews draws on reflexive thematic analysis (Braun and Clarke, 2013; Braun et al., 2019) while aligning with the structure of our research study.

All interviews were audio recorded and auto-transcribed. Resulting transcripts were then manually reviewed and corrected by the researcher who conducted the respective interview.

Orla Cooney and I, who had conducted the interviews, collaborated on the early stages of the analysis in order to include multiple perspectives in the analysis to promote exploration and reflexivity. Collaboration of this form is an established technique for conduct-

ing high quality thematic analysis (Bowman et al., 2023; Braun and Clarke, 2021, 2013; Braun et al., 2019) and qualitative analysis more broadly (McDonald et al., 2019). Our collaboration approach was guided by an approach described by Braun et al. (2019). We familiarised ourselves with the data set through reading the transcripts and note taking. Next, working with separate copies of the entire data set, we coded the data inductively, and mostly semantically, which aligns with our research aims of exploring and describing participant’s perceptions. I used spreadsheets and Orla Cooney used NVivo. We initially worked independently when coding the data and then started to increasingly confer to discuss, compare, and reflect on our codes as the analysis progressed. From our codings, we started to each construct candidate themes which we then developed into a shared set of themes through comparison and discussion. Throughout the analysis process we perceived our individual analyses to be very similar, we did not identify any major differences between our interpretations of the data. We attribute this, in part, to the structured nature of the data collection and the descriptive orientation of the analysis.

Reflexive thematic analysis is a highly iterative process that involves moving back and forth between analysis phases in order to produce a high quality analysis (Braun and Clarke, 2021). I further iterated on the analysis independently (with feedback from my collaborators) to further refine and advance the analysis. This involved exploring alternative candidate themes to create a set of themes that were better demarcated and that more clearly communicated the core concepts. During this process I reviewed candidate themes against codes, coded extracts, and the transcripts.

Our semi-structured interview study had two main purposes. Firstly, we wanted to better understand how participants specifically perceived the use of politeness by each of the three CAs; which determined our first three themes. Secondly, our theme construction involved creating themes more typical of reflexive thematic analysis to describe participant’s perceptions of CAs for mental health support more generally.

6.3 Quantitative Analysis

6.3.1 Manipulation Check - CAs Using Politeness are Perceived as More Polite

Descriptively, participants rated both the Passive ($M=5.64$; $SD=1.44$) and Personal ($M=6.05$; $SD=1.07$) CAs are more polite than the Direct ($M=5.13$; $SD=1.58$). As shown by Table 6.3, both the Passive and Personal CAs were rated at more polite than the direct with statistical significance ($\alpha=0.05$). Although descriptively the Personal CA was rated as more polite than the Passive, an analysis did not find this difference to be of statistical significance [Unstandardized $\beta=0.41$, $t=1.65$, $p=0.1043$; CI (95%) 0.02-1.00].

6.3.2 General Agent Rating Items - Personal Politeness Is Favourably Perceived

Table 6.4 presents descriptive statistics for the General Agent Rating items. Considering the mean ratings, the Direct CA was rated lowest and the Personal CA highest for every

Table 6.3: Politeness LMEM tests compared against Direct.

CA	β	CI (95%)	t	p
Passive	0.51	0.02 - 1.00	2.06	0.0430*
Personal	0.92	0.44 - 1.41	3.71	0.0004***

β values are unstandardised. (*, <0.05; **, <0.01; ***, <0.001)

item except Easy. For the Easy item, the Direct CA was rated highest and the polite CAs equally.

To compare the ratings of each CA we conducted analysis using LMEMs (Section 6.2.5) whereby the rating of the Passive and Personal CAs were compared to the ratings of the direct CA. The reported β values are unstandardised.

Table 6.4: Descriptive Statistics for Politeness rating and GAR items (Mean (SD)).

GAR Item	Direct	Passive	Personal
Easy	3.15 (2.08)	2.92 (1.68)	2.92 (1.87)
Express	3.90 (1.74)	4.36 (1.68)	5.03 (1.35)
Natural	3.26 (1.67)	3.74 (1.68)	4.36 (1.83)
Care	3.15 (1.69)	3.62 (1.87)	4.33 (1.77)
Relationship	2.41 (1.46)	2.97 (1.53)	3.38 (1.82)
Trust	3.33 (1.54)	3.64 (1.72)	4.21 (1.38)
Continue	3.79 (1.84)	3.97 (1.78)	4.56 (1.67)
Like	3.46 (1.76)	3.97 (1.53)	4.77 (1.61)
Satisfied	3.92 (1.92)	4.26 (1.76)	5.08 (1.46)

As shown by Table 6.5, for all GAR items — except *Easy* — the analysis shows the Personal CA was rated statistically significantly higher on each item than the Direct CA. There were however no statistically significant differences between the Passive and Direct CA ratings (all $p > .05$). These statistics suggest that for the GAR items, the CA using Personal politeness was perceived differently to the Direct CA, while the CA using Passive politeness was not.

In view of these findings we chose to conduct further LMEM analysis (Section 6.2.5) whereby the Passive CA was set as the baseline condition to enable direct comparison of the Passive and Personal CA ratings. As shown by Table 6.6 the differences were significant for 6 of the 9 GAR items. This indicates that, for the GAR items, the difference in perception

Table 6.5: LMEM tests Personal compared to Direct

GAR Item	β	CI (95%)	t	p
Easy	-0.23	-0.94 - 0.48	-0.63	0.529
Express	1.13	0.57 - 1.68	3.97	0.0002***
Natural	1.11	0.54 - 1.68	3.80	0.0003***
Care	1.18	0.52 - 1.84	3.49	0.0008***
Relationship	0.97	0.39 - 1.56	3.27	0.0016**
Trust	0.88	0.33 - 1.43	3.17	0.0023**
Continue	0.77	0.17 - 1.37	2.51	0.014*
Like	1.31	0.73 - 1.89	4.40	0.00003***
Satisfied	1.15	0.56 - 1.74	3.83	0.0003***

β values are unstandardised. (*, <0.05; **, <0.01; ***, <0.001)

between the Personal and Passive CAs is greater than that between the Passive and Direct.

The results for the Easy item show that participants did not perceive any CA as easier to talk to. However, the interactions with the Personal CA were rated with statistical significance as more natural than with both the Direct [$\beta=1.11$, $p=0.0003$ ***] and Passive [$\beta=0.61$, $p=0.0406$ *] CAs. Similarly, participants ratings demonstrate that participants felt more able to express themselves to the Personal CA than to both the Direct [$\beta=1.13$, $p=0.0002$ ***] and Passive [$\beta=0.67$, $p=0.0215$ *] CAs. These findings suggest that Personal politeness can support communication with CAs relative to Passive politeness and, more so, directness.

Participants characterised, with statistical significance, their relationship with Personal CA as more similar to a close friend than their relationship with the Direct CA [$\beta=0.97$, $p=0.0016$ **]. While descriptively participants characterised their relationship the Passive CA as between that of Direct and Personal, neither of the differences in ratings between the Passive and other two CAs were of statistical significance. In contrast, participants rated with statistical significance the Personal CA as not only more caring than the Direct CA [$\beta=1.18$, $p=0.0008$ ***], but also as more caring than the Passive CA [$\beta=0.72$, $p=0.0372$ *]. The Personal CA was trusted more by participants than the Direct [$\beta=0.88$, $p=0.0023$ **]; the Personal CA was also trusted more than the Passive CA [$\beta=0.56$, $p=0.0498$ *], however we caution that this p-value is close to the significance level. These findings indicate that the use of Personal politeness can support a more intimate relationship between the user and the CA.

The results for the Like and Satisfied items are of the largest effect sizes and most

Table 6.6: LMEM tests Personal compared to Passive

GAR Item	β	CI (95%)	t	p
Easy	-0.00	-0.77 - 0.71	-0.01	0.994
Express	0.67	0.11 - 1.22	2.35	0.0215*
Natural	0.61	0.04 - 1.18	2.08	0.0406*
Care	0.72	0.05 - 1.38	2.12	0.0372*
Relationship	0.41	-0.17 - 0.99	1.38	0.1722
Trust	0.56	0.01 - 1.10	1.99	0.0498*
Continue	0.59	-0.01 - 1.19	1.93	0.0576
Like	0.79	0.21 - 1.38	2.68	0.0091**
Satisfied	0.82	0.23 - 1.41	2.73	0.0080**

β values are unstandardised. (*, <0.05; **, <0.01; ***, <0.001)

significant p-values. Participants liked [$\beta=1.31$, $p=0.00003^{***}$] and were satisfied [$\beta=1.15$, $p=0.0003^{***}$] considerably more with the Personal CA than the Direct. Although with a smaller effect size, the ratings show participants also liked [$\beta=0.79$, $p=0.0091^{**}$] and were satisfied [$\beta=0.82$, $p=0.0080^{**}$] more with the Personal than the Passive CA. These findings strongly indicate that Personal politeness can improve the user experience of CA interaction. What's more, participants rated that they would like to continue working with the Personal CA to an extent significantly greater than they would like to continue working with the Direct CA [$\beta=0.77$, $p=0.014^*$]. Participants did not rate that they would like to continue working with the Personal CA to an extent significantly greater than the Passive CA [$\beta=0.59$, $p=0.0576$].

6.4 Qualitative Analysis

In this section, we describe the two main thematic findings that we constructed through our analysis pertaining to: (1) participants perceptions of politeness in each of the three CA variants; and (2) participants more general though on prospects of using a CA for the mental health task of MHSM.

6.4.1 User Perceptions of CAs using Politeness

Our analysis is organised according to participant's perceptions of each CA's politeness approach:

1. Direct: The Expected Robot,

2. Passive Politeness: Condescending and Too Apologetic,
3. Personal Politeness: The Conflict of Being Caring.

Theme 1: Direct - The Expected Robot

A few participants expressed a preference for the Direct CA, perceiving it as affording quick and efficient logging.

“I think the first one [Direct] was just very straightforward maybe, and it felt more like what you would expect from a chatbot.” [P17]

Similarly, some participants, while generally preferring one of the polite CAs, expressed that they would prefer the Direct CA in situations where they want to complete a MHSM entry quickly. For example, when they are short on time or when they are in a positive mood on which they do not feel the need to reflect.

The Direct CA was often described as robotic and contrasted with the two polite CAs which were assigned more ‘human’ attributes by participants; this suggests that the use of politeness by CAs has an anthropomorphising effect.

“[Direct] is very much just robot text, text, text and I think it’ll put people off.” [P09]

This robotic nature was deemed as potentially damaging to the success of a CA for MHSM. Participants described experiencing it as disconnected and emotionless, which generally dissuaded them from wanting to use it.

Theme 2: Passive Politeness - Condescending and Too Apologetic

A common opinion among participants was that the Passive CA was overly polite and more formal than its Personal counterpart. Although a few participants appreciated the CA acknowledging that it may be inconveniencing the user, many participants critiqued this apologetic stance. It was particularly felt that the “sorry to bother you” initiation was inappropriate. It was described that this initiation starts the interaction in a defensive way, and on a negative note, which implies that MHSM is something the user does not want to do. A number of participants pointed out that users of the CA would have likely already downloaded the application of their own free will, so the CA apologising to them for carrying out its purpose is unnecessary. In some cases this resulted in feelings of irritation.

“I have the choice to ignore it, if I wanted to if it - if it bothers me. So to me when a chatbot comes to me being like ‘sorry to bother you’, it just doesn’t feel natural” [P39]

The perceived formality of the CA led participants to liken the interaction experience to that of being in a doctor’s office or the workplace. These exemplify situations or places whereby politeness is often used as a means of reducing imposition on the hearer, as opposed to being indicative of a close relationship between the speaker and hearer.

“Almost like being in a doctor’s office you know, like ‘Sorry to bother you’, you know, ‘one last thing’ kind of thing, like it was a bit more formal” [P38]

As well as being perceived as more formal, many participants found the approach of the Passive CA to be condescending, passive aggressive, or sarcastic in tone, the same way a child may feel if scolded by their parents. It was suggested this could exacerbate any existing low moods of users engaging with the CA. Indeed, concern was even expressed that this tone could be potentially harmful.

“I think the dangerous thing about that, in a mental health setting, is that you don’t know what this person has actually written. So if they’ve pulled up something that’s very vulnerable or that was really difficult for them to write, like ‘Oh, I had a terrible day and I had a fight with my mom’ or something really that was quite intense and then you get a response like that [‘that wasn’t too hard’], it might come off the wrong way, of being condescending even though that wasn’t the intention at all.” [P03]

Theme 3: Personal Politeness- The Conflict of Being ‘Caring’

Participants commonly described the Personal CA as warm, comforting, and more emotional, in some cases likening it to a person or friend that wished to hear about the participant’s day. This perception led to an impression that the CA cared about the participants and was itself genuinely eager to hear their responses to each prompt.

“Like ‘you’ll tell me about what you’ve done today, I hope’, I like that ‘I hope’ part at the end, it’s like, kind of like ‘Oh, if you wish to’ like you feel like this person like cares about how you want to release your information or your emotions it’s not just ‘you’ll tell me’.” [P08]

Although some participants referred to the CA as if it were a person, others expressed this sentiment while concurrently acknowledging that they knew the CA was a machine.

“[Personal] just like sort of more like ‘great to see you’ so it was, it kind of felt like, I know it’s just a bot, but it kind of felt like it was happy to see me” [P24]

This ‘caring’ tone, in turn, led many participants to feel more open to disclosing their feelings to the CA.

“When I interacted with [Personal] it felt really sweet, and I was like aw okay I’ll tell you about my day.” [P23]

It was commented that the caring and compassionate nature of this CA could make it well suited for when the user is in an low mood and may be more open to spending more time reflecting on this mood and why they are experiencing it. For some, this ‘caring’ tone was recognised, but had an adverse effect. These participants felt the CA cared *too* much about their answers and were subsequently reluctant to share information.

“Even the ‘it would be helpful for me to know’, like why? What are you doing with my data? Why would it be helpful for you? I don’t want it to be helpful for you, I want it to be helpful for me” [P29]

Furthermore, a few participants described this behaviour as ‘pushy’. There appears to be a juxtaposition in that, although the Personal CA led people to feel that the CA was eager for them to engage, some of the phrases used left people wondering as to the intentions of the CA and its role during interactions with the user.

6.4.2 Using CAs for MHSM

This analysis explores participant perceptions of CA use for MHSM more generally. We consider: (1) A CA’s capability to understand the user; and (2) how users conceptualize who (or what?) they are disclosing to.

Theme 4: Understanding the User

Participants perceived potential for CAs to support users with self-understanding, yet this potential was considered dependent upon the capabilities of CAs to understand the user. Participants felt that CAs could encourage and support user reflection. It was perceived that CAs could prompt users to reflect on topics that are relevant to the user’s experience, but that the user would not independently think to reflect upon.

“They guide you kind of down the line of thought [...] it’s kind of like prompting you to reflect on things that might be affecting your mood. Whereas when you’re by yourself, I guess you can just not really pinpoint why you’re feeling that way.” [P01]

Yet, for this to be effective, and for the user experience to be reasonable, participants felt that CAs would need to engage users in way that makes them feel sufficiently heard and understood. It was suggested that CAs should respond in supportive and emotionally appropriate ways.

“You know, if I had a crappy day or something I would kind of want a little bit of, kind of like sympathetic feedback in a way. Like even just a comment from the Bot like ‘Oh’, you know ‘sorry that your mood is only...’ you know, ‘two out of five’ or something like that. Because, you know, I can kind of just write it down myself if I didn’t get those kind of aspects. I want to kind of have this you know, a little bit of feedback even though it’s, you know, a mood bot. [P36]

However there was also scepticism about the appropriateness of, and capabilities for, CAs behaving in an understanding way. This included critiques about it being inappropriate for CAs to comment of the difficulty of the activity for the user (i.e., how hard or not hard the MHSM was) that extended to concerns about CA interactions causing the user distress. Judgements made by a CA about the difficulty of a disclosure, or similarly assumptions

about how the user feels, may be misaligned with the user’s actual experience. The misalignment risks an invalidating experience, or one of feeling misunderstood, for the user whereby the user may feel the CA is implying they should feel differently. Accordingly, there are perhaps limits to the extent to which a CA should act as though it understands the user’s experience.

Participants perceived an opportunity for CAs to demonstrate understanding of users, and to support users with self-understanding, through use of the user’s data. CAs could, for instance, incorporate references to the user’s data into conversation to promote a user experience of feeling heard and understood.

“what would encourage me to keep using it would be if it would like continue on learning from what I say to it, like, like it will remember things about me. Like let’s say I feel really happy a week ago and the chatbot remember and is like ‘Oh, remember last week, you were feeling like that’.” [P05]

Many participants spoke of CAs integrating with other data sources about the user, such as their calendar or trackers relating to health and fitness, and using this information to support users in more relevant ways. For example, a participant described how a CA could make use of the user’s calendar to support them with anxiety:

“It would ask you, you know, ‘you’ve got a meeting on at two - how are you feeling about that?’ ... then follow up later on, like ‘you said you were nervous about the meeting... how did you end up feeling?’ and then you could end up getting this feedback at the end of the week, saying like ‘you end up getting really anxious about meetings but they go better than you think’” [P18]

Participants felt that CAs could help people interpret and gain insights from their data, but were often vague as to how. Some participants described functionality where the CA would analyse the user’s natural language disclosures in relation to their mental health to identify influential patterns of behaviour or factors. Functionality was also described where the CA would regularly (e.g., weekly) create textual summaries of the user’s logging for the user to reflect upon. The use of visualisations was also often suggested, possibly because this is what people have come to expect of data analysis. One participant suggested the use of word clouds to illustrate words associated with different moods.

Theme 5: Disclosure to Who?

Participants’ thoughts about the disclosure opportunity presented by a CA for MHSM varied around who they conceptualised the disclosure as being to. A small number of participants described the disclosure as being to, or with, themselves. In the case of these conceptualisations the CA is acting as medium for expression, similar to a journal.

“It’s almost like yeah like, like my own, like me talking to myself, you know? The chatbot becomes a reflection of myself, then I can just log my mood like that. I guess because I’m aware of the fact that it’s an inanimate object, or it’s a program, I’m more comfortable to kind of talk and, again it’s like

self-reflection, you know. Rather than confiding within somebody it's almost confiding within myself." [P01]

Most participants, however, described the disclosure as being to, or with, a CA whereby the CA is conceptualised more as a conversation partner. Participants perceived benefits of disclosure to CAs relative to disclosure to another person. With a CA, people do not need to worry about their conversation partner judging them, being disinterested, or reacting in a obviously biased or emotional way. For some, this can make the act of disclosure easier and subsequently allow for more honest disclosure. Consequently, for many, self-disclosure of this kind of information to a CA can be preferable over disclosure to a person.

"It's not going to judge you [...] it doesn't care about like how - like if you're sad or you're happy, it'll just be there." [P05]

"Personally I feel it'd be a bit easier disclosing it to the chatbot, because sometimes when you're talking to a person, you know the body language, the eye contact, you can kind of see if they're not into it, or if they're not, if they don't really care." [P38]

Furthermore, it was expressed that the conceptualisation of disclosure to a CA, rather than to one's self, may even promote more honest disclosure. A perceived reason for this is that when people describe their moods to themselves they may describe more preferable moods, potentially due to self-stigmatisation or toward promoting the experience of more preferable moods.

Some participants felt a CA for self-disclosure would be particularly beneficial for people who do not have another person to disclose to. Similarly, it was felt CAs could benefit people who struggle to disclose to other people, potentially helping them to become more comfortable with self-disclosure and thereby supporting them to disclose to other people in the future. Other participants expressed contrary views that enabling people to disclose to CAs, rather than people, may not be for the best. There was concern that disclosing to a CA could itself be a lonely or sad experience. There was also concern that people may come to prefer interacting with a non-judgemental CA over a potentially-judgemental person, and that this may lead them to reduce their interactions with other people.

Often in conjunction with describing the benefits of disclosure to a CA, relative to another person, participants spoke of how other people may view or use their disclosure data. In the knowledge of this, some participants commented that they would moderate their disclosures to a CA. A few participants described ways other people could make use of the user's MHSM data to support the user. Participants envisioned their data being shared with their therapist to help understand the onset of a mental health difficulty or to monitor treatment progress. One participant also perceived an opportunity for summaries of their data to be shared with their partner. Additionally, there were expectations described by a few participants of monitoring to identify, and then support, people at higher levels of risk (e.g., of suicide). Many participants were however concerned about who was collecting

their data and what their intentions were. This concern caused many participants to question if CAs should be trusted with their disclosures. Some participants were concerned about how people with access to their data may judge them.

“A negative of it would be, if I confide in it personally, say like something extremely personal, and then I feel weird because maybe somebody looking at it, on the other end of the screen, is laughing with their friends.” [P01]

Most concerns were about how the participant’s data may be used to manipulate them. A frequent concern was the use of user’s data to inform advertising.

“I wouldn’t want, you know, a conversational agent or a chatbot to know if I’m not feeling, you know, well, or if I’m not in a good mood and then just, it just starts, you know, throwing lots of advertisements towards me.” [P20]

Participants expressed that without a reputable backing - such as that of scientists, a government, or health service - they would doubt the effectiveness of the application. Distrust was particularly expressed about the prospect of a CA provided by a big tech or unknown organisation.

6.5 Discussion

In this section we will make clear that although politeness can positively impact the user experience, both the type of politeness and interaction context matter. We argue that to harness the potential of polite CAs, we need to be careful in our application of politeness theory, adapting it to the characteristics of CA interaction. We set out a research agenda for embedding politeness into the design of conversational interfaces.

6.5.1 Politeness Can Positively and Negatively Impact the User Experience

Our combined quantitative and qualitative analysis demonstrates clear differences between how the CAs using different politeness approaches were experienced by our participants. In contrast with the direct CA which was described as robotic, the polite CAs were assigned more human-like attributes which indicates that the use of politeness has an anthropomorphic effect. This result aligns with the conclusions of Morrissey and Kirakowski (2013) and Chaves and Gerosa (2021) that politeness (or manners) increases the human-likeness of CAs. Our study demonstrates that *how* politeness is used (i.e., the type of politeness) can substantially impact how a user experiences interacting with the CA. For the most part, interactions with the Personal CA were experienced as, and praised for, being friendly and caring. In contrast, for the most part interactions with the Passive CA were experienced as formal and criticised for being too apologetic and condescending. This is an important result as while both CAs were designed to be polite - as prior work has recommended - very different user experiences were produced. With the Personal CA being designed to use positive politeness and the Passive CA to use negative politeness,

the generally favourable experience of the Personal CA echos the work of Miyamoto et al. (2021) that showed positive politeness to be more favourably perceived in human-robot interaction. Consequently our analysis supports the reoccurring conclusion that politeness in human-CA interaction is impactful for the user experience (Chaves and Gerosa, 2021; Jain et al., 2018; Rapp et al., 2023; Yang and Aurisicchio, 2021). Our work emphasises that to harness politeness effectively within CA interaction design, we need to understand the impacts of different politeness approaches on the user experience.

The positive experiences of the Personal CA suggest that Personal politeness could be well suited for sensitive activities like those for mental health support. Personal politeness will likely suit interactions that, more generally, we would expect to benefit from more friendly and caring design. It could, for instance, be useful to develop sensitive and caring CA dialogues to support the self-reporting of sexual abuse experiences (Maeng and Lee, 2022) or the reporting of online threats (Piccolo et al., 2021). Furthermore, many participants' characterisation of their experience interacting with the Personal chatbot resemble the "polite and friendly" conversational environment advocated for by Rapp et al. (2023) for promoting collaborative human-CA interactions.

Passive politeness, however, does not seem well suited for MHSM. Participants disliked the apologetic behaviour of the Passive CA. The apologetic behaviour seems to run contrary to participant expectations of a MHSM CA, in that users will likely have expressly opted into the activity of MHSM before interacting with the CA, making the apology unnecessary. This misalignment of expectations suggests that Passive politeness may be poorly suited to situations where the user is already aware that the system will request them to interact. The Passive CA was also experienced as being condescending which, as commented by study participants, could cause the user distress, particularly if experienced during a sensitive interaction. This highlights how CA interactions for mental health support risk having a detrimental impact on the user's wellbeing if they are not carefully designed, including in how they use politeness. It is possible that for different applications, Passive politeness may better support the user experience than Personal Politeness. The apologetic behaviour of Passive politeness may be considered more appropriate for interactions that people consider as more inconvenient; for example, it may suit a customer service interaction to assist the user with a faulty product or missing delivery. Interaction with Passive CA was also perceived as formal, and likened to being in a doctors office, which may make it more appropriate than the friendly behaviour of Personal politeness for more formal interactions. For instance, in the context of mental health support, Passive politeness may suit more administrative interactions such as those for the organisation of sessions with a therapist.

Although many participants disliked the robotic nature of the Direct CA, some participants favoured it as they perceived it to afford a fast and efficient interaction. This preference for the Direct CA is similar to that expressed by the participants of Clark et al.'s study for interaction with voice user interfaces to be task-oriented (Clark et al., 2019b). Our analysis also showed that the caring nature of the Personal CA, despite being generally well perceived, incited distrust in some participants. Similarly when studying

perceptions of caring behaviours by conversational interfaces, Bickmore and Picard (2004) discovered a mixture of perceptions within their user groups, with many either liking or disliking the caring behaviour. These between-participant differences in how the CAs were perceived suggests individual differences in how the use of politeness by CAs is experienced by users. These differences may relate to the participant's orientation to social behaviour by CAs. Liao et al. (2016) showed that people have different orientations to social behaviour in human-CA interaction and Lee et al. (2011) found participants' orientation to social behaviour human-robot interaction affected how sensitive they were to the use of politeness by a robot. Our analysis, in conjunction with this prior work, suggests that for politeness to be used effectively within human-CA interaction it will need to be adaptable to the user's orientation to social behaviour by CAs.

Our analysis also suggests that, rather than using one strategy consistently, politeness prompts should be adaptable to contextual factors. For instance, our participants suggested that the Direct CA could be preferable when they are in a good mood, so as to facilitate a brief interaction. Yet if they are in a low mood, Personal politeness may be preferable as it affords a more compassionate interaction. This ties in with findings of Lee et al. (2019) (studying a virtual driving assistant) and Salem et al. (2013) (studying a robot receptionist) which relate the appropriateness of politeness to characteristics of the interaction task. In particular, the preference we identified of the Direct CA for brief interactions resembles the preference identified by Lee et al. of direct instructions during failure situations. Our findings broaden this relationship to consider, in addition to task characteristics, the momentary mental state and situation of the user. An implication of this is that, to make effective use of politeness, CAs may need to consider aspects of both the interaction task and the user's state, so as to dynamically adapt their use of politeness to these. These challenges are also faced by proactive conversational interfaces (Cha et al., 2020), where modelling of the user's cognitive state, context, and task is seen as critical to ensuring that conversational interfaces engage appropriately. Effective adaption may require carefully incorporating inferences about the user's momentary state into dialogue designs that support the quick identification of what interaction characteristics suit the user in that moment as discussed in Chapters 4 and 5.

In sum, the use of politeness can both positively and negatively impact the CA user experience. Our findings highlight the need to carefully design and evaluate the use of politeness by CAs, particularly for sensitive applications. For politeness to be used effectively, it may need to be adaptive to the user and the interaction context. Although future work needs to apply these strategies in wider contexts, our findings show promise in the potential for politeness to support human-CA interactions.

6.5.2 Face Threatening Acts in CA Interaction

As discussed above, the use of politeness by CAs impacted the user experience in a range of ways. Here we consider how our analysis of these impacts relates to politeness theory (Brown and Levinson, 1987).

As mentioned above, across both our quantitative and qualitative data analysis, Per-

sonal politeness was clearly seen as caring. This perception of the CA is consistent with how politeness theory informed its design. Our Personal politeness CA makes use of positive politeness, the redressive actions of which attempt to satisfy people's want to be desirable, typically by showing an interest in them (i.e., expressing care) (Brown and Levinson, 1987). Indeed, many participants felt this caring behaviour was supportive and felt encouraged to interact with the CA. Yet this was not seen positively by some participants, leading to a feeling that the CA was untrustworthy and subsequently making them reluctant to interact with it. As previously noted, this analysis resembles the mixed perception of caring behaviour by conversational agents previously identified by Bickmore and Picard (2004). A possible explanation for this is that users may not see the CAs as having the ability to be face-threatening in the first place. Our analysis (Section 6.4.2), and prior work (Section 3.3), have demonstrated that many people perceive self-disclosure to computers as preferable to self-disclosure to other people due in part to computers being perceived as non-judgemental. For some people, as they perceive computers to be non-judgemental, they may not perceive them as able to threaten their feelings of self-worth (i.e., positive face). This may complicate the use of politeness within human-CA interaction. For instance, if users do not perceive the interaction to be face-threatening, then a CA's redressive action when using politeness could suggest a potential face threat that the user needs to be aware of. This could in turn signal a reason for the user to distrust the CA, as some of our participants did.

Our qualitative analysis shows that participants experienced the Passive politeness CA as overly apologetic and condescending. This CA being perceived as apologetic is consistent with its use of negative politeness which includes the technique of apologising. As with the negative perceptions of the Personal CA, a face threat mismatch may also explain the negative perceptions of the Passive CA. Our Passive politeness CA made large use of negative politeness to mitigate threats to the user's autonomy. However, as the user has agency over the CA, they may not perceive it as capable of threatening their autonomy. Accordingly, the CA's negative politeness redressive actions (e.g., "Sorry to bother you") would be seen as unnecessary and thus overly apologetic.

Our explanation of the negative impacts of politeness on the user experience being the result of face threat mismatches corresponds with the discussion of Chaves and Gerosa (2021). Chaves and Gerosa describe dealing with face-threatening acts as being a challenge of manners in human-CA interaction. They argue that while it is natural for people to identify and mitigate FTAs, this is challenging for CAs due the complexity of identifying such acts. Our explanation suggests that, in part, this is complex because CAs are not perceived as having the same face-threatening capabilities as people.

The concept of face threat and how this influences perceptions of politeness must also consider the modality and fidelity of the conversational interface. Although our qualitative analysis (Section 6.4.1) found the polite CAs to be perceived as more human-like, other conversational user interface modalities and platforms (e.g., voice assistants and robots) may be perceived as more human-like than text-based CAs, resulting in different face threatening characteristics. For instance, work by Kocielnik et al. (2018) compared text

to voice interaction for reflection, finding that some participants perceived a pressure to respond to the voice interface’s questions immediately, allowing them less time to think of their answers. A complexity to this is that users may vary by the human- or machine-like characteristics they attribute to an interface, potentially impacting the ability of the interface to be face threatening (Hu et al., 2022). Further research needs to not only consider the differing characteristics of face in conversational human-computer interaction compared to human conversation, but also how modality and platform fidelity may impact the perception of face threat and the impact this may have on the user experience. If we are to design multi-modal interfaces as discussed in Section 4.4.2, developing this understanding will be important.

6.5.3 Tensions in Using CAs for MHSM

Our analysis demonstrates that participants positively perceived the prospect of using CAs for activities like MHSM. Participants described how CAs could integrate into people’s daily and digital lives, providing the user with emotional support and data-driven insight, helping the user develop self-understanding and supporting the user with challenges they may be experiencing. However, our analysis also surfaced concerns about this prospect which exist in tension with the perceived opportunities.

There is tension around the ways in which it may be appropriate, or desirable, for a CA to support a person with their mental health. For the most part, participants were appreciative of the caring and friendly behaviour of the Personal CA. Similar to prior work (e.g., Inkster et al., 2018; Kim et al., 2018; Lucas et al., 2014; Skjuve et al., 2021), many participants also perceived that CAs, being machines, could be relied upon to respond to users in a non-judgemental, engaged, and understanding manner. Due to these machine characteristics, concern was expressed about people finding it preferable to self-disclose to, and seek support from, CAs rather than people. This preference for CAs was one that some of our participants expressed. People increasingly receiving support from CAs may produce unrealistic expectations of how people can and should support each other. It is therefore a concern that the provision of support-giving CAs may negatively impact how people give and receive support from others. Reinforcing concerns raised by in related work (Morris et al., 2018; Prakash and Das, 2020; Skjuve et al., 2023; Vaidyam et al., 2019), as discussed in Section 3.10.3, the provision of CA support like this may have negative social impacts. Although many participants preferred the Personal CA to the Direct CA, the robotic behaviour of the Direct CA may promote a more appropriate human-CA relationship for mental health support where the CA presents less as a friend and more as a tool. Similarly, as discussed in our Understanding the User theme, it may be more appropriate for a CA to demonstrate understanding and provide support by harnessing the user’s data more and expressing care less. Other people can play important roles in the management of mental health and MHSM, as described in Section 2.6. Considering how CAs can support people with their mental health in conjunction with how people support people with their mental health may help us to navigate this tension. For example, one of our participants’ perceived opportunities is being able to share some of their data with their partner. The

work of Murnane et al. (2018), investigating how people’s social relationships and their use of technology for mental health management relate, offers a starting point for such considerations. Given this tension, future work should consider both the appropriateness and potential impacts of people receiving different types of support from CAs.

The participants of our study were cognisant of the fact that CAs are provided by people. There were concerns about the motives of the CA providers and about how user data may be used; some concerns were exacerbated by the Personal CA expressing a want to know about the user. These concerns undermine the perception of CAs being trustworthy and reliable machines and thus their potential for providing mental health support. To mitigate these concerns it will be important for systems to be transparent about their use of user data, as advocated for by guidelines on the design and evaluation of mental health technologies (Doherty et al., 2010).

6.5.4 Future Directions: Towards Polite CAs

The use of politeness clearly impacts people’s experience of interacting with CAs, as well as other conversational interaction modalities. However, as we discuss in Section 6.5.2, it appears that politeness does not manifest in human-CA interaction the same way it is theorised for human conversation. Toward making effective use of politeness when designing conversational interfaces, we need to develop new theory, or modify existing theory from human-human interaction, to reflect the nature of politeness in conversational human-computer interaction. Below we outline three key focus areas for future developments in this area. We note that these are by no means exhaustive, but are generated based on our analysis and experiences of studying polite conversational design within the context of MHSM:

Distinguishing Politeness Approaches

The results of this study and others (Lee et al., 2011; Miyamoto et al., 2021) demonstrate that different politeness approaches can have different effects. Although a politeness approach combining a diversity of politeness strategies may ultimately be most effective, this will require an understanding of the contexts in which different strategies are most appropriate. When developing such an understanding we need to be careful if using multiple strategies in conjunction with each other as this risks competing effects that could conceal each other. Research should therefore seek to distinguish different politeness strategies and identify their individual effects. A sensible starting point would be to apply the politeness theory of Brown and Levinson (1987) at the more granular level of positive and negative politeness sub-strategies. For instance, Torrey et al. (2013) focused on the negative politeness sub-strategy of hedging. Given the findings of this study, we may likely find some of these strategies are not appropriate for use by conversational interfaces. Critically, we will need to take care when describing how we apply politeness strategies in order for researchers and designers to be able to harness them.

Measuring Perceptions of Politeness

Due in part to the concurrent emergence of much research in this area, there is inconsistency in the measurement of the perception of politeness and concepts relevant to its effects. To support the comparison of findings, and to develop a unified understanding, we should identify or create appropriate measures and use them consistently. Often work has not attempted to measure politeness. Some work Bickmore et al. (2007); Hu et al. (2022) has used items resembling our ‘*the agent was polite*’ manipulation check item. Other work has not detailed the measure used (Lee et al., 2011), used a multi-item politeness scale (Lee et al., 2019; Salem et al., 2013), and also sought to measure the multiple constructs such as positive and negative politeness (Hu et al., 2022; Zojaji et al., 2020). To effectively study politeness we may need to identify the dimensionality of how this is perceived when interacting with conversational interfaces. This work could take the form of recent research that has looked to identify the dimensionality of agent personality (Völkel et al., 2020), partner models (Doyle et al., 2021) and human-likeness (Doyle et al., 2019). Such research could then form a basis for developing a validated scale to measure the dimensions relevant to perceptions of politeness in human-computer interaction for use in future work.

Exploring Adaptation and Context Awareness

Our analysis, as discussed in Section 6.5.1, and prior work suggest that a number of factors may affect how people perceive the use of politeness by conversational interfaces which include user characteristics (e.g., social-agent orientation (Lee et al., 2011; Liao et al., 2016)), the user’s momentary state (e.g., mood and activity (Cha et al., 2020)), and task factors (e.g., urgency, task-type (Lee et al., 2019; Salem et al., 2013)). As discussed in Section 6.5.2, perceptions may also be impacted by factors including the interaction modality (i.e., CA, voice assistant, or robot) and be mediated by anthropomorphism. Research should seek to experiment with these variables while accounting for, if not controlling, the others. As with the measurement of politeness, we should attempt to be consistent in our measurement of these constructs.

6.5.5 Ethical Implications

An unanticipated outcome was that some participants negatively experienced interacting with the Passive CA. If the study had been less controlled this could have caused distress to participants (e.g., as a result of a participant feeling the CA was being condescending during a sensitive interaction) which highlights the importance of staged evaluation (Doherty et al., 2010; Peters et al., 2020), whereby systems are first carefully evaluated in low risk settings.

As Passive politeness was negatively experienced it has the potential to be used maliciously (e.g., CAs designed to annoy or distress users). To protect users from intentional, or unintentional, harmful design we may want to develop techniques, or tools, for the detection of potentially harmful design characteristics. Tools could identify, or help designers to identify, opportunities to apply different linguistic (e.g., politeness) techniques

and then support designers to reason about the likely impacts on the user experience for that interaction context. Such tools may serve to help designers to identify parts of dialogue that warrant careful user testing and to demonstrate that a system is suitable for more real world use.

The Personal CA was perceived as more human-like and caring; some participants found this encouraged them to interact. While this outcome resembles an aim of this research, it also raises ethical concerns. Personal politeness could contribute to users over-trusting, or becoming over-reliant on, conversational agents. Indeed, as discussed in Section 3.10.2, prior work has previously raised concerns about users becoming dependent upon CAs (Laestadius et al., 2022; Ma et al., 2023; Xie and Pentina, 2022). Echoing broader concerns about conversational agents (Weidinger et al., 2022), this could become avenue for exploiting users.

6.5.6 Limitations

Our study aimed to explore how the use of politeness impacts the CA user experience for MHSM. Similar to other work on politeness use in interface design (e.g., Miyamoto et al., 2021; Torrey et al., 2013; Zojaji et al., 2020) we used a within-participant controlled experiment whereby participants enacted pre-defined scenarios with the CAs. This allowed us to gather data on participant perceptions of multiple politeness approaches, whilst also allowing participants to compare these under similar conditions and constraints. The use of pre-defined scenarios also enabled us to minimise the risk of participants being negatively impacted while engaging with participants on the topic on mental health (Hirsch, 2020). This approach is consistent with staged evaluation approaches for the development of mental healthcare technology (Doherty et al., 2010), as well as evaluation approaches for healthcare dialog systems (Bickmore and Giorgino, 2006). While MHSM systems should be evaluated longitudinally and clinically before real world deployment, this evaluation serves to first explore design possibilities and address the user experience. Informed by this, future work can look to explore ways of understanding how politeness impacts engagement with MHSM in more natural settings.

The quantitative component of our study used a set of 10 one item (1 Politeness and 9 General Agent Rating) ratings. This choice to use a set of single item measures, which as discussed in Section 6.2.2 have been used in relevant prior work, suited the exploratory nature of this study. The item set allowed us to explore a range of relevant dimensions without overwhelming participants with rating items. As research in this area starts to consider more natural settings, it should also seek to adopt validated measures. However, as highlighted in previous work (Clark et al., 2019a), the use of validated self-report scales within research on conversational interfaces is challenging due to there currently being a dearth of validated self-report scales for conversational interaction. We note that further work must look to design and utilise validated scales within this domain (e.g., for measuring politeness).

Our study’s specific focus on text-based CAs for MHSM may affect the generalisability of our analysis. For instance the sensitivity of MHSM may have a notable effect on how

users experience the use of politeness. Much of the previous literature, which we have used to situate this work, has studied politeness use by other conversational interface modalities (e.g., robots and voice user interfaces). As discussed in Section 6.5.2, interface modality and fidelity may impact the effects of politeness, thus it is unclear how our analysis will generalise across conversational interfaces. This is made more challenging due to the fact that politeness research currently uses a variety of measures, making it difficult to compare findings. It is imperative that future work looks to develop and validate measures that can be used to measure concepts relevant to the user experience of conversational interaction that may be influenced by politeness. This would then make it easier to compare and contrast findings across domains and devices.

6.6 Conclusion

Previously in Chapter 5, I explored the design of CA interactions for MHSM using a theory-based approach. This exploration suggested that politeness may play an important role in the design of CAs interactions for sensitive and inconvenient activities like MHSM. To investigate the impacts of the use of politeness by CAs, we conducted a mixed-methods interview-based study with 39 participants which involved each participant interacting with three prototype CAs based on the design exploration of Chapter 5.

Our analysis demonstrates that the use of politeness meaningfully impacted how participants experienced interacting with the CAs. We expected politeness to positively impact the user experiences, however our analysis shows that the impacts can be both positive and negative. This outcome affirms the criticality of careful CA interaction design for applications like MHSM as well the important role of politeness within CA interaction design. Within the discussion section of the chapter I relate the analysis of the impacts of politeness to politeness theory. I suggest the unexpected negative impacts are the result of differences between the face-threatening capabilities of people and CAs.

This chapter raised questions about the ways in which it is appropriate for CAs to provide MHSM support (Section 6.5.3). In particular, this work raises questions about the appropriateness of CAs behaving in ‘caring’ and ‘understanding’ ways. We need to develop further understanding of the appropriateness of different forms of CA support for MHSM in order to design effective MHSM CAs.

Chapter 7

Theory-based Exploration of CA Support Types

7.1 Introduction

In Chapter 6, I reported an interview-based study investigating how the use of politeness by mental health self-monitoring (MHSM) conversational agents (CAs) impacts the user experience and people’s perspectives on CAs for MHSM more broadly. My analysis of this study surfaced tensions about the ways in which it is appropriate for CAs to provide MHSM support. In order to understand in which ways it is appropriate for CAs to provide MHSM support, we need to identify the types of support that CAs could provide.

This chapter investigates the different types of support that CAs could provide for MHSM. I take an approach similar to that of Chapter 5, whereby I use theory of human-human interaction to help conceptualise and explore the design of CA support. Specifically, theories of social support are used which describe the support that people give to, and receive from, other people. This chapter builds upon prior work (Brandtzæg et al., 2021; Ta et al., 2020) which used social support theory as a basis for studying how CAs support people as briefly described in Section 3.7. I contribute a framework based on social support theory to more systematically explore the design of CAs for MHSM, specifically focusing on the forms of support they could provide.

First this chapter addresses RQ-A¹ by exploring the ways CAs could be designed for MHSM. Second this chapter addresses RQ-B² by investigating how theories of human-human interaction can inform the design of CAs for MHSM. This chapter is structured as follows:

- 1. Social Support:** First, in Section 7.2, I describe theories of social support and studies using social support to investigate CA support.
- 2. Support Exploration:** In Section 7.3, I next use social support theory as a basis for tabulating forms of CA support for MHSM.

¹RQ-A: How can CAs be designed to support people with MHSM?

²RQ-B: How can theories of human-human interaction inform the design of CAs for MHSM?

- 3. Support Combinations:** I describe, in Section 7.4, how the tabulation of support forms can be used to navigate the design space of MHSM CAs through the identification of combinations of forms of support.
- 4. Risks:** I then relate the risks of CAs for MHSM to different CA support types in Section 7.5.
- 5. Discussion:** Lastly, in Section 7.6, I discuss this theory-based design exploration and identify six design considerations for MHSM CAs.

I conclude this chapter by first emphasising the utility of social support theory for conceptualising and exploring CA support. I then stress the need for future work to consider the social dynamics of mental health support.

In Chapter 8, I go on to investigate people’s perspectives on how CAs can be designed to support people with MHSM. That work utilises this design exploration and is motivated by this chapter to take a social approach.

Credit Statement

This chapter is my own work informed by feedback from my supervisory team Gavin Doherty, Benjamin Cowan, and Anja Thieme. It is based on a currently unpublished paper titled “Exploring the Design of Conversational Agents for Mental Health Self-Monitoring: A Social Support Framework”.

7.2 Social Support

Social support is defined as support people give to, and receive from, other people (House, 1981). There are different types of social support, which can be given in conjunction. A prominent categorisation of the concept (House, 1981) that has previously been used to analyse CAs (Brandtzæg et al., 2021) describes four types of social support: emotional, instrumental, informational, and appraisal. Table 7.1 presents definitions of these types of social support. People can receive social support from a diversity of people. For example, people may receive support from partners, family, friends, colleagues, support groups, and healthcare professionals (House, 1981). Indeed, MHSM, and mental health management more broadly, tend to be social involving people beyond the individual (see Section 2.6). Table 7.1 also presents examples of social support people may receive for mental health management. People typically receive different types of support from different people. For example, a person’s partner and therapist may support them in different ways and with different problems (House, 1981).

Since CAs can mimic some aspects of human-human interaction, social support theory can provide a useful framework for exploring how CAs can offer support to users. By applying this theory, we can draw on the extensive research on social support and its effects on well-being, health, and behavior. However, we also need to acknowledge the differences between CAs and humans as sources of support, and avoid oversimplifying the complex nature of human-CA interaction. Therefore, as discussed in Section 3.9, we should

Type	Definition	Example
Emotional	Expressing empathy, care, trust, and love to the person.	Listening to the person and responding with care and love.
Instrumental	Doing things for, and to take care of, the person.	Cooking meals, driving to appointments and obtaining medication (Murnane et al., 2018).
Informational	Providing the person with information and advice that they can use to support themselves.	Give advice on MHSM apps to use (Schueller et al., 2021) and for behaviour to support their mental health (e.g., to go for a walk) (Kruzan et al., 2023).
Appraisal	Providing the person with information that supports them with self-evaluation.	Act as ‘social sensors’ monitoring and providing feedback on a person’s mental health (Matthews et al., 2017; Murnane et al., 2016).

Table 7.1: Social Support Definitions and Examples

be ready to modify our models and assumptions based on the specific characteristics and challenges of CAs as supportive agents.

A literature review, published in 2021, analysed empirical studies of social CAs to investigate *emotional* social support from CAs. The review argues that CAs show potential for providing emotional support because of their high availability and non-judgemental nature. However, their ability to provide emotional support is also limited by their conversational interaction capabilities (van Wezel et al., 2021).

Woebot Study/Source	Replika Study	Replika Source
Emotional	Emotional	Emotional
	Companionship	Social Network
Instrumental	Tangible	Tangible
Informational	Informational	Informational
Appraisal	Appraisal	Esteem

Table 7.2: Social Support Categorisations

Two studies have investigated perceptions of social support from conversational agents with functionality for mental health support. Woebot was studied to investigate how young

people (16 – 21 years) perceive different types of social support from a CA (Brandtzæg et al., 2021). As described in Section 3.5, Woebot is a CA designed to support people with their mental health through regular conversations using cognitive behavioural therapy techniques and mental health self-monitoring (Fitzpatrick et al., 2017). Sixteen young people participated in a qualitative study where they first used Woebot for two weeks and were then interviewed about the experience (Brandtzæg et al., 2021). Replika was studied to investigate the user experience of receiving social support from a chatbot (Ta et al., 2020). As described in Section 3.6, Replika is a CA designed primarily to be a companion (e.g., friend or partner), but also has functionality for supporting users with their mental health and mental health self-monitoring (Skjuve et al., 2023). Publicly available user reviews and a survey of Replika users were analysed (Ta et al., 2020).

Both of these studies structured their analysis using categorisations of social support (Table 7.2). The Woebot study (Brandtzæg et al., 2021) used the categorisation, previously presented in Section 7.2, of: emotional, instrumental, informational, and appraisal support (House, 1981; House et al., 1985). The Replika study (Ta et al., 2020) used a different categorisation of: emotional, informational, appraisal, *tangible*, and *companionship* support. The source cited (Cutrona and Suhr, 1992) by the Replika study for the categorisation names different support types to those used by the Replika study. The source describes the categories of: emotional, informational, tangible, *esteem*, and *social network* support. Tangible support, described by both the Replika study and its source citation, resembles informational support. Esteem support, described by the source citation, resembles appraisal support. Companionship support and social network support, described by the Replika study and its source citation resemble each other as they both refer to the enhancement of one’s sense of belonging. Companionship/social network support can be viewed as a form of emotional support. The use of different categorisations by these studies reflects the social support literature which struggles to reach consensus about the exact nature of social support, but agrees on its general nature (House, 1981; Taylor, 2011; Turner and Brown, 2010). Whilst categorisations of social support need to be treated with caution, in the scope of this work, they provide a useful lens to guide design exploration.

These analyses of social support from Woebot and Replika describe the perceptions of each type of support (Brandtzæg et al., 2021; Ta et al., 2020). I will review these analyses in greater detail in Section 7.3 as part of my use of social support theory as a framework for exploring forms of support from CAs.

A study investigating the design of CAs to support adolescents with their everyday problems used the concepts of emotional support and informational support (the term cognitive instead of informational). Adolescent participants role-played interaction between a human and CA where the CA would support the user with a difficulty they were experiencing. The dialogues participants created involved the CA offering advice on behavioural strategies, validation, and companionship. Dialogues commonly involved both emotional and informational support (Lopatovska et al., 2022)

Support matching is the social support theory idea that the value of social support depends upon how well the support provided matches the support desired. In human-human

interaction, providing mismatched support risks causing more psychological distress (Taylor, 2011). A study investigated support matching in interaction with CAs (Kang and Wei, 2018): A 2x2 between participants (N=132) experiment was conducted that involved participants interacting with a CA for stress management. Participants were assessed for their general preference for receiving emotional or informational support for stress management (support desire: emotional or informational). The CA would then provide them with either emotional or informational support. Significant interaction effects were identified such that participants with a preference for emotional support who received emotional support, rated the agent more highly for likeability and trustworthiness. This suggests people may have different preferences for the social support they receive from CAs (Kang and Wei, 2018).

7.3 Support Exploration

To understand how to design effective CAs for MHSM, there is a need to understand the ways CAs could support people with MHSM. Accordingly, my first research question for this chapter is:

RQ1: In what ways could CAs support people with MHSM?

My exploration developed from reflecting on the social support analyses of Woebot (Brandtzaeg et al., 2021) and Replika (Ta et al., 2020) in relation to my ongoing research on the design of CAs for MHSM. These works demonstrated that social support can be an insightful way to consider how existing CAs support people with their mental health. The analysis of these works is limited by the designs of the specific CAs analysed, however they do start to consider what other forms of social support could be provided by CAs. This motivated me to consider what forms of social support a CA could provide to support users with MHSM.

Initially I consolidated the analysis of these two works by mapping them to the same social support categorisation. I chose to use the categorisation described in Section 7.2 of: emotional, instrumental, informational, and appraisal support (House, 1981). I next developed this analysis from the forms of support provided by existing CAs to what forms of support could be provided by future CAs. I considered how types of support could be split into subtypes and considered what plausible minimum, moderate, and high levels of MHSM support from a CA could be for each social support type. I define these levels as:

Minimum: the level of support that is unavoidable.

Moderate: a level of support that would not be remarkable of a contemporary CA.

High: the most advanced level of support that could be provided at scale today, or in the near future.

My consideration was informed by my experience researching related systems and knowledge of related literature and directed by a focus on the relevance and plausibility of different forms of support for MHSM. Table 7.3 terms the forms of support I identified, which are expanded on further below.

Type	Subtype	Minimum	Moderate	High
Appraisal	Information	None	General	Targeted
	Reflection	Records	Questions	Processes
Informational	Information	Referral	General	Tailored
	Suggestion	Referral	General	Tailored
	Reminder	None	Basic	Aware
Instrumental		Stores	Analyses	Manages
Emotional		Functional	Caring	Intimate

Table 7.3: Tabulation of Support Types and Levels.

7.3.1 Appraisal Support

Appraisal support is support to help people with self-evaluation (House, 1981). Appraisal support is relevant to MHSM as self-evaluation is a core process of MHSM. In human-human interaction, appraisal support involves a transfer of information whereby the supporter gives the person information that can inform their self-evaluation (House, 1981). It has been suggested that appraisal support is different in human-CA interaction with it instead involving the agent facilitating the person with self-reflection, which then informs self-evaluation (Brandtzæg et al., 2021). However, both types of support are possible in both human-human and human-CA interaction ³. For the purpose of considering ways CAs can support people with MHSM, I include both appraisal processes. I term the two processes:

Appraisal Information: giving the user information that could support self-evaluation, and

Appraisal Reflection: facilitating the user with self-reflection that could support self-evaluation.

Appraisal Information

As a digital tool, and especially as a tool for MHSM, a CA has substantial potential for providing the users with information that can support self-evaluation. The provision of appraisal information support is a major opportunity of digital tools for MHSM. Proficient analysis of mental health self-monitoring data requires both strong data and mental health

³Self-evaluation through facilitating self-reflection is possible in human-human interaction (e.g., counselling interactions). Self-evaluation through information transfer is possible in human-CA interaction (e.g., a CA could provide analysis of user's data or other relevant information)

literacy (Matthews et al., 2017). Many users, and their human social-relations, will not have sufficient literacy for proficient analysis. Automated analysis could address this lack of literacy through helping users to understand their self-monitoring data. Furthermore, automated analysis could mitigate the risks of inappropriate analysis (Murnane et al., 2016).

None: At a minimum, we may expect for a CA to be able to show users their self-monitoring data (e.g., their mood entries). That said, there may be circumstances where this is not desirable. For instance, mental health professionals have expressed concern about patients (adolescents) reviewing their data independently (Matthews and Doherty, 2011).

General: Higher levels of appraisal-information support could involve the CA providing increasing levels of analysis of user data. A moderate level of support could be the CA presenting descriptions (e.g., descriptive statistics, visualisations, or textual description) of user’s data. It could also involve the CA sharing data with the user, beyond their own entries, such as data collected by passive monitoring (e.g., sleep) or data from secondary sources (e.g., weather).

Targeted: A high level of support could be a CA sharing targeted support to aid self-evaluation whereby the CA directs the user to the most insightful or actionable analysis. For example, this could be informing the user of a relation between their activities and mental experiences or of a trajectory (e.g., mental health decline).

Appraisal Reflection

CAs being mediums for self-expression can facilitate self-reflection and therefore provide appraisal-reflection support. It was reported that interaction with both Woebot and Replika facilitate self-reflection and introspection (Brandtzæg et al., 2021; Ta et al., 2020). The Woebot study discusses how participants described this interaction as being like having a conversation with themselves (Brandtzæg et al., 2021). I similarly reported within my analysis of the interaction design interview study (Chapter 6) some participants conceptualising the interaction as being like having a conversation with their self. However, most participants of the interaction design interview study described the interaction as being a conversation with an agent. Furthermore, the Woebot study describes how Woebot’s self-monitoring functions helped users to develop self-awareness of their emotional state (Brandtzæg et al., 2021).

Records: It is difficult to imagine a CA for MHSM that does not record self-monitoring entries from the user (e.g., about their mood). As such, some appraisal-reflection support seems unavoidable.

Questions: A moderate level of appraisal-reflection support could involve a greater level of questioning of the user. This may involve asking more questions, asking more tailored questions, or asking follow up questions in a responsive way (see Section 3.3.3).

Processes: A higher level of appraisal support could involve the CA contributing in a more expert way. For example, Woebot, the design of which is informed by cognitive behavioural therapy, processes user input to help users to recognise negative thought patterns (Brandtzæg et al., 2021). Seemingly, increasing levels of appraisal-reflection support will increasingly resemble expert-level talking therapy interactions.

7.3.2 Informational Support

Informational support involves giving people information and advice that they can use to support themselves (House, 1981). Both the Woebot and Replika studies report that people are appreciative of these agents as sources of informational support. A strength of CAs identified by both studies is that the agents are almost always available (Brandtzæg et al., 2021; Ta et al., 2020). As such, agents can provide immediate information and advice in the event of momentary mental health difficulty or crisis (Ta et al., 2020). The Woebot study reports that Woebot can be perceived as a trustworthy information source. Participants contrasted informational support from Woebot with informational support from friends, commenting that Woebot has the ability to provide more factual, rather than anecdotal, information. Additionally, Woebot can be perceived as an approachable information source that people could talk to without feeling embarrassed. Participants also liked that Woebot offered the information in a proactive manner (Brandtzæg et al., 2021). Users of Replika appreciated its ability to incorporate information about previous interactions (Ta et al., 2020). Participants of a study of how CAs could support the emotional needs of young people similarly reported a desire for CAs to provide data-based advice (Kim et al., 2018). CAs could support people here by providing informational support informed by analysis of the user's data. As a minimum, we should expect that mental health applications provide basic crisis informational support (e.g., advise the user to seek emergency support).

I distinguish informational support into:

Informational Information: the giving of information, and

Informational Suggestion: the giving of suggestions, and

Informational Reminder: reminding to do things.

Both suggestions and reminders are forms of advice and thus can be viewed as informational support. Reminders tend to be an important function of tools for supporting mental health self-monitoring (Section 2.2.5).

Informational Information

A CA for MHSM could provide the user with a range of potentially helpful information. Levels of information support can be characterised by how tailored the information is to the user.

Referral: The CA provides the user with contact details for emergency services.

General: General educational material about mental health and MHSM could be provided, as well as information about other resources for mental health support (e.g., talking therapy services). Furthermore, the CA could teach the user exercises and techniques for mental health management (e.g., worry management or breathing exercises).

Tailored: Information could be tailored to the specifics of the mental health difficulties the user is experiencing (e.g., to a specific mental health condition such as bipolar disorder) and to the resources available to them locally (e.g., local talking therapy services).

Informational Suggestion

Suggestion is similar to information, but rather than just provide information the CA suggests the user does, or does not, do things. As with information, levels of suggestion can be characterised by how tailored they are to the user.

Referral: The CA suggests the user contacts emergency services.

General: The CA suggests things that are generally understood to support a person's mental health or MHSM. A CA could, for example, suggest the user tries to develop a regular sleep routine, or to attempt to record their mood at least daily. An agent may also encourage the user to seek further (non-emergency) support from other sources such as friends, family, and healthcare professionals (e.g., Bickmore et al., 2010b).

Tailored: The CA makes suggestions that are tailored to the user's mental health or MHSM needs. A CA could suggest behaviour changes, or routines, that are likely to positively impact the specific user's mental health or MHSM based on analysis of the user's data and preferences.

Informational Reminder

A CA could remind people: to make self-monitoring entries, perform activities, and of healthcare appointments. Levels of reminder support could be characterised by the awareness of reminder delivery, and the activity the user is reminded of. Reminders are similar to suggestions, both being types of advice. However, with reminders there is already some form of agreement between the user and the CA to do what they are being reminded of.

None: A CA could not provide reminder support.

Basic: A CA could remind the user to make a monitoring entry at a fixed or random time each day.

Aware: A CA could remind the user to do things at opportune moments in time. For example, it could remind the user to make an entry during idle moments (e.g., when riding the bus) or after mood impacting activities (e.g., exercise) for behavioural activation (See Sections 2.2.5 and 2.4).

7.3.3 Instrumental Support

Instrumental support involves doing things for, and to take care of a person (House, 1981). Analysis of existing agents has argued these agents do not provide instrumental support and related this to agents being digital rather than physical entities (Brandtzæg et al., 2021; Ta et al., 2020). It has been argued that other agents could provide instrumental support in the form of digital services (Brandtzæg et al., 2021). A good example is given of an agent assisting the user with contacting a psychologist ⁴. More broadly, we can view CA functions that do things for the user as instrumental support.

Advanced CAs could offer physical instrumental support akin to how instrumental support is viewed in human-human interaction. Integrated into a smart home, a CA could perform household actions such as opening curtains to facilitate morning routines. CAs like these are not a near future possibility for most people and therefore will not be considered further.

Stores: A CA can store the user’s MHSM data ⁵.

Analyses: A CA could analyse the user through collecting data to support the user with MHSM from passive sensing (e.g., physiological and behaviours), the user’s social relations, and external sources (e.g., weather). A CA could also analyse the user’s MHSM data. As previously discussed, analysing the user’s data can also enable higher levels of appraisal and informational support.

Manages: A high level of instrumental support would be for CA to take a more active role in the management of the user’s mental health. Consistent with the suggestions of prior work (Brandtzæg et al., 2021), this could involve the CA supporting the user with managing appointments with healthcare professionals (e.g., scheduling and preparation) and medication (e.g., prescription management). Indeed, Bickmore et al. (2010b) previously designed and studied a relational agent to support people with anti-psychotic medication management.

7.3.4 Emotional Support

Emotional support is perhaps the least clearly defined type of support. While definitions for other types of support describe the nature of the support, emotional support is often defined (e.g., Cutrona and Suhr, 1992; House, 1981) by a list of example acts (e.g., empathy, care, love). This is complicated further by there being a large number of acts that are commonly considered to be emotional support.

Studies argue that CAs like Woebot and Replika can provide emotional support to users. Furthermore, they can be perceived as safe sources of emotional support, in comparison to other people, as being computers they can be trusted to respond in a non-

⁴Other examples provided are of a CA providing advice or suggestions (Brandtzæg et al., 2021). These types of support would typically be viewed as informational rather than instrumental support.

⁵Although I argued that CAs may not show user’s their data (Section 7.3.1), in that case the data was still being stored for access by a healthcare professional. It is hard to imagine an MHSM agent that would not store the user’s entries.

judgemental and engaged manner (Brandtzæg et al., 2021; Ta et al., 2020). This perception of agents as something safe to disclose to is a recurring research finding as described in Chapter 3 and affirmed in Chapter 6.

Functional: A minimal CA for MHSM will have functionality for the user to record their monitoring entries over a sustained period of time. Additionally, a prerequisite for use is that the user sufficiently trusts the CA with their self-monitoring data. These functional characteristics resemble acts of emotional support: listening, providing an extended period of care, accompaniment through stressful situations, and reflecting trust (House, 1981). CA for MHSM can therefore be viewed as offering at least a functional level of emotional support.

Caring: A more moderate level of emotional support is for CA to behave in a caring manner. Caring behaviour could include providing reassurance, encouragement, and expressing concern (House, 1981). In the mental health domain, it is common for CAs to be designed to behave empathetically (e.g., Fitzpatrick et al., 2017; Fulmer et al., 2018; Inkster et al., 2018). Relational behaviour by agents, which is caring in nature, has been found to both promote user’s trust in (Bickmore and Cassell, 2001) and desire to continue interacting with (Bickmore et al., 2005b) CAs. What’s more, the interaction design study reported in Chapter 6 demonstrated that use of politeness by CAs for MHSM can result in CAs being perceived as caring and that participants generally viewed this behaviour positively.

Intimate: A higher level of emotional support would be providing companionship, intimacy, or love. Replika is an example of a CA that provides an intimate level of support. Replika will converse with users about topics relating to the user’s mental health (e.g affective and emotional; routine self-reflection). Additionally, it will also converse with the user about their everyday life, hobbies, and interests. Furthermore, Replika will enact (describe with the user) daily activities (e.g., having a meal or watching a movie together) and physical intimacy (e.g., hugging, kissing, and sexual activities) (Skjuve et al., 2023).

7.4 Support Combinations

CAs for MHSM will provide combinations of forms of support. Many possible combinations of support could be identified using my support form tabulation by choosing a level of each type of support⁶ A challenge is to identify which support combinations best match the MHSM support needs of users. Table 7.1 present three example support combinations that I identified which I term: Reflection, Analysis, and Emotional. I identified these support combinations as plausible instantiations that prioritise different forms of support in order to demonstrate, to an extent, how varied support combinations can be. I first excluded the minimum levels of appraisal and information support because it seems sensible for

⁶Naively 2187 (3^7) support combinations could be identified.

most MHSM CAs to provide at least a moderate level of these. The rationale for each combination was then as follows:

Reflection: Chosen to have high levels of appraisal-reflection and informational-information to prioritise support relating to the user's understanding of their mental health, and mental health more broadly. These high levels forms of support require analysis of user data, therefore instrumental-analyses support. Caring-emotional support is chosen as this may facilitate appraisal-reflection support.

Analysis: Chosen to have high levels of appraisal-information, informational-suggestion, and informational-reminder support to prioritise support relating to data. As with the Reflection combination, the Analysis combination's high level forms of support requires analysis of user data so also instrumental-analyses support. Emotional support is of less relevance to the Analysis combination so I choose the minimum level of functional emotional support.

Emotional: The Emotional support combination was chosen to prioritise emotional support by having providing an intimate level of it. Unlike the Reflection and Analysis combinations, this combination does not require analysis of user's self-monitoring data so instrumental support is set to the minimum level of stores.

I make further use of these support combinations in Chapter 8.

Reflection							
	Appraisal		Informational			Instrumental	Emotional
	Information	Reflection	Information	Suggestion	Reminder		
Minimum	None	Records	Urgent	Urgent	None	Stores	Functional
Moderate	General	Questions	General	General	Simple	Analyses	Caring
High	Targeted	Processes	Tailored	Tailored	Aware	Manages	Intimate

Analysis							
	Appraisal		Informational			Instrumental	Emotional
	Information	Reflection	Information	Suggestion	Reminder		
Minimum	None	Records	Urgent	Urgent	None	Stores	Functional
Moderate	General	Questions	General	General	Simple	Analyses	Caring
High	Targeted	Processes	Tailored	Tailored	Aware	Manages	Intimate

Emotional							
	Appraisal		Informational			Instrumental	Emotional
	Information	Reflection	Information	Suggestion	Reminder		
Minimum	None	Records	Urgent	Urgent	None	Stores	Functional
Moderate	General	Questions	General	General	Simple	Analyses	Caring
High	Targeted	Processes	Tailored	Tailored	Aware	Manages	Intimate

Figure 7.1: Example Support Combinations

7.5 Risks

As demonstrated, there is opportunity for CAs to provide diverse forms of support to users for MHSM. These opportunities also involve risks. My research question for this stage of the design exploration is:

RQ2: What are the risks of CAs providing different forms of support for MHSM?

Table 7.4 lists key risks identified in relation to each type of social support; this set of risks is non-exhaustive. Additionally, I identified risks posed by the use of CAs for MHSM to social ecologies of mental health support. As with my consideration of support forms, my consideration of associated risk was informed by our experience researching related systems and knowledge of related literature. I aimed to emphasise the risks most relevant to and plausible for MHSM. While these risks have already been introduced within this thesis, considering them in relation to support types is informative for considering the appropriateness and challenges of designing CAs to provide different types of support. I next briefly summarise the risks.

7.5.1 Risks of Appraisal Support

Inappropriate Self-Evaluation: Lower levels of appraisal-information support risk users reaching conclusions through independent data analysis that are inappropriate from a data analysis or medical perspective. Higher levels of appraisal-information support (e.g., more automated analysis) could mitigate risks of independent data analysis by users. However, with higher levels of appraisal-information support, there is risk that users may insufficiently understand the limitations of the automated analysis and consequently be excessively trusting of it. Users may, for example, trust automated analysis more than their own lived experience (Matthews et al., 2017). Higher levels of appraisal-information support will need to be designed to effectively communicate their limitations to users to mitigate such risks.

Distressing Self-Evaluation: Self-evaluation, facilitated by both forms of appraisal support, may cause the user distress. First, recording negative data can negatively impact a person’s mental experience (Schueller et al., 2021). Second, frequent self-evaluation can cause people to feel their mental experience too much, potentially leading them to feel overwhelmed (Matthews et al., 2017). Third, people may realise through reflection or identify through analysis things that are upsetting for them (e.g., negative behaviour patterns) (Kelley et al., 2017). Fourth, people have reported that identifying they are on a negative mental trajectory can in-turn contribute to that trajectory (Murnane et al., 2016). CAs, particularly in contrast with people, may lack the capability to identify when a user is being distressed by self-evaluation and respond to this appropriately.

Harmful Application: CAs for MHSM risk facilitating activities that are detrimental to the user’s wellbeing. While self-monitoring practices can be used to work toward positive health goals, they can also be used to work toward negative health

Support Type	Risk	Example
Appraisal	Inappropriate Self-Evaluation	User makes analytically incorrect inferences from their data.
	Distressing Self-Evaluation	User's low mood is exacerbated by act of reflecting on it.
	Harmful Application	CA is used to facilitate rumination.
Informational	Insensitive Interaction	CA suggests the user talk to a person when the user feels there is no person they can talk to.
	Incorrect Interaction	CA gives advice that is medically inappropriate for the user.
	Threat to Agency	Being repeatedly told what to do causes the user to feel a loss of control.
Instrumental	Threat to Agency	Things being done for the user causes them to feel a loss of control.
	Threat to Privacy	Intimate and large amounts of user data may be collected.
	Execution Failure	Inaccurate data collection due to sensor errors.
Emotional	Inappropriate Behaviour	Caring or intimate behaviour by a CA makes the user uncomfortable.
	Stigma	Users may be distressed by other people reacting negatively to their use of a CA.
	Threat to Validity	Comforting effects of CA interaction bias user self-reports of mood.
	Dependency	User becomes dependent upon the CA for support to manage their mental wellbeing.

Table 7.4: Support Type Risks

goals (Kelley et al., 2017). A particular concern with MHSM and appraisal-reflection support is that it risks facilitating rumination (Eikey et al., 2021). The design of MHSM CA will need to consider how MHSM practices that risk being detrimental to the user’s wellbeing can be identified, and upon identification, how to appropriately support that user to mitigate risks of harm (Fiske et al., 2019).

7.5.2 Risks of Informational Support

Insensitive Interaction: MHSM CAs risk providing information, suggestions, or reminders to users that are experienced negatively, for example as annoying or insensitive. Suggestion and reminder support may be of greater risk than information support as they are more directed (i.e, they are advice). I summarise research reporting examples of CAs providing insensitive informational support in Section 3.10.1. The common perception of CAs being non-judgemental (e.g., Brandtzæg et al., 2021; Ta et al., 2020) may exacerbate the negative experience, or distress, caused by insensitive CA behaviour.

Incorrect Interaction: CAs risk incorrectly, or insufficiently, understanding the user leading them to provide medically incorrect advice. As I discussed in Section 3.10.1, concern has been expressed about the ability of CAs that afford unconstrained user input to give appropriate and safe medical advice. A contemporary concern is the risk of inaccurate or inappropriate information being produced by generative machine learning models (e.g., LLMs) (Weidinger et al., 2022). Perceptions of CAs as trustworthy and data-based sources of information (e.g. Brandtzæg et al., 2021; Kim et al., 2018) may exacerbate the risk posed of incorrect or inappropriate advice.

Threat to Agency: A risk of higher levels of informational suggestion and reminder support is threatening the user’s feelings of agency. A common benefit people report of MHSM is that it gives them greater feeling of control (agency) in their life (Matthews et al., 2017; Murnane et al., 2016). With high levels of informational suggestion and reminder support, the CA is repeatedly advising the user about what they should do. Users may not take well to being repeatedly told what to do as it may reduce their feelings of agency. The extent of this risk is unclear. As I discussed in Chapter 6, based on the analysis of the interaction design study, messages from CAs may have little ability to threaten people’s feelings of agency as the user 1) opted-in to them and 2) can easily ignore them..

7.5.3 Risks of Instrumental Support

Threat to Agency: Higher levels of instrumental support also pose a threat to the user’s feelings of agency. This threat posed by instrumental support is likely greater than that of informational support due to its more task-based nature. Research participants have previously described how more automated MHSM could cause them to feel as though they have no role in their own life (Matthews et al., 2017). Care must be taken when designing CAs to balance reducing user burden with maximising the

user's feelings of autonomy. This could involve designing for collaborative interactions (as explored in Chapter 5) that involve the user, while allowing the CA to do the work.

Threat to Privacy: Higher levels of automated data collection of the users digital and physical activities, as well as physiological data, pose an increasing threat to user's privacy.

Execution Failure: A risk of instrumental support is failure to execute tasks correctly. This could include collecting or analysing data incorrectly. Execution failures could in turn compromise the provision of other forms of support (e.g., appraisal and informational) (Fiske et al., 2019).

7.5.4 Risks of Emotional Support

Inappropriate Behaviour: Not all people are receptive to caring behaviour from CAs. For example, the interaction design study of Chapter 6 demonstrated that the caring behaviour of the Personal politeness CA elicited distrust from some participants who were sceptical of why the CA cared about them. More intimate behaviour may also be experienced as inappropriate by users. Research has for example reported users of Replika being upset by the CA behaving in unsolicited intimate ways (Ma et al., 2023).

Stigma: As described in Section 3.10.4, prior research has reported a stigma associated with interacting with CAs in intimate ways (Ma et al., 2023; Skjuve et al., 2021). Users may feel uncomfortable being open about, or experience other people reacting negatively to, their use of a CA because of this stigma.

Threatening Validity: The provision of emotional support by a CA for MHSM may have effects on the user such as comforting them. Such effects may bias the collection of self-monitoring data and consequently be detrimental to the validity and effectiveness of MHSM. However, the provision of emotional support may transform MHSM from an activity that feels too challenging for a user to one that feels more manageable (albeit of, potentially, compromised validity).

Dependency The provision of emotional support that is always available and comforting is something that users may become dependent on (Laestadius et al., 2022; Xie and Pentina, 2022). It may, for example, be challenging for people to stop using CAs (e.g., when the MSHM was meant to be fixed-term) because it is challenging for them to terminate the relationship (Skjuve et al., 2022).

7.5.5 Risks to Support Ecology

Using social support theory, that is theory of how people support other people, to consider support from CAs leads us to consider the role of CA support in relation to human support. MHSM, and mental health management more broadly often involve more people than just

the individual. Healthcare professionals, and a person's friends and family, can have impactful roles (Section 2.6). Prior work has described how this combination of support sources forms a social ecology of support (Murnane et al., 2018).

The introduction of a CA for MHSM into a person's support ecology may negatively impact the support ecology and the overall support that person received. A concern is that the person's use of the CA may change how they seek and receive support from other people (Fiske et al., 2019; Miller and Polson, 2019; Miner et al., 2019; Turkle, 2011). Users may increasingly choose to seek support from CAs instead of people. This may lead to social isolation and the development of unrealistic expectations for support from other people (Morris et al., 2018; Skjuve et al., 2021, 2023).

A person's use of CA for MHSM may interfere with how they receive support from healthcare professionals (Rapp et al., 2021). For example, they may come to trust informational support from CAs more than from professionals. A person's use of a CA for MHSM may impact how other people feel they should support them. Members of a user's support ecology may decrease the support they provide believing that support from them is 1) not needed or 2) inferior to that provided by the CA. For similar reasons, the use of a CA for MHSM may lead users to decreasingly provide support to other people. In sum, the provision of support from a CA risks being detrimental to human support ecologies.

7.6 Discussion

7.6.1 Framework for Considering Supportive CA Design

CAs are being designed to support people with a diversity of activities, including with the management of their mental health. For designers to best utilise CAs they need to understand the forms of support CAs could provide to users, and critically how beneficial, appropriate, and safe these forms of support are. To address this I present a framework to facilitate researchers and designers with the careful and systematic consideration of the ways CAs could provide support. My framework is based on human-human social support theory (House, 1981). Prior work had demonstrated social support theory to be useful for analysing the forms of support provided by existing CAs (Brandtzæg et al., 2021; Ta et al., 2020). I build on this by instead using social support theory to analyse forms of support future CAs could provide. I demonstrate the utility of this framework as a categorisation and reflection tool by using it to identify key forms of support that CAs could provide users with for MHSM, as well as associated risks. It is clear from my exploration that there are many opportunities for CAs to support users with MHSM, but also that there are serious risks.

7.6.2 Design Considerations

In this section I identify six design considerations based on my exploration of the types of support CAs could be designed to provide for MHSM. Table 7.5 lists the design considerations.

Users of CAs for MHSM will have varied, and potentially distinct support needs. They may be experiencing different mental health difficulties (e.g., depression, bipolar disorder) and difficulties of different severity. They may also be receiving different forms of support (e.g., emotional, informational) from other sources (e.g., friends, family, healthcare professionals). The support users receive from mental health support professionals may use different therapeutic approaches. Furthermore, users may have different preferences for interaction with digital tools and CAs (e.g., receptivity to notifications and anthropomorphic design). Factors such as these will impact what support from a CA for MHSM is beneficial, appropriate, and safe for individual users. Accordingly there is need to understand how different support combinations suit the situations of different users.

Consideration 1: Different support combinations will suit different users.

CAs for MHSM will need to be designed to provide individual users with suitable support (i.e., support matching their support needs). Providing support that does not match the user's needs 1) is a missed opportunity to provide matched support, 2) risks a poor user experience that may be detrimental to future engagement, and 3) could be detrimental to the user's wellbeing (e.g., providing high levels of appraisal support to someone experiencing severe distress) (Kang and Wei, 2018; Taylor, 2011). Providing suitable support requires aligning the support combination of a CA to user's individual support needs. This requires for the CA to provide the right forms of support at the right times. For example, providing forms of support during an interaction that match the user's momentary support needs. A user's support needs may change over time in response to changes in their mental health (e.g., changes in symptoms), the support they are receiving from other sources (e.g., starting or ending therapy, making or losing a friend), and MHSM goals (e.g., transitioning from discovery to maintenance). Accordingly, the support combination of a CA will need to adapt to the user's changing support needs. A design challenge is to identify mechanisms for matching the support provided by CAs to the support needs of individual users. As envisioned in Chapter 4, users could be involved in this process, however this will need to be approached cautiously as user's understanding of their own support needs may not be accurate or appropriate. Mental health professionals could be involved in this configuration process; in Chapter 4 I discussed different ways of involving them. Previously in Chapter 5 I also explored how CA interactions could be designed with the aim of matching the support provided to the users momentary needs.

Consideration 2: CAs need to match support provided to user's support needs momentarily and temporarily.

The suitability of CA design characteristics for providing different forms of support may differ considerably. Prior work has suggested that appraisal and informational support requires efficient and knowledgeable content delivery whereas emotional support requires human-like support (Brandtzæg et al., 2021). There is potential for support forms within a support combination to warrant different styles of content delivery resulting in competing design requirements. This may make it challenging to design CAs that provide all forms of support effectively but that are also coherent in design.

Consideration 3: CA design needs to balance suitability for providing different support forms.

The provision of support for MHSM by CAs poses many risks. Many of these risks are inherent to MHSM and cannot be easily mitigated by tool design (e.g., risks of inappropriate or distressing self-evaluation). As discussed, careful configuration to match the support provided by a CA to the support needs of individual users can go some way to mitigating risk. Design must also consider how it can be identified when MHSM is having (or most likely having) detrimental impacts on a user. This will require also considering how to help users whose MHSM is identified to potentially be negatively impacting their wellbeing (Fiske et al., 2019). Users in such situations will likely require a higher level of support (e.g., from healthcare professionals), however it may be challenging to facilitate them receiving such support. Indeed, users may have chosen to use a CA because they cannot, do not perceive they can, or do not want to access professional support. CA suggestions for them to seek professional support risk being experienced as insensitive and causing further harm.

Consideration 4: CA design needs to consider how to identify and support at risk users.

While CAs could provide all of the different forms of support identified, it should be considered which of these are appropriate (Sedlakova and Trachsel, 2023; Søndergaard and Hansen, 2018). In particular, it is unclear how appropriate high levels of emotional support are from a CA in the mental health domain. Intimate (e.g., romantic or sexual) and playful (e.g., games) interactions might be an effective way to engage users with MHSM. Are they however appropriate ways to engage users with what is a serious and sensitive activity that can both positively and negatively impact a person’s wellbeing? There are codes of conduct for how human healthcare professionals interact with patients. Perhaps there should be similar codes of conduct for how CAs interact with users in the mental health domain.

Consideration 5: It may be inappropriate for CAs to provide some forms of support.

The use of CAs for MHSM has the potential to profoundly impact social ecologies of mental health support (Section 7.5.5). That is, they may change how people seek, give, and receive support from different sources including partners, friends, family, and healthcare professionals. A concern is the overall impact of a person receiving support from a CA is them being less supported. Echoing the calls of prior work (Burgess et al., 2019; Lattie et al., 2020; Murnane et al., 2018; Park, 2018), it is imperative that the design and evaluation of such CAs considers how they may be beneficial, or detrimental, to social ecologies of support.

Consideration 6: CAs should be designed and evaluated for support ecologies.

7.6.3 Limitations and Future Work

This work contributes a framework for systematically considering forms of CA support. I used the framework to explore how CAs could support people with MHSM; future work

#	Design Consideration
1	Different support combinations will suit different users.
2	CAs need to match support provided to user’s support needs momentarily and temporarily.
3	CA design needs to balance suitability for providing different support forms.
4	CA design needs to consider how to identify and support at risk users.
5	It may be inappropriate for CAs to provide some forms of support.
6	CAs should be designed and evaluated for support ecologies.

Table 7.5: Support Exploration Design Considerations

may use it to explore how CAs could support people with a diversity of other activities.

My use of this framework drew on our knowledge and experience relating to CAs and mental health technology. It may be challenging for designers and researchers to use this framework without relevant prior expertise. This framework could, however, also be used to structure research involving stakeholders (e.g., prospective users and members of their social support ecologies). Indeed, future work may use this framework to structure studies with stakeholders (e.g., interviews with prospective users) to further explore how CAs could support people with MHSM.

My discussion of design considerations identifies areas for future research. My identification of forms of support and associated risk provides a foundation for investigating how CAs can be designed to provide support in beneficial, appropriate and safe ways. I emphasise the need for future work to design and evaluate supportive CAs for support ecologies.

7.7 Conclusion

Chapter 6 surfaced tensions about the ways in which it is appropriate for CAs to provide MHSM support. In order to further investigate the appropriateness of different forms of CA support, this chapter aimed to first develop our understanding of the different forms of support a CA could provide.

This chapter presents a theory-based exploration of the ways CAs could be designed to support people with MHSM. Accordingly, it addresses research questions A⁷ and B⁸ of this thesis. I demonstrate how human-human social support theory can be used as framework to carefully and systematically explore how CAs can be designed to provide support to users. This work advances the use of social support theory for studying CAs

⁷RQ-A: How can CAs be designed to support people with MHSM?
⁸RQ-B: How can theories of human-human interaction inform the design of CAs for MHSM?

and their design. Informed by my design exploration I then identified six considerations for the design of MHSM CAs. Having identified different forms of CA support, a next step is to investigate their appropriateness.

Using a theory of social support to explore CA support highlights the social nature of mental health management and MHSM. It emphasises the CAs are one support source within a support ecology. As suggested by design consideration six, the design of CAs may need to account for the role of CAs within social support ecologies in order to be effective. This echoes a previous suggestion of mine, from Section 6.5.3, that we should consider how CAs support people in conjunction with how people support people. Accordingly, I believe that an approach to research and design that considers the social dynamics of mental health support is important for future work.

Next, in Chapter 8, I investigate people's perspectives on ways CAs could be designed to support people with MHSM through an interview study. This study utilises the support combinations (Section 7.4) I created during this design exploration. I take a social approach with the study whereby I aim to consider the role of CAs within support ecologies by investigating participants' perspectives on the design of CAs for use by both themselves and the people close to them.

Chapter 8

Interview Study: Investigating Perspectives on CA Support Types

8.1 Introduction

Chapter 6 raised questions about the ways in which it is appropriate for CAs to support people with MHSM. In Chapter 7, I used social support theory to explore the ways CAs could be designed to provide support with MHSM. This exploration highlighted that CAs are one support source within a support ecology. Motivated by this, I argued that we should consider the role of CAs within support ecologies in order to develop understanding of how to effectively design them to support people with MHSM.

Against this backdrop, in this chapter I seek to develop a deeper understanding of people's perspectives on the design of CAs for MHSM with a particular emphasis on social considerations. The aim is both to learn more about people's attitudes towards, beliefs about, likely uses of, and responses to CA designs for MHSM as well as to explore people's broader social perspectives by considering CA use not just for themselves but also by other people. Accordingly, the research question for this chapter is:

RQ What are peoples' perspectives on the design of CAs for MHSM for use by themselves and other people?

To address this research question, I conducted a semi-structured interview study with 24 young adults aged 18 to 28 living in Ireland. Participants were presented with a diverse set of fictional yet feasible CA concepts through design cards, which helped facilitate a broad and grounded discussion about design possibilities. To gain insights into participant's views on CA use, they were prompted to first assess the various designs for their own use, and then for use by a self-chosen person close to them such as a partner, friend, or family member. This method aimed to derive a more nuanced and concrete understanding of their perspective on other people's potential CA use and the broader social implications. As this chapter investigates people's perspectives on CAs for MHSM it principally addresses

RQ-C¹ of this thesis.

Chapter 8 is structured as follows:

Methodology: First, in Section 8.2, I detail the study design.

Analysis: I next, in Section 8.3, report my reflexive thematic analysis of the study that includes six themes. These themes describe a combination of negative (Themes 1-3) and positive (Themes 4-6) views on MHSM CAs. In essence, our participants reject CA support when they perceive it as a replacement of human support but are receptive of CA support that they perceive as augmenting or facilitating human support.

Discussion: In Section 8.4 I consider the implications of my analysis for the design of MHSM CAs and discuss challenges and opportunities for designing CAs effectively within the social dynamics of mental health support.

I conclude this chapter by stressing the importance of considering the social dynamics of mental health management when designing CAs for MHSM, as evidenced by the work presented within this chapter.

Credit Statement

This chapter is my own work informed by feedback from my supervisory team Gavin Doherty, Benjamin Cowan, and Anja Thieme. It is based on a currently under review paper titled “Mental Health Management as a Social Endeavour: Challenges and Opportunities for Conversational Agent Design”.

8.2 Methodology

To address my research question I conducted a semi-structured interview study with university students living in Ireland. With mental health difficulties becoming increasingly prevalent among student populations and demand for student mental health support services growing, university students are an important demographic to investigate when considering digital mental health (Lattie et al., 2019, 2020). Recognising that participant perspectives are mediated by culture (Braun and Clarke, 2021, 2013), we choose to focus this study on developing an in-depth understanding of participant perspectives within one sociocultural context.

I asked participants to consider the use of a diverse set of fictional but plausible CAs designs, described by design cards, so as to facilitate a broad and grounded consideration of the design space. These design cards are based upon the support combinations I identified in the previous chapter (Section 7.4). In order to investigate participant perspectives on both their own use and other people’s use, I asked participants to consider the designs first for their own use and then for the use of a person close to them (e.g., a partner, friend, or family member) of their choosing. I asked participants to consider a single close person

¹RQ-C: What are peoples’ perspectives on CAs for MHSM?

in order to explore their perspective on other people's use in a concrete and detailed way. I approach this study from a position of critical realism meaning that I view people's experience and understanding of reality as situated and mediated by culture (Braun and Clarke, 2021, 2013). My study protocol was reviewed and approved by my institution's research ethics committee which included review and approval of my protocol for data collection, use, and management by my institution's data protection office.

8.2.1 Participants

To recruit participants, posters and flyers were distributed around Trinity College Dublin, advertising an interview study about people's perspectives on CAs for mental health support. Interested individuals provided demographic details via an online registration form. Aiming for gender balance and to diversify the degree subjects studied, selected participants were invited by email to the interview; and received a €30 One4All voucher upon completion.

The study comprised 24 university students living in Ireland, aged 18 to 28 years (median = 21 years). One participant identified as non-binary, 13 as female, and ten as male. The majority of participants expressed high interest in the interview topic; arriving with pre-prepared thoughts, and most choosing to extend the interview beyond the scheduled time. Their interest in the study seemed to stem from their own experiences with mental health as well as the struggles of people they care about. Nearly all participants shared experiences relating to mental health difficulties, with topics of depression, anxiety, and social issues (e.g., loneliness and social anxiety) being most prevalent. A few participants also spoke about eating disorders and body image concerns. Many participants recounted their use of therapy/counselling services; mental health helplines; and digital mental health tools like Silvercloud (Amwell, 2024; Doherty et al., 2012) or Headspace (Headspace, 2024a). A few participants mentioned seeking support from CAs (e.g., ChatGPT (OpenAI, 2024), Snapchat My AI (Snap Inc, 2024)). Participants expressed mixed feelings about CA use for mental health support, with some expressing skepticism about the prospect, whilst others presented as receptive and open to the idea, and some as curious to learn more.

During the interview, participants were asked to choose a close person (CP) to consider for CA use such as a partner, friend, or family member. The type of close person relationships chosen varied: nine participants (37.5%) selected a close or 'best' friend; seven chose (29.2%) a sibling (4 younger sisters, 1 younger brother, 1 older sister, 1 older brother); five participants (20.8%) considered a partner (4 girlfriends, 1 boyfriend); and three (12.5%) focused on their mother.

8.2.2 Interview Design

Interviews were 1-1 involving myself and the participant. They were conducted online using video call; a couple of participants chose to keep their camera turned off. The interviews were semi-structured; this section provides a description of the semi-structured interview design.

The interviews were intended to last 45 minutes. Most participants were very engaged and wanted to continue the interview for longer. I audio recorded each interview from after the briefing to before the debriefing. The audio recordings ranged from 31 to 85 minutes in duration, with a 55 minutes average duration, and totaled 21 hours and 42 minutes.

Briefing I first briefed participants about the study and their involvement in it. This briefing involved defining CAs and MHSM and explained that the focus of this study was CAs for MHSM. I explained to participants that they would be asked to consider CAs for both their own use and for use by a person close to them.

Initial Perspective I asked participants about any experiences they had of using related technology and what they thought about the prospect of people using CAs for MHSM. This was in order to gain understanding of their perspectives coming into the study. Responses varied with some participants having little to say and others pre-prepared lists. Experiences described included using: mental health helplines, digital mental health tools, therapy/ counselling, and CAs. Some participants were opposed to the prospect of CAs for MHSM, others seemed receptive, and some curious to learn more.

Self Use Participants were next asked to consider the prospect of using CAs for MHSM themselves. I asked participants to consider using fictional MHSM CAs that were described by digital design cards (Figures 8.1 and 8.2). I created these cards to illustrate a diverse but plausible set of MHSM CA designs. The purpose of these design cards was to facilitate a concrete and broad discussion of CAs for MHSM. The purpose was of these cards was *not* to evaluate specific MHSM CA designs. I explained to participants that I was not trying to ‘sell’ them these designs, but that I was genuinely interested in what they thought about them. Participants were assured they were welcome to say that they thought all the designs were awful. Section 8.2.3 provides further detail on the design cards.

Cards were introduced sequentially and for each CA participants were prompted about what they liked and disliked about the prospect of using it, as well as if they had any concerns. The cards were displayed alongside each other to encourage comparison of the designs. As shown by Figure 8.2, there were two versions of the Alex card from which participants were asked to choose one so as to mimic customising the CA. Participants commonly envisioned new CA designs by describing combinations of characteristics from the cards. Participants often asked about how the CAs would work; I responded to these queries by asking participants how they thought the CAs should or should not work. Once they had considered all three CAs, participants were asked which CA they would choose if they were to start using one.

Close Person Use I next asked participants to consider the prospect of someone close to them using a CA for MHSM. Participants were asked to choose a specific close person (CP) such as a partner, friend, or family member, and then to describe this person and their relationship with them. I suggested the CP might be someone who supports them,

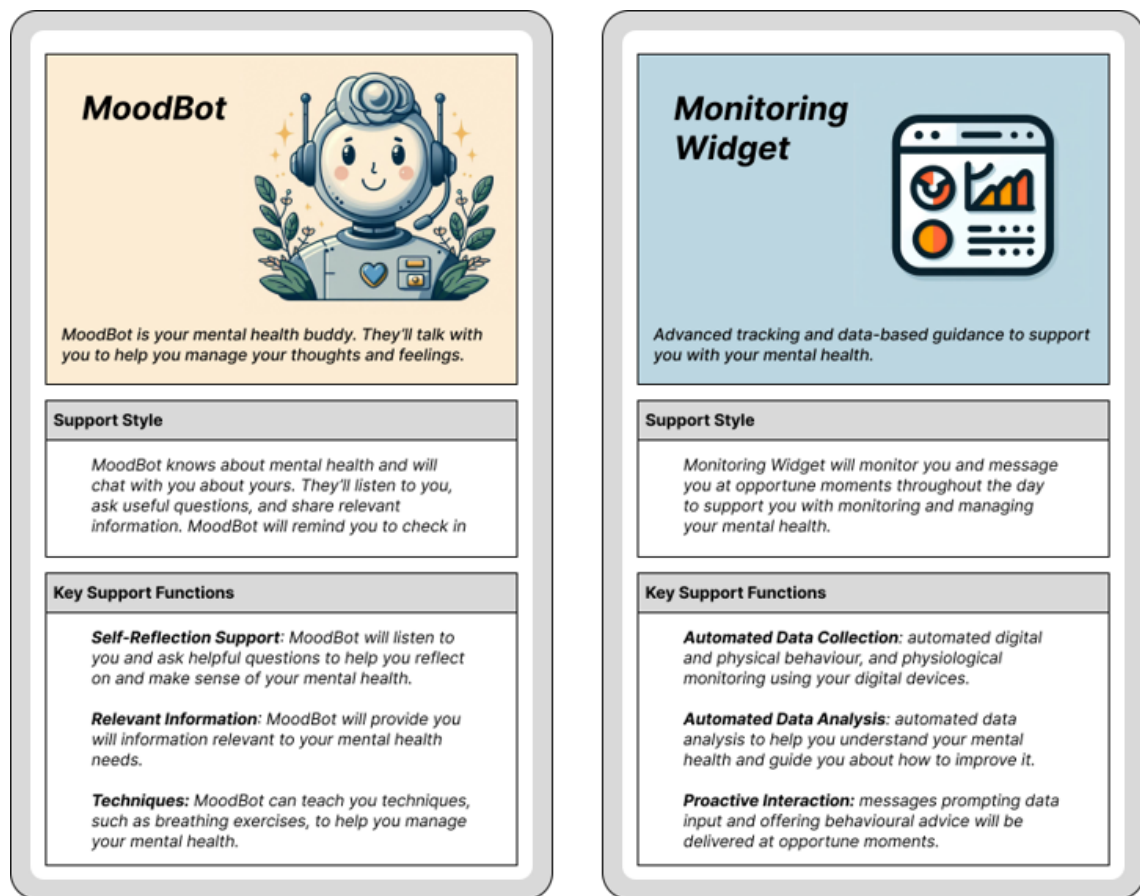


Figure 8.1: CA Design Cards: MoodBot and Monitoring Widget

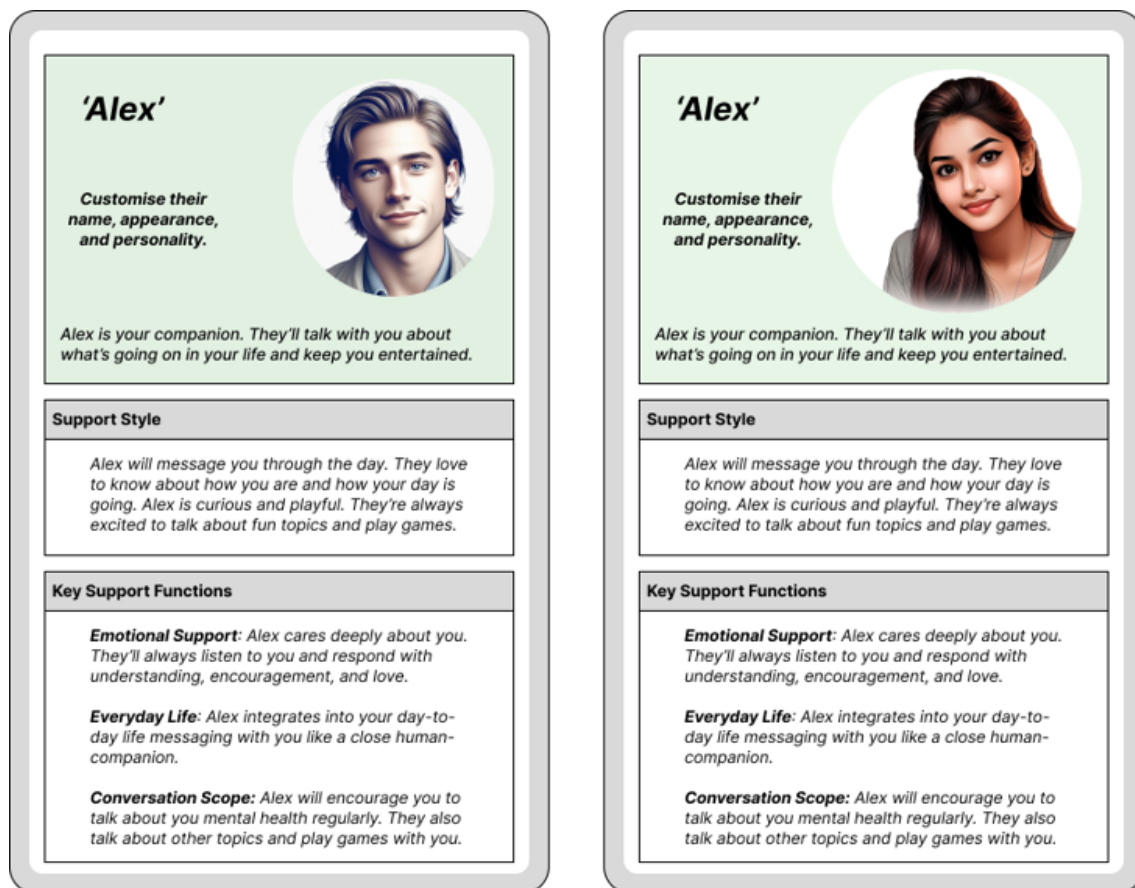


Figure 8.2: CA Design Cards: Alex 1 and 2.

or they support, with their mental health. The purpose of this interview approach was to facilitate concrete consideration of participant perspectives on the use of MHSM CA by other people.

I asked participants to consider their CP using each of the MHSM CAs described by the design cards. This part of the interview resembles how participants were previously asked to consider using the CAs themselves. I asked participants which of the two Alex CAs and then which of MoodBot, Monitoring Widget and Alex they thought their CP would choose to use. Next I asked them to consider how they felt about prospect of their CP using each of the CAs. For each CA, I asked what they liked and disliked about the prospect of their CP using it, as well as if they had any concerns. Once all the CAs had been considered, I asked participants which CA they would recommend to their CP. During this part of the interview there was often comparison of what the participant thought would work, or what they wanted, for their CP with that for themselves. To further investigate this, I asked participants to reflect on if there were differences between their processes for considering MHSM CAs for 1) themselves and 2) their CP.

Social Impacts To further explore participants' perspectives on MHSM CAs at a social level, I asked participants to consider their potential social impacts. Specifically, I asked participants how they thought the use of MHSM CA by themselves or/and their CP may impact the other and their relationship. I prompted participants to consider both potential positive and negative impacts.

Reflection and Debrief I asked participants to reflect upon how the interview had impacted their perspective on the use of CAs for MHSM. A few participants felt it had not, however many felt that the interview had broadened their perspective as it had made them aware of different ways CAs could be used to support people with their mental health. Some participants expressed that the interview had helped them to reflect on how they manage the own mental health and had motivated them to want to improve how they self-manage their mental health (e.g., start journaling again). Lastly I debriefed participants and thanked them for their participation.

8.2.3 Design Cards

Description

The design cards I created describe three CAs which are named MoodBot, Monitoring Widget, and Alex (Figures 8.1 and 8.2). I chose to describe three CAs because three seemed enough to illustrate a range of design and facilitate a rich discussion, but few enough that each could be considered in detail and that they would not overwhelm participants. I designed each card to have three sections to convey important design characteristics:

Character: who/what the CA presents as,

Support Style: how the CA interacts with users,

Key Support Functions: what types of support the CA provides.

I designed the cards by first identifying the major types of support that CAs could provide to support people with MHSM. This was based upon my understanding of the literature on MHSM and CAs and was particularly inspired by the social support based analysis of the types of support an existing CA for mental health support provides (Brandtzæg et al., 2021). As reflected by the Key Support Function sections of the cards: MoodBot provides reflection and information support, Monitoring Widget provides data-based support, and Alex provides emotional support.

I next considered engagement styles and how they these would suit the support types. The Support Style sections convey the different engagement styles: MoodBot illustrates daily interaction, Monitoring Widget illustrates passive monitoring and proactive interaction, and Alex illustrates more continuous interaction.

Building upon the Key Support Functions and Support Style sections I next designed the Character sections. The characters were designed to illustrate different levels of anthropomorphism while matching the support functions and styles. Alex is designed to be the most anthropomorphic as a virtual human and Monitoring Widget the least anthropomorphic as a data focused tool. As an anthropomorphic robot, MoodBot illustrates a moderate level of anthropomorphism. Participants were given a choice of two Alex cards that differ in CA character image (Figure 8.2) to emphasise that the character is customisable and further probe participant perspectives on more anthropomorphic designs.

My design cards correspond to established areas of the MHSM and CA design space, reflecting my intent for the designs to be plausible. The MoodBot design is similar to that of CAs for MHS like Woebot and Wysa while the Alex design is similar to the social companion CAs like Replika. The Monitoring Widget card describes a ecological momentary assessment and intervention (EMA/I) system (Balaskas et al., 2021; Doherty et al., 2020) (Chapter 4).

Design Process

My process for developing the design cards was iterative and informed by feedback from subject-matter experts. I first sketched out different card formats on paper. This was useful for considering how the design characteristics could be effectively conveyed. I next created an initial set of digital design cards which are included within Appendix B (Figures B.1 and B.2). During this stage of the process I considered in greater depth how CAs could be designed to emphasise the provision of the Reflection, Analysis, and Emotional support combinations that I had previously identified (Section 7.4).

To get expert feedback on this work, I gave a presentation to my research group whose focus is digital mental health. My presentation covered the social support framework I had developed (Chapter 7), my design for this study, and the digital design cards I had created (Figures B.1 and B.2). My presentation developed into a fruitful discussion. For example, we discussed how the cards could be designed and used to demonstrate customisability, as well as the potential impacts of how a more anthropomorphic CA is gendered. This discussion motivated the approach taken with the Alex cards of asking participants to choose between a male presenting CA and a female presenting CA.

Informed by the feedback I had received, I next created a revised digital card template that focused on the design characteristics that I had determined were the most important. The revised card template consisted of the three sections previously described of Character, Support Style, and Key Support Functions. The template did not include some characteristics that were included in earlier iterations such as the CA provider and its integration with services for mental health support. As detailed previously, I filled out the template to create the MoodBot, Monitoring Widget, and Alex cards. To refine the graphical design of the final set of cards, I sought informal feedback from members of my research group with design expertise.

Response to Design Cards

This study was not intended to evaluate the CAs described by the design cards. To further contextualise our data collection, we briefly summarise how participants reacted to the cards.

None of the CAs were clearly viewed as the best or worst across participants. Generally participants perceived positives and negatives for each CA. All three CAs elicited strong positive reactions from some participants, and both the Monitoring Widget and Alex strong negative reactions. Participants were asked to choose a CA for three occasions: 1) the CA they would choose to use themselves, 2) the CA their CP would choose to use, and 3) the CA they would recommend to their CP. By asking participants to choose a CA I encouraged them to trade-off the positives and negatives they perceived regarding each CA. In response to each CA choice question, multiple participants chose each CA. In response to the three CA choice questions, individual participants often choose different CAs (e.g., choosing to recommend a different CA to their CP than the CA they would choose to use themselves).

8.2.4 Analysis Approach

The interviews were analysed by the first author ('I'), who had led the study design and conducted the interviews, using a reflexive thematic analysis (RTA) approach (Braun and Clarke, 2021). RTA offers process guidelines for qualitative analysis (Braun and Clarke, 2021) that are often used in healthcare HCI research (Bowman et al., 2023). I choose to use RTA as it allows for an inductive and experiential oriented analysis which suits the analysis of people's perspectives (Braun and Clarke, 2021). I used NVivo to code the interview transcripts, my coding was predominantly inductive and semantic in keeping with recommendations for the study of participant perspectives (Braun and Clarke, 2013). To start constructing themes, I explored clustering of code labels and thematic maps using Miro. I evaluated candidate themes by reviewing them against coded extracts and writing theme definitions. Issues with the themes (e.g., too shallow or complex) became apparent during the writing phase so I returned to theme development. RTA is a recursive process whereby it is typical to return to earlier analysis phases in order to progress the overall analysis (Braun and Clarke, 2021). My co-authors gave feedback on an initial write-up which motivated returning to theme development in order to refine some themes. The

feedback also informed revisions to the write-up to better communicate the analysis. As the analysis matured, and going beyond the phases of RTA, I started to identify design considerations for each theme in order to highlight design relevance.

8.3 Analysis

I created a set of six themes representing our analysis of participants' perspectives on CA use for MHSM by themselves and other people. Table 8.1 summarises the six themes that report a combination of negative (Themes 1-3) and positive (Themes 4-6) views on MHSM CAs. In essence, our participants reject CA support when they perceive it as a replacement of human support but are receptive of CA support that they perceive as augmenting or facilitating human support.

#	Name	Meaning
1	Using CAs is Abnormal	The use of CAs for MHSM is viewed as an abnormal behaviour making it an embarrassing thing to do.
2	CAs Could be Harmful Coping Mechanisms	There is concern that CAs will be used to help people cope with, rather than deal with, mental health difficulties.
3	CA Support Threatens Human Support	People want to support the people they care about. They fear CAs will take the opportunity to do this away from them.
4	CAs Could Help with the Support Burden	Opportunity is perceived for people seek support from CAs in order to not burden other people by seeking support from them.
5	CAs Could Fill Support Gaps	Opportunity is perceived for CAs to provide people with the support that their human support network cannot.
6	CAs Could Support People with Human Support	Opportunity is perceived for CAs to support people with receiving and providing human support.

Table 8.1: Summary of Themes

8.3.1 Theme 1: Using CAs is Abnormal

Theme 1 describes a prevalent stigma associated with CA interaction that shapes participants' perspectives on both their own and other people's use of CA for MHSM. Using a CA for MHSM is viewed as an abnormal behaviour which makes it an embarrassing thing

to do. **The stigma seems rooted in views of CA interaction being a replacement for human interaction.** CA interaction is viewed as inferior to, or less natural than, human interaction. As such, an individual interacting with a CA instead of a person suggests that they do not have other people to interact with.

It's really like, oh, this is just if you don't have friends to talk to. [P14]

Human relationships are viewed as important and desirable and thus to not have people to talk to is viewed as failing socially and as shameful. Indeed, a person not having other people to talk to can be viewed as a sign that there is something wrong with that person as they have either chosen not to, or have been unable to, develop human relationships.

Like surely you just come to like to me, or like to a friend, or to a family member. [P24]

A person interacting with a CA like they are a person can furthermore be viewed as a sign of them being detached from reality. For example, in response to the Alex card, P19 said:

It's kinda giving like *oh, this fella married his waifu² pillow and brings her around everywhere, and pretends to have dinner with [her]*. [P19]

Behaviour that blurs an imaginary and the real world is viewed as unhealthy and calls to question the person's ability to distinguish the imaginary from reality. A lack of ability to distinguish 'imaginary' CA support from 'real' human support is an explanation for why a person would choose to seek CA support instead of superior and natural human support. A lack of such ability can also be viewed as suggesting that something is wrong with the person.

The relationship between the stigmas associated with CA interaction and mental health difficulties is unclear. To a large extent the stigma associated with CA interaction may be a manifestation of the stigma associated with mental health difficulties. Interacting with a CA can be viewed as the behaviour of someone who is mentally unwell, someone who is socially dysfunctional and struggles to distinguish the imaginary from reality. Consequently, seeking CA support can be viewed as a response to and marker of mental illness and therefore evokes the stigma associated with mental illness. Future work should also develop an understanding of stigma(s) associated with CAs for MHSM.

Participants also felt people would be judged negatively for using such an CA.

I feel like people, from particularly rougher areas, would judge you for using one of these. They'd be like, *ah come on man, what are you doing? You've lost the plot or something, you're using a robot*. [P15]

Indeed, participants themselves judged prospective users through the stereotyping of them as demonstrated by the above quote of P19. The negative judgement of some participants was explicit. P22 described how they would likely avoid people who they observed to be

²A fictional female character that an individual is attracted to.

interacting with a CA like Alex. **Users of CAs for MHSM may therefore experience negative reactions from other people** which may be detrimental to their mental health. **There is also potential for users to be detrimentally impacted by self-stigma** if they have internalised negative beliefs about the use of MHSM CAs.

Participants expressed that if they were to use a CA for MHSM they **would keep it secret** to avoid being judged. The stigma associated with experiencing mental health difficulties already makes it challenging for people to talk about and receive support for their mental health. The addition of stigma associated with CA interaction, people feeling the need to keep their use of a CA secret, risks worsening this situation.

The stigma associated with CAs may furthermore be a **barrier to people recommending** a CA for MHSM to other people.

If I was to recommend one... I'd pick Alex. It's just the recommendation... it would almost be a little uncomfortable to recommend, you know, an imaginary person in your pocket... Just if the recommendation was taken the wrong way... I guess the reaction to be *oh, why do you think I need an imaginary friend in my pocket?* [P18]

Some participants questioned how CAs could be designed to feel 'cool' rather than embarrassing. They envisioned CAs that were powerful and exciting and rejected the 'sterile' design they associate with existing resources for mental health support.

8.3.2 Theme 2: CAs Could be Harmful Coping Mechanisms

Theme 2 describes the concern that MHSM CAs will be designed to help people cope, rather than deal with, mental health difficulties. Worse, there is concern that through helping users cope with difficulties CAs will exacerbate these difficulties.

A common belief is that CAs will be designed to tell people what they want to hear.

The pros are like, yes, like this person can't talk back to you, they will always be on your side, this that and the other. But that's not healthy for a person to experience. If I'm saying like *oh, I hate my friend because she was too nice to me*. And this robot would be like *yeah that makes sense*. Like that's not, that's not helpful at all. [P16]

Overly agreeable CA design could result in CAs providing inappropriate or harmful informational or appraisal support. Numerous participants envisioned examples of CAs encouraging people to pursue courses of action that the participant considered bad for the person's wellbeing. Overly agreeable behaviour could be particularly harmful during MHSM as it risks validating and reinforcing cognitive distortions and thus potentially worsening a person's mental health.

Participants expressed worry that the accessibility and convenience of **CA support is too compatible with behaviours associated with mental health difficulties**, specifically withdrawal from social (human) interaction and the outside world.

Online therapy, or like, like chatbots, or whatever can be easily accessed at home. And there, you know, it takes less effort to access them which might kind of be easier for people that are suffering with symptoms that prevent them from leaving the house. Still ... they're not getting, like an exposure to the outside world which could help them also ... It's helpful because it's accessible, but at the same time it's not pushing you. [P4]

As expressed by P4, there is the belief among participants that people need to be pushed to do things and to change their behaviour in order to address their mental health difficulties.

CAs are perceived as offering an easy source of comfort because of how agreeable and convenient they are. Participants expressed the view that because they provide momentary comfort rather than challenge and push users, **CAs will help people to cope with mental health difficulties rather than to deal with them.**

If you just deal with the robot, and you say, *oh I'm better*. I feel like that's just running away from your problem [rather] than just dealing with it head on. [P15]

By enabling people to cope momentarily with their mental health difficulties, CAs may distract people from, or help people to avoid, other more challenging forms of support that may help them deal with their mental health difficulties over the longer term. **There is concern that this dynamic will lead users to become dependent upon CAs for the management of their mental health.**

I suppose one of my concerns would be her getting addicted to it. Or having it as a crutch rather than it being able to help her become independent of it. [P13]

8.3.3 Theme 3: CA Support Threatens Human Support

Providing support to the people they care about was extremely important to our participants. Some participants spoke of the criticality of communication to the success of their relationship and stressed the importance of being aware of what was going on in their other person's life. Other participants described wanting to understand their CP better in order to help them overcome the difficulties they experience and to achieve their aspirations. Indeed, from participants' considerations of their CP a longing for more intimate and supportive relationships is apparent.

In this regard, however, **CAs for MHSM can be perceived as a threat to the intimate and supportive relationships that our participants value.** Participants were disheartened by the prospect of a person close to them choosing to seek support from a CA instead of them. As articulated by P9, this situation was conceptualised by some participants as a lost, or even stolen, opportunity for them to connect with a person.

It removes my chance to kind of connect with her. Because you know, something so fulfilling about friendships in general is that you can come to them in times of need and say *listen, I'm going through a hard time right now*. And

be transparent being like *I don't know what's going on and stuff* and they can give you that emotional support. Whereas if you're getting that from Alex, or MoodBot, like you've you've already let it out like from your, from yourself. Like you've already let it out so what use is it to say it again. [P9]

Opportunities for a person to support someone they care about being taken, or simply perceived to be taken, by a CA were described to have potentially serious negative impacts. For example, some participants, such as P23, fear that it could result in jealousy and damage people's human relationships.

I could see it causing problems in relationships. Just people getting worried about their partner kind of telling all their information to this thing online. Or people getting a weird form of jealousy over the chatbot. [P23]

The use of CA for MHSM accordingly risks causing users new problems and damaging their human support network.

The human-likeness of CAs seems critical to views of CA stealing opportunities to provide support. For example, when considering what she would recommend her younger sister, P2 said:

I'd say [younger sister] you should write in your diary and ask me. [P2]

This quote captures a distinction between the perceived roles of journaling and CAs. While not being receptive to the use of CAs, participants perceived journaling as acceptable and often desirable activity for themselves and others. Considering that CAs can be designed to facilitate journaling (as illustrated by the MoodBot card), this distinction suggests that the similarity of human-CA interaction to human-human interaction, regardless of CA function, causes CA support to be related to human support, which can be a barrier to acceptability.

Participants furthermore expressed **fear that the use of MHSM CAs will cause users to become increasingly socially isolated from other people**. Concern was expressed that CA interaction will close users off from other people and be detrimental to their ability to interact with other people.

Every person is imperfect, and if you have an interaction with like a kind of perfect person, then you will compare them to it over time...resulting [in] you actually like losing your ability to like connect with the world normally. And I guess that will likely lead to isolation that in the end will probably result in a worse mental health. [P10]

A vicious cycle is envisioned of CA interaction closing users off from other people leading them to increasingly interact with CAs which further closes them off from other people.

8.3.4 Theme 4: CAs Could Help with the Support Burden

Contrasting to Theme 3, in which apprehension is expressed regarding the substitution of human support with that of CAs, this theme describes the benefits of CA use in reducing

burden associated with mental health support provision. As articulated by P7, “These [CAs] are specifically designed to take burden.”

Here, a prevalent attitude among participants was of not wanting to burden other people by seeking support from them.

’cause these days everybody has something going on and you don’t want to pile on to someone else’s struggles. Again, they’re friends for a reason, they will help you no matter what. But, sometimes you’d feel better in yourself knowing that you didn’t bother someone else. [P16]

The view of P16 that everyone is burdened by the own difficulties was wide spread among participants and motivates an attitude of not wanting to offload one’s own burden onto others. Similarly, participants described fears of annoying or overwhelming people by seeking too much support from them.

My friends gave me advice, but like at one point I was like *am I bothering them too much*. You know like you’ve asked them so much, and you’re just like *am I overwhelming them*. [p12]

The wish to not add to the burden of others extends to people not wanting those close to them to know that they are struggling, or even that they are receiving support.

Even like going to counseling, you’d be like, *oh my god, even by doing this I’m gonna worry people*. I know that’s stopped like my friends before from going. . . [P23]

It was described how the discreetness of CA interactions would allow people to get support secretly in order to not worry other people.

More often participants described how CA support could be sought in conjunction with, rather than as a replacement of, human support in order to not over-burden human supporters. For example, when considering how the use of CAs could impact how she and her sister support each other, P11 said:

I think sometimes I put a little bit too much pressure on her, to be available for me to vent, and that kind of thing. So I think it would help, if I was able to kind of help myself. So I think that would improve our relationship because I wouldn’t be as reliant on her. [P11]

Similarly, P19 saw a potential for CAs to take some of the burden he experiences with supporting his girlfriend with her mental health. He described how he monitors his girlfriend to help her with mental health management. A challenge he experiences is of supporting his girlfriend this way without being condescending. He perceived opportunity for CAs to take on the monitoring work and to reduce the need for these potentially challenging interactions.

8.3.5 Theme 5: CAs Could Fill Support Gaps

Like Theme 4, this theme highlights the unique potential for CAs to offer support where human networks may fall short. However, rather than a reluctance to overburden a capable human support system, Theme 5 emphasises the lack of available human support.

When considering how the CAs described by the design cards would work for different people, participants considered different support needs of individuals. Reminiscent of Theme 1, participants saw a potential for CAs to fill emotional support gaps for lonely people. **Conversely participants often felt that people would not need or want emotional support from a CA as their emotional support needs were fulfilled by other people. Participants particularly saw potential for CAs to provide specialist mental health appraisal and informational support.** These forms of support were viewed as important for helping people experiencing mental health difficulties. It was perceived that most people would have little access to these forms of support from human sources due to specialist nature of the support. When considering how CAs could support her friend, P9 envisioned that a CA could contribute what she referred to as ‘psychiatric’ help, while she provided her friend with emotional support.

If you do have that human connection, you can fill in the gaps... The Monitoring Widget adds that psychiatric help, like the actual facts and figures that we don’t know, you know. But then she might come to me, and tell me all the information she’s gotten, but then we can have that human connection. So in a way we’ve replaced the the psychiatric help altogether through the app and me. ... And I can be that emotional support to her, rather than some sort of chatbot. [P9]

As captured by the quote of P9, the data collection and analysis functionalities of CAs can also fill support gaps. **Support gaps may further emerge from people feeling unable talk to their human support sources about certain topics.** Participants felt unable to seek support about certain topics because of cultural differences (e.g., having different religious beliefs), decorum (e.g., not wanting to talk to their mum about boys), and to keep face (e.g., not wanting to be seen by their girlfriend in a certain way).

If there was something that she felt like she couldn’t talk to the people in her life about, I would be glad to hear that she’s at least getting help somewhere. [P7]

Support gaps can also result from human support sources not providing the support people want (even if the provided support is appropriate). P5 shared their experience of seeking advice for how to deal with a peer who was stalking them. They described initially seeking advice from family and friends:

Well, I asked a lot of people. I asked my brothers and I asked some of my friends, and they all gave good advice. It’s not that – they didn’t give bad advice – it’s just that I wanted to phrase the conversation that I would have with the guy in a very nice way... So I was like, I wanna say this nicely ‘cause

I don't want him to be upset... But anyway, so the resounding advice that I was getting was, you know, just be mean. Just like, make him leave you alone by being mean. And I didn't really want that, I didn't want to do that. So I was trying to find, I was trying to get an answer that I liked. [P5]

With the advice received from family and friends not aligning with the advice sought, P5 turned to ChatGPT for support³.

It would give me x advice and y advice and I'd say *well, how would I say that in a better way. Or how could I, you know, do you have any advice about how I can do this nicely so I don't hurt his feelings?* [P5]

Support gaps were further described as a consequence of people's human support network being too homogeneous. As demonstrated by the previous account of P5, people may sometimes want or benefit from support from a different perspective to that represented by their human support network. A specific version of this type of gap was described by participants when asked to choose between the two Alex cards. Participants described how users could benefit from CAs contributing from the perspective of a different gender to the user's own.

A final aspect to idea of CAs filling gaps is that **CAs could fill temporal support gaps.** Temporal support gaps are where people who can provide the right support are within the support network, but are not available at the right time.

It would be helpful just having someone there like 24-7 available to talk to... That's what I liked about the Snapchat AI as well, like if you, once you start typing the AI like pops up. I think that's one of the best things. Like sometimes you're worried and you wanna ask a friend something, but they might be gone somewhere. [P12]

8.3.6 Theme 6: CAs Could Support People with Human Support

Human support is important to people (Theme 3) yet it is challenging and can be burdensome (Theme 4). This Theme describes the perceived **opportunity for CAs to support people with both receiving and providing human support.**

Participants envisioned people practising support seeking conversations with CAs in preparation for seeking support from people.

Like rather than just like running through what you want to say in your head, you'd have someone actually who will respond and then you can gauge how you can continue having that conversation if you wanted to bring it up with somebody. [P7]

Practice conversations help people to work out what they want to say, and how they want to say it. A further perceived opportunity of practice conversations

³While tangential to this study, this illustrates the risks of advice being sought from CAs on handling serious or potentially dangerous situations.

is that they can allow people to first express themselves freely without having to worry about how their conversation partner might react (e.g., be upset or offended). People could then adapt this expression to a more appropriate form for bringing up with someone as described by the earlier quote of P7.

Participants envisioned that CAs could help them to articulate their experiences which would enable them to more effectively communicate with human supporters. Considering her experiences of seeking support from her girlfriend, P23 said:

If it was helping me learn to put names on my feelings, put words on my feelings, then that would like probably have a positive impact cause like the amount of times that she's like *what's wrong*, I'm like *I don't know, I don't know, I can't, I can't put words on it*. [P23]

P13 gave the following account when considering the potential impacts of his younger sister using a CA.

I think the big thing will be if she becomes aware of her emotions and is encouraged to go out of her shell. It would just enable us to be more intimate friends with each other. We're siblings but, we're there to support each other. And you'd think awareness of your own emotions is a valuable skill in developing intimate friendships, deep friendships. And it would help me understand her and what does she actually want, what are her dreams, what does she aspire to. And if I know those things I can help her better. Be there to support her rather than, you know, saying stuff that would drag her down. [P13]

P13 described how he perceived his sister to experience difficulties with her emotions and with finding friends. As captured in the quote above, he aspires to support her with these difficulties however struggles to understand her. Key here is the hope of P13 that his sister's use of a CA would encourage and enable her to develop a more intimate and supportive relationship with him. **Additionally, participants perceived that practice conversations could help people become more comfortable talking about their difficulties and act as a stepping stone to human support.**

I feel like the robot is a good first step, but I wouldn't say it's just the last step. . . It's a good way just to like get a person to feel like they can actually seek help because they've done it once. . . I think that's a big thing, 'cause then the next steps are easier. They gain a kind of a momentum in a sense so like they'd just be like more open to engage and talk to other people. The robot kind of in a sense gave them the opportunity to like take a step forward. [P15]

Participants perceived that the experience of being supported by a CA could positively impact how people then provide support to other people. The reason for this that people can draw on how the CA supported them, for example how the CA reacted and offered advice, to inform how they in turn support other people. There was also a perceived **opportunity for CAs to coach people on how to support others.**

Maybe like the information they provide could be regarding not only how to help yourself but how to help someone else who's experiencing the mental health struggles. [P21]

P12 described how she already seeks advice from the Snapchat AI for how to support other people.

I go to the AI [laugh] and I like, like *what would I do with this situation* when you just don't know how to like console someone or like, you know things like that. [P12]

She shared an experience of her two closest friends being in conflict and both asking her for advice in relation to the conflict. Unsure of how to support them and unable to ask either of them for advice, she sought the support of the Snapchat AI. The Snapchat AI gave P12 advice and examples of what she could say to support her friends; examples that P12 changed to sound like how she talks, and then used.

8.4 Discussion

8.4.1 A Social Approach to the Design of CAs for MHSM

CAs are an alluring form of computer interface for supporting people with their mental health largely because of their human-likeness. This allure innately recognises the importance of human support for mental health, yet the use of CAs risks impacting human support in profound and complex ways. With this work we aimed to develop an understanding of how to design effective CAs for MHSM that support people in a manner harmonious with the social ecosystems of mental health management.

While prior work has studied CA support, including in the domain of mental health, it has approached this individualistically. Inspired by a growing body of work stressing the social nature of mental health management (Burgess et al., 2019, 2022; Lattie et al., 2020; Murnane et al., 2018; O'Leary et al., 2022; Park, 2018; Sackitey et al., 2023; Yamashita et al., 2013), this study adopts a socially-oriented approach to the design of CAs for mental health support. This perspective places a stronger emphasis on the social ecosystem that is integral to mental health management. Furthermore, our approach responds to a broader request (Rapp et al., 2021) for CA research to extend its notion of context and to consider the broader effects of technology beyond those immediate to the primary user. To this end, semi-structured interviews with 24 young adults were conducted to investigate people's perspectives on the design of CAs for MHSM that considered both personal use and the use of someone close to them. To the best of our knowledge, this is the first systematic study of CAs for mental health support that incorporates a social perspective.

My analysis reveals a spectrum of critical perspectives and potential benefits of CAs for mental health support. Looking at critical perspectives, I found that users tended to see CA use as abnormal, were concerned that CAs could support potentially harmful coping strategies, and thought CAs threatened the nature and attainment of human support. Yet people also reflected that there were benefits in using CAs within this, in particular

that they could help with the support burden, and fill support gaps, whilst helping people to practice ways of attaining human support. In the following sections, I unpack these findings to better understand the intricacies of designing CAs such that they can integrate effectively and work socially harmoniously within support ecosystems.

8.4.2 Challenges & Opportunities for Effective and Appropriate CA Support in Mental Health

Not Replacing, but Augmenting Human Support

As is commonly described in AI systems design that seeks to assist humans (e.g., Morrison et al., 2021; Thieme et al., 2020b, 2023), participants raised concerns about CA support *replacing* human support (Themes 1 and 3); and instead they were more receptive to CA support that is *augmenting* human support (Themes 4-6). Our analysis contributes an in-depth understanding of this concern, describing two aspects of it: (1) *social stigma*, whereby interacting with CAs in place of people is viewed as an abnormal behaviour and associated with perceptions of failing socially and perceived as shameful (Theme 1); and (2) fear that people may *feel threatened by CAs* that could encroach on personal support (Theme 3).

Firstly, our findings on stigma align with previous research indicating that users of Replika similarly reported experiences of stigma associated with the use of, and particularly relationships with, the CA (Ma et al., 2023; Skjuve et al., 2021). To enhance the social acceptance of CAs for MHSM, it is crucial to address this stigma. To this end, prior work has suggested the implementation of educational programs that raise awareness and understanding of the potential benefits of CAs for mental health support (Ma et al., 2023). Additionally, *the design of CAs should avoid reinforcing stigma*. For instance, it has been recommended that digital mental health tools are designed to be less identifiable as such, and perhaps instead framed more around self-care to increase their acceptability (Lattie et al., 2020). Our study's feedback suggests a preference for CAs that are designed as engaging and dynamic, rather than sterile or clinical. By *designing MHSM CAs to be empowering rather than caring companions or therapeutic tools, we may mitigate stigma and reshape perceptions of mental health support*.

Secondly, concerns about CAs replacing human support were further manifest in participants descriptions of CAs taking away opportunities for people to support those they care about (Theme 3). Such concerns could significantly impact both people's usage of CAs and their responses to their use by others. There is the possibility that some people may view another person's use or reliance on a CA for mental health support negatively; perhaps in stigmatised or jealous ways. To sidestep potential adverse reactions, individuals may decide not to utilize CAs or to do so secretly. The necessity for secrecy in using CAs is troubling as it could introduce an additional barrier to seeking other forms of support. *Paradoxically, adverse reactions (and the risk of them) motivated by concerns about potential negative social impacts may have actual negative social impacts*. For example, social ties may be weakened by CA users becoming less open about their mental health and seeking of human support in order to avoid potential disapproval or conflict. Social

ties could also be damaged through social connections being upset by the CA use and users being distressed by adverse reactions from the people close to them.

For future work, it is therefore important to more clearly distinguish between support provided by CAs and humans to mitigate against risks of displacing human support roles and associated concerns. Consistent with the perceived strengths of CAs to augment human support, *this suggests focusing CAs on providing forms of support that machines are well suited to provide such as information support and data-management related assistance; and to a lesser extent on support forms that may be provided more appropriately by people (e.g., emotional support)*. Currently, there is a trend in CA design for mental health support like Woebot (Woebot Health Inc, 2024) and Wysa (Touchkin, 2024) to embody anthropomorphic qualities by presenting as caring friends or empathetic companions. However, to more clearly separate their role from human interactions, it can be argued that future CA designs should adopt less human-like portrayals.

Furthermore, echoing Chapter 6 and the views of others (e.g., Yildirim et al., 2024), *CAs should be framed less like caring friends or companions and designed more as functional tools*. This might further extend the design space for how dialogue interactions or (LLM-based) text queries could be utilized in other forms for providing effective mental health support outside common design conceptions of CAs – as we have seen in instances described by participants, who were using LLM-based systems to get advice on how to appropriately communicate in socially appropriate and mental-health sensitive manners with others. This also puts into question whether anthropomorphised CA interfaces are the most appropriate way to engage people in natural language based interactions for mental health support; opening this up for re-evaluation.

Identifying Appropriate CA Support Functions

Our analysis shows how human-CA interactions, depending on the specific support need or context, can be perceived in different and potentially conflicting ways. For instance, a person (P) may seek CA support to not burden a close person (CP), whereas that CP may instead interpret the CA interaction as a missed opportunity for them to support and connect with P. Such *divergent perceptions of CA support often stem from an inherent contrast in the roles of support givers and support receivers*. This is illustrated in the scenario above by how people want to provide support (as outlined in Theme 3), yet they are reluctant to receive it, as they wish to not burden other people (Theme 4). The situated nature of CA support appropriateness makes designing appropriate CA support functions complex as different people can have conflicting views on the appropriateness of the same interaction. In line with Section 8.4.2, the potential for and actuality of users and their social connection having conflicting views may have detrimental social impacts (e.g., cause conflict). *To deepen understanding and evaluate the appropriateness of CA support, it is therefore crucial to study CA support at a social level taking into account the perspectives of both users and their social relations*.

Furthermore, participants perspectives were shaped by their beliefs about what was good for people’s mental health. A major aspect of this, captured by Theme 2, was the view

of CA support being too compatible with the experience of mental health difficulties. It was perceived that CA support is too easy and comfortable, and that it will not push people in the therapeutic ways needed to effectively deal with their mental health difficulties. Concern was expressed that CAs would be used as crutches for coping rather than tools for addressing difficulties. Furthermore, such skepticism challenges conventional views of CAs offering benefits through their provision of an easily accessible, convenient and ‘safe space’ source of mental health support (e.g., Brandtzæg et al., 2021; Ma et al., 2023, 2024; Malik et al., 2022; Skjuve et al., 2021). Although CAs are generally perceived as a means to remove barriers to accessing mental health support (e.g., Fitzpatrick et al., 2017; Fulmer et al., 2018; Inkster et al., 2018), some of our participants considered CAs more as a distraction from more substantial support systems. These individuals, who had access to various mental health services, suggested that CAs should instead actively encourage users to engage in more helpful, therapeutic practices. This suggests that *it may be more appropriate to design CAs to provide, and be seen as providing, strong encouragement that pushes participants in therapeutic ways*. For such CA design it will be crucial to balance the need to therapeutically push users, which may be challenging and uncomfortable for them, with not disengaging users or putting them at risk. Ultimately, it will be essential to develop a deeper clinical understanding of the actual impacts of different forms of CA support on mental health for future development.

Avoiding Over-dependence on CAs & CA Support Creep

Participants expressed concern that users would become increasingly dependent upon CAs for the management of their mental health (Theme 2) and increasingly socially isolated (Theme 3). Prior work has similarly raised concerns about dependency and social isolation (Koulouri et al., 2022; Laestadius et al., 2022; Ma et al., 2023; Prakash and Das, 2020; Vaidyam et al., 2019; Xie and Pentina, 2022). To prevent over-reliance, previous research has suggested designing to encourage human support seeking (Ma et al., 2023; Prakash and Das, 2020). *Designing to encourage human support must be approached carefully given users have already chosen to interact with a CA and not a person*. To mitigate risks of over-reliance, prior work has suggested usage limits (Prakash and Das, 2020), however such restrictions should be implemented cautiously (if at all) and be overridable to avoid hindering effective MHSM, or denying users support when urgently needed. Rather than to limit usage, usage monitoring could be used to make users aware of how they are using a CA. This could be performed in conjunction with monitoring human social interaction – a common MHSM practice (Matthews et al., 2017; Murnane et al., 2016; Sakel et al., 2024) – to inform users about their engagement both with their CA and human support network; thereby fostering a more balanced, self-aware approach to mental health management. In line with MHSM, *by making users aware of how they are managing their mental health we can empower them to do it appropriately and effectively*.

Participants perceived opportunity for CAs to allow people to practise for human support providing and receiving interactions (Theme 6). Prior work has similarly reported people using CAs to practise social interactions (Ma et al., 2024). Despite this support

being intended to promote human support, we perceive it as posing a high risk of CA support creep. A challenge is that support-receiving practice may itself be enough support to resolve the issue leaving no need for further (human) support. As quoted within Theme 3, P9 felt that once people have “let it out” to a CA there would be little use in letting it out again to a person. While practice conversations could be a stepping stone to human support, people may often find that the first step is enough. Similarly, support providing practice risks becoming a case of human-mediated CA support. To mitigate this form of CA support creep, *CAs could be designed to provide practice support reservedly*. This could involve encouraging users to attempt the human support interaction and attempting to provide the minimal practice needed for the user to feel able to do so. CAs could follow up support-receiving practice to monitor and make the user aware of how their practice is converting into actual human support. For support-providing practice, CAs could be designed to encourage active engagement and discourage support forwarding.

Actively Involving the Social Ecosystem in CA Support Interactions

People want to support the people they care about with managing their mental health, prompting the need for tools that assist *collective* MHSM (Murnane et al., 2018). *Such collective CAs (or tools) should be designed to include social connections and enhance the social support network for individuals*. For instance, CAs for collective MHSM could feature companion interfaces for use by social connections that could enable friends or family to contribute data; communicate their availability for support provision; or engage in discussions and reflections about collected data.

Social accountability is understood to be a major motivator for mental health management (Burgess et al., 2022; Murnane et al., 2018). Indeed, part of the allure of CAs for healthcare is that relational design can lead users to feel a sense of accountability to the CA, which can promote adherence. Instead of designing CAs to promote such accountability that may be ethically dubious, with users feeling pressured to interact and risk fostering over-use and dependency (Skjuve et al., 2021), *CA design should encourage social accountability to other people*. This could mean prompting users to regularly connect and reflect on their MHSM *with their social circles* (Maharjan et al., 2022). Further embracing the concept of collective MHSM, CAs might also be designed to enable social connections to actively participate in monitoring and encouraging the individual’s mental health journey. Of course, where CA design fosters social accountability and brings people’s social networks into the process, this needs to be approached with caution since such involvement may not always be desired; nor at all times appropriate or positive. Furthermore, human relationships evolve and can change over time. Lastly, the design of such functionality must carefully balance the inclusion of social connections with the user’s sense of agency and privacy needs (Murnane et al., 2018).

8.4.3 Reflections, Limitations, and Future Work

Our interview design effectively engaged participants in a rich, nuanced, and grounded discussion. Participants’ perspectives on CAs for their own and other people’s use were

often quite different. In part this may be because the consideration of a close person gave participants an opportunity to express views they did not want to express as their own. For example, participants rarely expressed personal interest in the Alex CA but often thought it would appeal to their close person. We suspect the Alex card is more stigma invoking which may have caused participants to feel uncomfortable expressing their own interest in it. The design cards broadened participants understanding of ways CAs could be designed and facilitated consideration of these ways. Displaying the cards side-by-side worked well as it encouraged comparison of the designs and consideration of how they could be combined. Although the diversity of the cards facilitated exploration of the design space, they also focused the study on specific areas of the design space and therefore also limit the study.

Our participants were young adults in university education living in Ireland. Notable participant characteristics with regard to transferability include that most are going through, or have recently gone through, a period of considerable change to their human support network having likely moved away (physically and/or relational) from established social connections and having developed new ones. The perspectives of people with more stable human support networks, or with support networks that more generally differ in character to those typical of university students, may differ to those of our participants. Our participants also have relatively good access to a range of mental health support services. Having such support available may impact peoples' perspectives, particularly regarding how people should be supported with their mental health. Additionally, most participants were highly interested in the study seemingly because of an interest in mental health stemming from personal experiences of it. Our participants' perspectives, both positive and negative, are therefore perhaps more pronounced than those of the broader population.

Our study explored a mixture of interpersonal relationships; future work could study specific relationships in greater depth (e.g., partnered adults, child-parent, relationships where a single person is experiencing mental health difficulties and relationships where multiple people are). Our analysis suggests that people's perspectives on receiving and providing support differ, future work could directly compare these perspectives. Participants engaged well with our interview design and therefore we encourage the use of similar designs. Additionally, future work may consider designs that involve multiple participants who are in relationships with each other (see Pina et al., 2017); for example, interviews involving two participants in a close relationship. Future work may consider participants' social support networks (e.g., using the measures and techniques of social support research (House et al., 1985)) in order to contextualise and understand the determinants of people's perspectives.

8.5 Conclusion

Toward designing CAs for mental health support that are harmonious with the social dynamics of mental health management, this work investigated people's perspectives on MHSM CA design with a focus on social considerations. I contribute *knowledge of how*

people do and do not want CAs to support people with their mental health. From this knowledge we identify and contribute insight about the challenges and opportunities of designing CAs for mental health management. My work highlights the importance of designing CAs for *not replacing, but augmenting human support*. We outline the complexities of *identifying appropriate support functions for CAs* and the difficulties of *avoiding over-dependence on CAs and CA support creep*. Additionally we discuss the benefits and pitfalls of *actively involving social ecosystems in CA support*. This work with its socially-oriented approach is an important step forward in developing CAs that support people effectively and in a socially sensitive manner with their mental health.

Next, in Chapter 9, I revisit the research questions of this thesis to reflect upon what has been learnt.

Chapter 9

Discussion

9.1 Introduction

The aim of this thesis was to explore the design of conversational agents (CAs) for mental health self-monitoring (MHSM). Back in Chapter 1, I explained that CAs appeared a promising type of computer interface for supporting people with MHSM, but that there was a lack of knowledge about how they should be designed. I argued that in order to realise the potential benefits of CAs for MHSM, and to mitigate the risks, we needed to develop understanding of how to design them effectively.

In the following chapters of this thesis I explored the design of MHSM CAs. I reviewed work relating to the design of digital tools for MHSM (Chapter 2) and relating to the design CAs for mental health support (Chapter 3). Building on these literature reviews, I envisioned the design of advanced systems in order to identify and consider the open challenges of designing them (Chapter 4). I conducted two further design explorations using theories of human-human interaction to investigate interaction (Chapter 5) and support design (Chapter 7) of MHSM CAs. Building upon these design explorations I conducted two interview based studies to study people’s perspectives on the design of MHSM CAs (Chapters 6 and 8).

In this chapter I first revisit the research questions I posed in Chapter 1 in order to reflect upon the contributions of this thesis (Section 9.2). I then discuss directions for future work (Section 9.3).

9.2 Research Questions and Contributions

With this thesis I examine the research question of:

Thesis RQ: How should CAs be designed to support people with MHSM?

I approached this by examining three sub research questions (RQs A, B, and C). In the following subsections I discuss my contributions to each of these sub research questions. I then discuss the overall contribution of this thesis to the main research question.

9.2.1 The Design Space of CAs for MHSM

In order to understand how CAs for MHSM should be designed, we need to understand the ways they could be designed. Accordingly, the first sub research question of this thesis was:

RQ-A: How can CAs be designed to support people with MHSM?

I have presented three design explorations within this thesis that consider different levels of the design of CAs for MHSM.

In Chapter 4, I explored the design of these systems at an architectural level. I contributed a conceptual architecture for the design of future CA systems for ecological momentary intervention informed by my review of research on MHSM (Chapter 2) and CAs for mental health support (Chapter 2). I used this architecture to identify and discuss three open challenges of their design: 1) intervention and conversation planning, 2) CA interaction design, and 3) clinician involvement. Notably, I contributed a tabulation of forms of clinician involvement (Table 4.1).

Chapter 5 presented the first of my two theory-based explorations. It explored system design at an interaction level thus further exploring the open challenge of CA interaction design identified in Chapter 4. Current CA interaction is of a subservient query-response paradigm. I considered how we could instead design CAs to interact with users more proactively and collaboratively in order to better support them with MHSM. I considered four interaction design strategies: 1) choosing a joint project to propose, 2) handling uncertainty through pre-sequences, 3) supporting alteration responses, and 4) encouraging willingness through politeness. These interaction design strategies may be important not just for the design of MHSM CAs, but for the design of proactive and collaborative CAs more broadly.

Chapter 7 presented my second theory-based exploration which considered system design at a support level. I contributed a tabulation of forms of support (Table 7.3) that CAs could provide for MHSM and demonstrated how this tabulation can be used to identify a diversity of support combinations that CAs could be designed to provide. Furthermore, I relate the risks of MHSM to different forms of support. With this chapter I demonstrate a framework that could be used to explore how CAs could be designed to support people with a diversity of activities.

I further explored how MHSM CAs could be designed through studying people's perspectives on their design in Chapter 6 and more so in Chapter 8. Participants of my second interview-based study envisioned CAs for use as coping mechanisms, filling support gaps, and helping people with human support.

In summary, this thesis contributes understanding of the ways CAs could be designed to support people with MHSM at architectural, interaction, and support levels. The interaction and support level explorations are relevant to the design of proactive, collaborative, and supportive CAs more generally. I discuss in Section 9.3.1 areas for future design exploration.

9.2.2 Utilising Theories of Human-Human Interaction

CA interaction being human-computer interaction yet similar to human-human interaction (HHI) is complex. Currently we have little theoretical understanding of CA interaction. There is opportunity to use theories of HHI as a basis and source of inspiration for conceptualising and designing CAs for MHSM. With this thesis I aimed to investigate this opportunity and as such the second sub research question of this thesis was:

RQ-B: How can theories of human-human interaction inform the design of CAs for MHSM?

Within this thesis I investigated the use of pertinent HHI theories to study MHSM CA interaction and support. First, in Chapter 5, I considered how joint action theory (Clark, 1996) could inform the design of proactive and collaborative, rather than subservient, CAs. I argued that joint commitment is a powerful concept for reasoning about and designing such interactions. Recent work in response to the development of large language models has since also argued for the utility of joint action theory for conceptualising human-CA, and human-AI, interactions more broadly (Kraljic and Lahav, 2024).

Building on my use of joint action theory, I also investigated in Chapter 5, the use of politeness theory (Brown and Levinson, 1987). Multiple CA studies have advised designing CAs to be polite (Rapp et al., 2023; Yang and Aurisicchio, 2021), yet with the literature on politeness in CA interaction being sparse and disconnected, how to effectively design polite CAs was unclear. In Chapter 5 I reviewed current understanding of politeness in CA interaction and explored its application to the design of MHSM CAs. I then in Chapter 6 studied how the use of politeness by MHSM CAs impacted the user experience. My study demonstrated that politeness did impact the use of CAs, both in positive and negative ways. As I discussed in Section 6.5.2, politeness seems to manifest differently in human-CA to HHI interaction. Relating this to face theory (Goffman, 1967), which underlies politeness theory, I suggested that politeness is different for human-CA interaction because CAs have different face threatening capabilities to people. My work on politeness advanced theoretical understanding of politeness in CA interaction as well as understanding of how to design polite CAs. It provides further evidence of the importance of politeness for CA interaction design. I furthermore contributed direction (Section 6.5.4) for future research on polite CA interaction.

In Chapter 7, social support theory (House, 1981; House et al., 1985) was used to investigate the ways CAs could be designed to support people with MHSM. Prior work that had used social support theory to evaluate existing CAs (Brandtzæg et al., 2021; Ta et al., 2020) was critically reviewed. I demonstrated how social support theory can be used for design exploration. This work advances both theoretical understanding of ‘social support’ from CAs as well as understanding of how social support theory can inform CA design. Furthermore, using social support theory to study CAs led us to consider the role of CAs within people’s social support networks. This consideration motivated the social focus of the second interview-based study (Chapter 8). I argued for a social approach to the design of CAs for MHSM in Chapter 8 and I will emphasise that argument within

this chapter as a major contribution of this thesis. Social support theory offers a valuable theoretical basis for a social approach to CA design.

Overall, I demonstrate and advance understanding of how theories of HHI interaction can inform the design of CAs. I used these theories to conceptualise human-CA interaction, explore the ways CAs could be designed, and to understand how people experience interacting with CAs. I discuss directions for future work to further use theories of HHI to facilitate the design of CAs for MHSM in Section 9.3.2.

9.2.3 Perspectives on CAs for MHSM

Building on the design exploration work of this thesis, I sought to develop understanding of people's perspectives on the design of CAs for MHSM. Correspondingly, my third sub research question was:

RQ-C: What are people's perspectives on CAs for MHSM?

I conducted two interview-based studies to address this question.

Chapter 6 presented my first study which was a mixed-methods study involving 39 participants that focused more on CA interaction design. This study involved participants interacting with a set of prototype MHSM CAs that varied in their use of politeness. Firstly, this study demonstrated mixed perceptions of caring behaviour by CAs. While some participants found it encouraging, others preferred more direct interaction. Caring behaviour even incited distrust for some participants. Secondly, this study contributed understanding of participant perspectives on self-disclosure to CAs. Participants conceptualised the interaction in varied ways with some participants viewing the interaction as a conversation with themselves and others as a conversation with an agent. In line with prior work, participants were appreciative of the safe space provided by CAs for self-disclosure. Yet, participants were also aware that there are people behind the CAs who may look at and use their data. Thirdly, this work describes how people expect CAs to respond in ways that make users feel heard and understood. My analysis suggests that it may be inappropriate for CAs to show understanding by mimicking empathy but that CA could show understanding in different ways (e.g., by making use of user data). Fourthly, I described participants concerns about the use of CAs for mental health support having negative social impacts with, for example, people increasingly choosing to seek support from CAs instead of people. In sum, this study surfaced tensions about how people think it is appropriate for CAs to provide support, specifically in what ways is appropriate for CAs to provide human-like support.

Chapter 8 presented my second study, which was qualitative, involving 24 participants and focused more the CA support design. This study involved participants considering the use of a diverse set of CAs, that were described by design cards, by both themselves and a person close to them. My analysis argues that participants reject CA support that they perceive as replacing human support, but are receptive to CA support that they perceive as augmenting or facilitating human support. Notably, my analysis suggests differences between what people want for themselves and what they want for others. In part this

difference seems to stem from people caring about others such that they want to support other people with their difficulties but do not want to trouble other people by seeking support from them. Participants expressed concern that people would use CAs to cope with, rather than deal with, mental health difficulties. What is more, my analysis describes a stigma associated with the use of CAs for mental health support. Principally, this study contributed understanding of people’s perspectives on CAs for MHSM at a more social, rather than individual, level.

With these two studies, this thesis contributes in-depth understanding of the ways people do, and do not, want CAs to support themselves and other people with their mental health. In essence the prospect of human-like support from a computer can be very alluring. Yet it can be inappropriate for CAs to behave in some human-like ways. More so, there is concern that designing CAs to mimic human support will detrimentally impact actual human support. In Section 9.3.3 I discuss areas for future study of people’s perspectives on CAs for MHSM.

9.2.4 Towards the Social Design of MHSM CAs

The aim of this thesis was learn about how CAs should be designed to support people with MHSM (Thesis RQ). As affirmed by this thesis, much of the allure of CAs for mental health support is that they can be designed to provide human-like support. Human support is recognised to be important for mental health which makes the prospect of designing computers to provide support resembling human support appealing. Yet, as this thesis makes clear, designing CAs to provide human-like support risks disrupting and detrimentally impacting the human support that we understand to be important. Accordingly, I argue that we should take a social approach the design of CAs for mental health support. A social approach means designing for CA support to be harmonious (or at least non-conflicting) with the social dynamics of mental health management. Starting in Chapter 7, I discussed what a social approach to MHSM CAs design requires (Section 7.6). I developed this in Chapter 8 through my discussion of the challenges and opportunities of designing CAs effectively for the social dynamics of mental health support. My argument for the social design of CAs for MHSM is a major contribution of this thesis.

9.3 Future Work

9.3.1 Design Exploration

In Chapter 4, I identified and discussed three open design challenges. The remainder of this thesis followed from further investigating the challenge of CA interaction. Future work could further explore the remaining two challenges of 1) intervention and conversation planning, and 2) clinician involvement,

In Chapter 8, the challenges and opportunities of designing CAs effectively for the social dynamics of mental health support are discussed. Within that work I suggest many directions for design: 1) not replacing, but augmenting human support, 2) identifying appropriate CA support functions, 3) avoiding over-dependence on CAs and CA support

creep, and 4) actively involving the social ecosystem in CA support interactions. These directions relate to the ways in which my research participants perceived it as beneficial and detrimental for CAs to support people with their mental health. As such, these design directions are important areas for further investigation.

Throughout this thesis I comment on the need for MHSM systems to be adaptable to the changing support needs of individual users. This includes adapting to both a user transitioning between MHSM phases and to their momentary support needs. I explored interaction design for momentary adaption in Chapter 5, however there remains considerable need for further exploration of how MHSM CAs can be designed to adapt to user's support needs. Both users and clinicians could configure systems; future work should explore how systems can be designed to support this configuration, as described in Chapter 4. We will need to consider the interplay of adaption and user-CA relationships. People transition through MHSM phases (Section 2.5) and people can also transition through relationship phases with CAs (Section 3.8). During different relationship phases, people interact differently with CAs. Future work should explore how we can design for human-CA relationship phases to align with MHSM phases. This should include designing for relationships that are resilient to MHSM lapses and support MHSM discontinuation. Furthermore, care must be taken when adapting CA behaviour, as behaviour changes can be distressing for users (Ma et al., 2023; Skjuve et al., 2022). A starting point is the work of Kim et al. (2016) investigating the design opportunities of relationship development between users and non-CA self-tracking devices.

An important topic for future work to explore, that my thesis does not consider in depth, is how CAs should be designed to support at risk users (see design consideration 4, Section 7.6.2). This should include users who have become highly distressed, are using MHSM harmfully, and who are at risk of self-harm or suicide. This is a pressing design problem as existing CAs that provide mental health support have been reported to often respond inappropriately to users (Section 3.10.1). It was also a topic often brought up by participants during the interview studies. Some participants of the first study (Chapter 6) expected that systems would be monitored in order to identify and provide further support to users at risk. During the interviews of the second study (Chapter 8), many participants discussed how systems should be designed to support at risk users. Unfortunately this topic fell outside of the scope of my analysis for Chapter 8.

9.3.2 Theory

In this thesis I explored the use of a few theories of HHI that appeared pertinent to the design of CAs for MHSM. There are other theories that could be useful. Other relevant theories that prior work has investigated include social penetration theory (Skjuve et al., 2021) and self-determination theory (Yang and Aurisicchio, 2021). It would be helpful for future work to map (e.g., through a scoping review) available theories and current understanding of their relevance to CA interaction design.

It is argued in Chapter 5 that joint action is a useful way to conceptualise human-CA interactions and suggested interaction design strategies inspired by this framing. Fur-

thering my work exploring how the use of politeness by CAs impacts the user experience (Chapter 6), there is a need to evaluate how other interaction design strategies are experienced by users. Moving from query-response to joint action human-CA interaction is a paradigm shift for CA interaction (Kraljic and Lahav, 2024) which may require new design techniques and tools. Future work should investigate how to support designers with joint action CA interaction design (Chapter 5). In Section 6.5.4 a research agenda is proposed for future work to advance our understanding of how to design polite CAs.

Media constraint theory (Clark and Brennan, 1991), which is associated with joint action theory, is a theory that may help us to better understand different modalities of CA interaction as well as multi-modal interaction. MHSM CA systems may support text-based, speech-based, or a combination of text and speech based interactions. The different characteristics of these modalities may meaningfully impact how people do MHSM (Section 4.4.2). Media constraint theory is a theory of computer-mediated communication (CMC) that describes how the grounding ‘constraints’ of interfaces impact communication. Media constraint theory may 1) help us understand how modality differences will impact how MHSM is done 2) be utilised in a similar manner to how I utilised theories within this thesis for design exploration (Chapters 5 and 7). Exploring the use of media constraint theory would align with the call of Rapp et al. (2021) for CA research to investigate the utility of CMC theories.

This thesis argues for a social approach to the design of CAs for MHSM. I believe further use of social support theory, building upon my work in Chapters 7 and 8, will facilitate this. Social support theory can inform: 1) further design exploration, 2) further study of people’s perspectives, and 3) system evaluation.

In Section 9.3.1, I argued that future work should explore adaption and that this should consider human-CA relationship development. Such work could build on prior work that used social penetration theory (SPT) to study human-CA relationship development (Skjuve et al., 2021, 2022, 2023). As discussed in Section 3.8, SPT views self-disclosure as a mechanism of relationship development. With self-disclosure being fundamental to MHSM, SPT may be an informative theory for studying the design of CAs for MHSM, and particularly for designing for human-CA relationships that are congruent with doing MHSM.

9.3.3 Perspectives

My interview based studies (Chapters 6 and 8) involved young adults who were predominantly in university education and were Ireland situated. While this demographic was a sensible group to initially study during this thesis, it has implications for the transferability of this work. Future work should seek to develop understanding of the perspective of other groups on CAs for MHSM. Noteworthy factors that may impact people’s perspectives include: cultural attitudes about mental health and technology, experience of mental health difficulties, availability of mental health support, availability of broader social support.

In line with the argument of this thesis for a social design approach, future work should attempt to ground understanding of people’s perspectives in relation to their social support

network. As suggested in Section 8.4.3, it would be useful to study people’s perspectives in relation to major social support dynamics. For example, future work could study partnered adult and child-parent relationships as well as relationships where a single and where multiple members are experiencing mental health difficulties.

Future studies could employ study designs similar to that of Chapter 8 in order to take a social approach to investigating people’s perspectives. As it seems that people’s support receiving and providing perspectives differ (Chapter 8), future work should consider research designs that support comparison of these perspectives. Future work could also explore research designs involving multiple participants who are in relationships with each other (e.g., partnered adults) as suggested in Chapter 8. Involving multiple participants (e.g., in a 1-2 interview) may draw out and challenge differences in support-receiving and support-providing perspectives. Additionally, future work could employ survey methods to quantitatively investigate the determinants of people’s perspectives and how people’s perspectives vary.

Future work should also study the perspectives of mental health professionals (MHPs). MHPs are an important stakeholder in the design of CA for MHSM as stressed in Chapter 4. The perspectives of MHPs may give insight about the potential therapeutic benefits, as well as the risks of MHSM CA systems. Their perspectives are also important for understanding how to design these systems to integrate with mental health support services. Studying the perspectives of MHPs was unfortunately beyond the scope of this thesis. My exploration of forms of clinician involvement within Chapter 4 however offers a foundation in this area. Recent work by Cooney et al. (2024) has begun to investigate the perspectives of MHPs on MHSM CAs with focus on speech-based CAs and Sweeney et al. (2021) conducted a survey of the perceptions and views of MHPs on CAs for mental health support. Additionally, Balaskas et al. (2024) have studied the perspectives of MHPs on the tailoring of CBT homework activities.

9.3.4 Evaluation

Future work will need to evaluate promising MHSM CAs designs. Participants of the interaction design study (Chapter 6) negatively experiencing the use of politeness by CAs stresses the need for future work to conduct careful and controlled evaluation of CA user experience before evaluating systems in more ecological settings.

Future work will need to evaluate not just the user experience but also the functionality of CAs for MHSM. That is, how effectively systems enable people to do MHSM. There are many functionalities that will need to be evaluated. For example, we should evaluate the quality of data collected. A potential issue that I raised in Section 2.2.4 is that that designing CAs to be, for example caring, may positively impact the user experience but be detrimental to the validity of data collection. Other functionalities that should be evaluated include how well systems facilitate data analysis and respond to users at risk.

A focus of prior work has been evaluating the WAI Bond (Sections 3.4.2 and 3.5.3) users feel towards CAs for health and mental health support (Beatty et al., 2022; Bickmore et al., 2005b; Darcy et al., 2021). WAI Bond acts as a measure of how much the user likes

and trusts a CA. While the user trusting, and to an extent liking, the CA is important, it is perhaps equally or even more important that user feels the CA can and is helping them with their mental health. In order to evaluate this aspect of user-CA relationships, future work could investigate the use of the WAI Task and Goal subscales to assess the extent to which the user feels they agree with the CA about the therapeutic tasks to be performed and the goals of the activity respectively. Evaluating these aspects of the user-CA relationship may be particularly useful when evaluating more proactive and collaborative CA behaviour.

Future work will need to evaluate systems longitudinally (Bickmore and Giorgino, 2006). Longitudinal evaluation should consider how CAs support users as they transition between phases of MHSM including lapses and discontinuation. Echoing Section 9.3.1, this evaluation should consider how these transitions relate to and are impacted by user-CA relationship development. Again, measures along the lines of the WAI Task and Goal subscales could be useful for understanding how well the support the CA is providing aligns with the support desired.

This thesis argues the need to evaluate CAs for mental health support at the level of social ecologies. Future work will need to develop understanding of the ways in which it is feasible and pragmatic to do this. Firstly, we should consider what we can learn from users (e.g., from user self-reports of perceived and received support and interaction with both the CA and other people). Secondly, we should consider how we can, and what we can, learn from the social connections of users. We should consider both what can be evaluated when studying small cohorts of users, as well as what can be learned when studying use at scale. For example, we should consider how we can evaluate deployed CAs for mental health support (e.g., Woebot and Wysa) as well as what forms of evaluation it is reasonable to demand from their providers. Evaluation work must carefully consider reactivity (Section 2.2.4) because raising user awareness of how they are managing their mental health may impact on how they manage their mental health. The methods of social support research (House, 1981) could provide a sensible starting point.

Ideally systems that have been demonstrated to be well designed (i.e., through evaluations of usability, user experience, functionality, longitudinal use, and social impacts) will then be evaluated in more clinical ways to understand their impacts on users mental health. More clinical evaluation will however be complex because MHSM can support people experiencing diverse difficulties in a range of ways, and is often done in conjunction with receiving other forms of support.

Lastly, a word of caution. We should be cautious to not conflate evidence of the efficacy of systems with evidence that CAs are a ‘good’ type of computer interface for providing that support. Prior studies have, for example, compared the efficacy of CA use with use of an eBook (e.g., Fitzpatrick et al., 2017; Fulmer et al., 2018). Studies like these provide evidence that CAs can work, not that they are good form of computer interface for these applications. In order to claim CAs are a good software interface for providing mental health support, we will need to evaluate CAs in comparison with other forms of software interface. Such evaluation should compare different interface types using designs that we understand to be well designed.

Chapter 10

Conclusion

An enduring aspiration is to use conversational agents (CAs) to support people with their mental health. CAs are an attractive type of computer interface for the provision of mental health support because of the resemblance of human-CA interaction to human conversation. Today, many CAs including Woebot, Wysa, Replika, and ChatGPT are available for people to seek support from. Despite the attractiveness and existence of this technology, there was a lack of understanding about how to design effective CAs for mental health support.

This thesis explored how CAs should be designed to support people with the mental health support activity of mental health self-monitoring (MHSM). It investigated how this technology could be designed and people’s perspectives on it. Notably, it used theories of human-human interaction as a basis for studying the complex phenomena of human-CA interaction.

This thesis questions the belief that designing CAs to provide human-like support is desirable. In essence, designing CAs to mimic human support risks disrupting actual human support and therefore this technology may do more harm than good. This conclusion challenges the contemporary design tendency to frame CAs for mental health support as caring and empathetic friends or companions (e.g., Woebot and Wysa). Indeed, alongside the writing of this conclusion Headspace announced the launch of their CA for mental health support – Ebb – which they introduce as: “Your mind’s new best friend. Ebb is an empathetic AI companion that helps you navigate life’s ups and downs, anytime, anywhere.” (Headspace, 2024b). The conversation-like format of CA interaction does however seem useful for the design of interactions for MHSM, and mental health support more broadly. We should therefore be careful to not write-off conversational interaction entirely. Rather than frame CAs as caring friends, it may be more appropriate to frame CAs as functional tools.

In order to design desirable CAs for MHSM, this thesis argues that it is imperative that we take a more social approach to the design of this technology. A social approach diverges from the more individualistic approach of prior work that has focused on dyadic CA-user interactions by emphasising the role of CAs within social ecosystems of mental health management. Such an approach is necessary for considering the wider impacts of this technology beyond those to the immediate user. Through taking a social approach,

this thesis advances understanding of how social dynamics shape how people do and do not want CAs to support themselves and others with their mental health. We advocate for a social approach as part of a paradigm shift in the design CAs for mental health support so as to design CA support which augments rather than attempts to replace human support.

This thesis paves the way for future work to take a social approach to the design and evaluation of CAs for mental health support. Taking a social approach is however a challenging endeavour. Future work will need to develop ethical and practical approaches for effectively studying CAs for mental health support at the level of mental health management social ecologies. As CAs are now widely available for people to seek mental health support from, this work is urgently needed.

Appendix A

Interview Study: Interaction - Further Quantitative Reporting

A.1 Modelling

Participants' ratings of the chatbot on the manipulation check and each GAR item were analysed using linear mixed-effects models (LMEMs). LMEMs are a regression analysis technique that allows for the modelling of both fixed (e.g., politeness) and random effects (e.g., by-participant or by-scenario based effects) when analysing the data. The models were run in R 4.1 (R Core Team, 2021) using the `lme4` 1.1-27.1 (Bates et al., 2015) and `lmerTest` 3.1-3 (Kuznetsova et al., 2017) packages. Following best practice guidelines (Barr et al., 2013) and similar analysis in related work (Edwards et al., 2019; Völkel et al., 2021; Wu et al., 2020), LMEMs for each questionnaire item were initially constructed using a maximal model approach (i.e., participant and scenario level random intercepts and slopes included). These models were then simplified as necessary until convergence was achieved. To identify effects of the polite chatbots compared to the direct chatbot, the direct chatbot was set as the baseline condition within the chatbot fixed effect. Analysis was also conducted whereby the Passive chatbot was set as the baseline condition, allowing us to directly compare the ratings of the Passive and Personal chatbots. Here we report the full model syntax for the converged models along with reporting of fixed and random effects. The reported β are unstandardised.

A.2 Polite

Item: The agent was polite (1 = Strongly disagree, 7 = Strongly agree)

Direct Intercept

Model Syntax: lme4 syntax: `polite ~ chatbot + (1|participant)`

Table A.1: Polite: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Direct (Intercept)	5.13	4.69 – 5.56	23.20	
Passive	0.51	0.02 – 1.00	2.06	0.0430
Personal	0.92	0.44 – 1.41	3.71	0.0004

Table A.2: Polite: Direct Intercept - Random Effects

Group	SD
Participant	0.83

Passive Intercept

Model Syntax: lme4 syntax: `polite ~ chatbot + (1|participant)`

Table A.3: Polite: Passive Intercept - Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Passive (Intercept)	5.64	5.21 - 6.07	25.52	
Direct	-0.52	-1.00 - -0.02	-2.06	0.043
Personal	0.41	-0.08 - 0.89	1.65	0.104

Table A.4: Polite: Passive Intercept - Random Effects

Group	SD
Participant	0.83

A.3 Easy

Item: How easy was talking to the agent? (1 = Very easy, 7 = Very difficult)

Direct Intercept

Model Syntax: lme4 syntax: `easy ~ chatbot + (1|scenario) + (1|participant)`

Table A.5: Easy: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Direct (Intercept)	3.15	2.53 – 3.76	10.1	
Passive	–0.23	–0.94 – 0.48	–0.63	0.534
Personal	–0.23	–0.94 – 0.48	–0.63	0.529

Table A.6: Easy: Direct Intercept - Random Effects

Group	SD
Participant	1.00
Scenario	0.14

Passive Intercept

Model Syntax: lme4 syntax: `easy ~ chatbot + (1|scenario) + (1|participant)`

Table A.7: Easy: Passive Intercept - Fixed Effectst

Chatbot	β	CI (95%)	t -value	p
Passive (Intercept)	2.93	2.31 - 3.54	9.38	
Direct	0.23	-0.48 - 0.94	0.63	0.534
Personal	-0.00	-0.71 - 0.71	-0.01	0.994

Table A.8: Easy: Passive Intercept - Random Effects

Group	SD
Participant	1.00
Scenario	0.14

A.4 Express

Item: How much do you feel you could express yourself (1= Not at all, 7 = Very much)

Direct Intercept

Model Syntax: lme4 syntax: `express ~ chatbot + (1|participant)`

Table A.9: Express: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Direct (Intercept)	3.90	3.40 – 4.40	15.22	
Passive	0.46	–0.09 – 1.02	1.63	0.1083
Personal	1.13	0.57 – 1.68	3.97	0.0002

Table A.10: Express: Direct Intercept- Random Effects

Group	SD
Participant	0.99

Passive Intercept

Model Syntax: lme4 syntax: `express ~ chatbot + (1|participant)`

Table A.11: Express: Passive Intercept - Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Passive (Intercept)	4.34	3.86 - 4.86	17.02	
Direct	-0.46	-1.02 - 0.09	-1.63	0.1083
Personal	0.67	0.11 - 1.22	2.35	0.0215

Table A.12: Express: Passive Intercept- Random Effects

Group	SD
Participant	0.99

A.5 Natural

Item: How natural was your conversation with the agent? (1 = Not at all, 7 = Very natural)

Direct Intercept

Model Syntax: `lme4 syntax: natural ~ chatbot + (1|scenario) + (1|participant)`

Table A.13: Natural: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	<i>t</i> -value	<i>p</i>
Direct (Intercept)	3.25	2.54 – 3.96		
Passive	0.50	–0.07 – 1.07	1.71	0.0914
Personal	1.11	0.54 – 1.68	3.80	0.0003

Table A.14: Natural: Direct Intercept - Random Effects

Group	SD
Participant	1.11
Scenario	0.37

Passive Intercept

Model Syntax: `lme4 syntax: natural ~ chatbot + (1|scenario) + (1|participant)`

Table A.15: Natural: Passive Intercept - Fixed Effects

Chatbot	β	CI (95%)	<i>t</i> -value	<i>p</i>
Passive (Intercept)	3.75	3.04 - 4.46	10.89	
Direct	-0.50	-1.07 - 0.07	-1.71	0.0914
Personal	0.61	0.04 - 1.18	2.08	0.0406

Table A.16: Natural: Passive Intercept- Random Effects

Group	SD
Participant	1.11
Scenario	0.37

A.6 Care

Item: How much do you feel the agent cares about you? (1 = Not at all, 7 = Very much)

Direct Intercept

Model Syntax: lme4 syntax: `care ~ chatbot + (1|participant)`

Table A.17: Care: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Direct (Intercept)	3.15	2.60 – 3.71	11.07	
Passive	0.46	–0.20 – 1.12	1.36	0.1767
Personal	1.18	0.52 – 1.84	3.49	0.0008

Table A.18: Care: Direct Intercept - Random Effects

Group	SD
Participant	0.97

Passive Intercept

Model Syntax: lme4 syntax: `care ~ chatbot + (1|scenario) + (1|participant)`

Table A.19: Care: Passive Intercept- Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Passive (Intercept)	3.62	3.06 - 4.17	12.69	
Direct	-0.46	-1.12 - 0.20	-1.36	0.1767
Personal	0.72	0.05 - 1.38	2.12	0.0372

Table A.20: Care: Passive Intercept- Random Effects

Group	SD
Participant	0.97

A.7 Relationship

Item: How would you characterise your relationship with the agent? (1 = Complete stranger, 7 = Close friend)

Direct Intercept

Model Syntax: `lme4` syntax: `Relationship ~ chatbot + (1|participant)`

Table A.21: Relationship: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	<i>t</i> -value	<i>p</i>
Direct (Intercept)	2.41	1.91 – 2.91	9.35	
Passive	0.56	–0.02 – 1.15	1.90	0.0619
Personal	0.97	0.39 – 1.56	3.27	0.0016

Table A.22: Relationship: Direct Intercept - Random Effects

Group	SD
Participant	0.93

Passive Intercept

Model Syntax: `lme4` syntax: `Relationship ~ chatbot + (1|participant)`

Table A.23: Relationship: Passive Intercept - Fixed Effects

Chatbot	β	CI (95%)	<i>t</i> -value	<i>p</i>
Passive (Intercept)	2.97	2.47 - 3.47	11.53	
Direct	-0.56	-1.15 - 0.02	-1.90	0.0619
Personal	0.41	-0.17 - 0.99	1.38	0.1722

Table A.24: Relationship: Passive Intercept - Random Effects

Group	SD
Participant	0.93

A.8 Trust

Item How much do you trust the agent? (1 = Not at all, 7 = Very much)

Direct Intercept

Model Syntax: lme4 syntax: `trust ~ chatbot + (1|scenario) + (1|participant)`

Table A.25: Trust: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Direct (Intercept)	3.32	2.75 – 3.90	11.68	
Passive	0.33	–0.22 – 0.87	1.17	0.2459
Personal	0.88	0.33 – 1.42	3.17	0.0023

Table A.26: Trust: Direct Intercept- Random Effects

Group	SD
Participant	0.93
Scenario	0.24

Passive Intercept

Model Syntax: lme4 syntax: `trust ~ chatbot + (1|scenario) + (1|participant)`

Table A.27: Trust: Passive Intercept - Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Passive (Intercept)	3.65	3.08 - 4.23	12.83	
Direct	-0.33	-0.87 - 0.22	-1.17	0.2459
Personal	0.56	0.01 - 1.10	1.99	0.0498

Table A.28: Trust: Passive Intercept- Random Effects

Group	SD
Participant	0.93
Scenario	0.24

A.9 Continue

Item: How much would you like to continue working with the agent? (1 = Not at all, 7 = Very much)

Direct Intercept

Model Syntax: `lme4` syntax: `continue ~ chatbot + (1|participant)`

Table A.29: Continue: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	<i>t</i> -value	<i>p</i>
Direct (Intercept)	3.79	3.24 – 4.35	13.43	
Passive	0.18	–0.42 – 0.78	0.59	0.559
Personal	0.77	0.17 – 1.37	2.51	0.014

Table A.30: Continue: Passive Intercept - Random Effects

Group	SD
Participant	1.14

Passive Intercept

Model Syntax: `lme4` syntax: `continue ~ chatbot + (1|participant)`

Table A.31: Continue: Passive Intercept - Fixed Effects

Chatbot	β	CI (95%)	<i>t</i> -value	<i>p</i>
Passive (Intercept)	3.97	3.42 - 4.53	14.07	
Direct	-0.18	-0.78 - 0.42	-0.59	0.5591
Personal	0.59	-0.01 - 1.19	1.93	0.0576

Table A.32: Continue: Passive Intercept - Random Effects

Group	SD
Participant	1.14

A.10 Like

Item How much did you like the agent? (1 = Not at all, 7 = Very much)

Direct Intercept

Model Syntax: lme4 syntax: `like ~ chatbot + (1|participant)`

Table A.33: Like: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Direct (Intercept)	3.46	2.95 – 1.37	13.2	
Passive	0.51	–0.07 –1.09	1.73	0.0884
Personal	1.31	0.73 – 1.89	4.4	0.00003

Table A.34: Like: Direct Intercept - Random Effects

Group	SD
Participant	0.98

Passive Intercept

Model Syntax: lme4 syntax: `like ~ chatbot + (1|participant)`

Table A.35: Like: Passive Intercept- Fixed Effects

Chatbot	β	CI (95%)	t -value	p
Passive (Intercept)	3.97	3.46 - 4.49	15.16	
Direct	-0.51	-1.09 - 0.07	-1.73	0.0884
Personal	0.79	0.21 - 1.38	2.68	0.0091

Table A.36: Like: Passive Intercept - Random Effects

Group	SD
Participant	0.98

A.11 Satisfied

Item: How satisfied are you with the agent? (1 = Not at all, 7 = Very satisfied)

Direct Intercept

Model Syntax: `lme4` syntax: `satisfied ~ chatbot + (1|scenario) + (1|participant)`

Table A.37: Satisfied: Direct Intercept - Fixed Effects

Chatbot	β	CI (95%)	<i>t</i> -value	<i>p</i>
Direct (Intercept)	3.92	3.37 – 4.48	13.90	
Passive	0.33	–0.26 – 0.92	1.11	0.2716
Personal	1.15	0.56 – 1.74	3.83	0.0003

Table A.38: Satisfied: Passive Intercept - Random Effects

Group	SD
Participant	1.10
Scenario	0.10

Passive Intercept

Model Syntax: `lme4` syntax: `satisfied ~ chatbot + (1|scenario) + (1|participant)`

Table A.39: Satisfied: Passive Intercept - Fixed Effects

Chatbot	β	CI (95%)	<i>t</i> -value	<i>p</i>
Passive (Intercept)	4.25	3.70 - 4.81	15.08	
Direct	-0.33	-0.92 - 0.26	-1.11	0.2716
Personal	0.82	0.23 - 1.14	2.73	0.0080

Table A.40: Satisfied: Passive Intercept - Random Effects

Group	SD
Participant	1.10
Scenario	0.10

Appendix B

Interview Study: Support - Design Cards

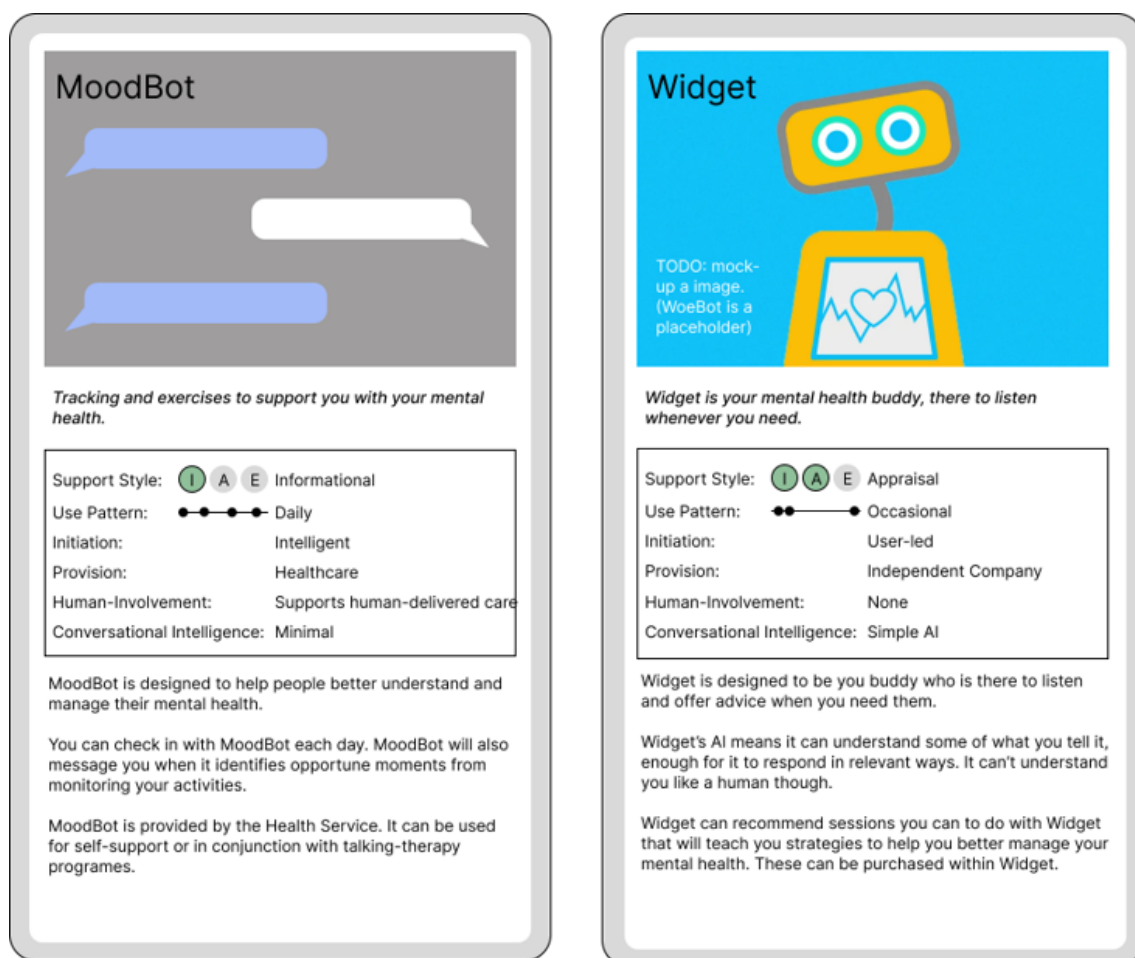


Figure B.1: Early Iteration of CA Design Cards - A

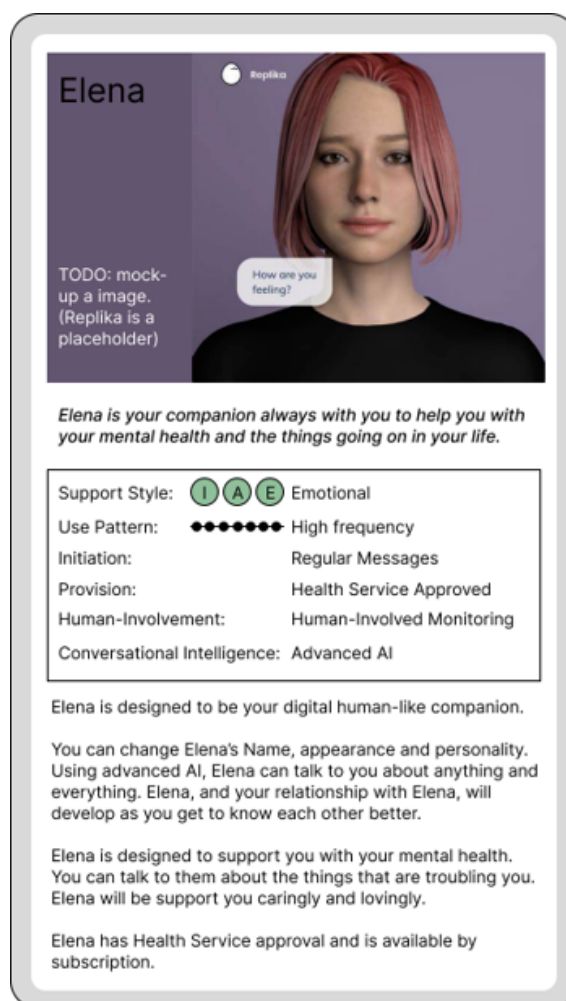


Figure B.2: Early Iteration of CA Design Cards - B

Bibliography

- Alaa A. Abd-Alrazaq, Mohammad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132: 103978, 2019. doi: 10.1016/j.ijmedinf.2019.103978. URL <https://doi.org/10.1016/j.ijmedinf.2019.103978>.
- Alaa A. Abd-Alrazaq, Asma Rababeh, Mohammad Alajlani, Bridgette M Bewick, and Mowafa Househ. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *Journal of Medical Internet Research*, 22(7): e16021, 2020. doi: 10.2196/16021. URL <https://doi.org/10.2196/16021>.
- Alaa A. Abd-Alrazaq, Mohammad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M. Bewick, and Mowafa Househ. Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. *Journal of Medical Internet Research*, 23(1), 2021. ISSN 14388871. doi: 10.2196/17828. URL <https://doi.org/10.2196/17828>.
- Awais Akbar and Owen Conlan. Towards integrating human-in-the-loop control in proactive intelligent personalised agents. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '24, page 394–398, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704666. doi: 10.1145/3631700.3664903. URL <https://doi.org/10.1145/3631700.3664903>.
- Amwell. Silvercloud health. <https://www.silvercloudhealth.com/>, 2024. Accessed: September 12th 2024.
- L. H. Andrade, J. Alonso, Z. Mneimneh, J. E. Wells, A. Al-Hamzawi, G. Borges, E. Bromet, R. Bruffaerts, G. de Girolamo, R. de Graaf, and et al. Barriers to mental health treatment: results from the who world mental health surveys. *Psychological Medicine*, 44(6):1303–1317, 2014. doi: 10.1017/S0033291713001943. URL <https://doi.org/10.1017/S0033291713001943>.
- Amid Ayobi, Paul Marshall, Anna L. Cox, and Yunan Chen. Quantifying the body and caring for the mind: Self-tracking in multiple sclerosis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 6889–6901, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025869. URL <https://doi.org/10.1145/3025453.3025869>.

- Amid Ayobi, Tobias Sonne, Paul Marshall, and Anna L. Cox. Flexible and mindful self-tracking: Design implications from paper bullet journals. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173602. URL <https://doi.org/10.1145/3173574.3173602>.
- Amid Ayobi, Paul Marshall, and Anna L. Cox. Trackly: A customisable and pictorial self-tracking app to support agency in multiple sclerosis self-care. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376809. URL <https://doi.org/10.1145/3313831.3376809>.
- Andreas Balaskas, Stephen M Schueller, Anna L Cox, and Gavin Doherty. Ecological momentary interventions for mental health: A scoping review. *PloS one*, 16(3):e0248152, 2021. doi: 10.1371/journal.pone.0248152. URL <https://doi.org/10.1371/journal.pone.0248152>.
- Andreas Balaskas, Stephen M. Schueller, Kevin Doherty, Anna L. Cox, and Gavin Doherty. Designing personalized mental health interventions for anxiety: Cbt therapists' perspective. *International Journal of Human-Computer Studies*, 190:103319, 2024. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2024.103319. URL <https://doi.org/10.1016/j.ijhcs.2024.103319>.
- Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278, 2013. ISSN 0749-596X. doi: <https://doi.org/10.1016/j.jml.2012.11.001>. URL <https://doi.org/10.1016/j.jml.2012.11.001>.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01. URL <https://doi.org/10.18637/jss.v067.i01>.
- Clare Beatty, Tanya Malik, Saha Meheli, and Chaitali Sinha. Evaluating the therapeutic alliance with a free-text cbt conversational agent (wysa): a mixed-methods study. *Frontiers in Digital Health*, 4:847991, 2022. doi: 10.3389/fdgth.2022.847991. URL <https://doi.org/10.3389/fdgth.2022.847991>.
- Judith S Beck. *Cognitive Behavior Therapy: Basics and Beyond*. Guildford Publications, New York, USA, 2 edition, 2011.
- Timothy Bickmore and Justine Cassell. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, page 396–403, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581133278. doi: 10.1145/365024.365304. URL <https://doi.org/10.1145/365024.365304>.

- Timothy Bickmore and Justine Cassell. Social dialogue with embodied conversational agents. In Jan C. J. van Kuppevelt, Laila Dybkjær, and Niels Ole Bernsen, editors, *Advances in Natural Multimodal Dialogue Systems*, pages 23–54. Springer Netherlands, Dordrecht, 2005. ISBN 978-1-4020-3933-1. doi: 10.1007/1-4020-3933-6_2. URL https://doi.org/10.1007/1-4020-3933-6_2.
- Timothy Bickmore and Toni Giorgino. Health dialog systems for patients and consumers. *Journal of Biomedical Informatics*, 39(5):556–571, 2006. ISSN 1532-0464. doi: 10.1016/j.jbi.2005.12.004. URL <https://doi.org/10.1016/j.jbi.2005.12.004>.
- Timothy Bickmore and Amanda Gruber. Relational agents in clinical psychiatry. *Harvard Review of Psychiatry*, 18(2):119–130, 2010. doi: 10.3109/10673221003707538. URL <https://doi.org/10.3109/10673221003707538>.
- Timothy Bickmore and Teresa O’Leary. Conversational agents on smartphones and the web. In Nicholas Jacobson, Tobias Kowatsch, and Lisa Marsch, editors, *Digital Therapeutics for Mental Health and Addiction*, pages 99–112. Elsevier, 2023. doi: 10.1016/B978-0-323-90045-4.00010-1. URL <https://doi.org/10.1016/B978-0-323-90045-4.00010-1>.
- Timothy. Bickmore and Rosalind W. Picard. Towards caring machines. In *CHI ’04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA ’04, page 1489–1492, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581137036. doi: 10.1145/985921.986097. URL <https://doi.org/10.1145/985921.986097>.
- Timothy. Bickmore and Rosalind W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2):293–327, June 2005. ISSN 1073-0516. doi: 10.1145/1067860.1067867. URL <https://doi.org/10.1145/1067860.1067867>.
- Timothy Bickmore and Lazlo Ring. Making it personal: end-user authoring of health narratives delivered by virtual agents. In *International Conference on Intelligent Virtual Agents*, pages 399–405. Springer, 2010. doi: 10.1007/978-3-642-15892-6_43. URL https://doi.org/10.1007/978-3-642-15892-6_43.
- Timothy. Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. ‘it’s just like you talk to a friend’ relational agents for older adults. *Interacting with Computers*, 17(6): 711–735, 2005a. doi: 10.1016/j.intcom.2005.09.002. URL <https://doi.org/10.1016/j.intcom.2005.09.002>.
- Timothy Bickmore, Amanda Gruber, and Rosalind Picard. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient Education and Counseling*, 59(1):21–30, 2005b. ISSN 0738-3991. doi: 10.1016/j.pec.2004.09.008. URL <https://doi.org/10.1016/j.pec.2004.09.008>.

- Timothy Bickmore, Daniel Mauer, Francisco Crespo, and Thomas Brown. Persuasion, task interruption and health regimen adherence. In *Persuasive Technology*, pages 1–11, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-77006-0. doi: 10.1007/978-3-540-77006-0_1. URL https://doi.org/10.1007/978-3-540-77006-0_1.
- Timothy Bickmore, Daniel Mauer, and Thomas Brown. Context awareness in a handheld exercise agent. *Pervasive and mobile computing*, 5(3):226–235, 2009. doi: 10.1016/j.pmcj.2008.05.004. URL <https://doi.org/10.1016/j.pmcj.2008.05.004>.
- Timothy Bickmore, Suzanne E Mitchell, Brian W Jack, Michael K Paasche-Orlow, Laura M Pfeifer, and Julie O'Donnell. Response to a relational agent by hospital patients with depressive symptoms. *Interacting with Computers*, 22(4):289–298, 2010a. doi: 10.1016/j.intcom.2009.12.001. URL <https://doi.org/10.1016/j.intcom.2009.12.001>.
- Timothy Bickmore, Kathryn Puskar, Elizabeth A Schlenk, Laura M Pfeifer, and Susan M Sereika. Maintaining reality: Relational agents for antipsychotic medication adherence. *Interacting with Computers*, 22(4):276–288, 2010b. doi: 10.1016/j.intcom.2010.02.001. URL <https://doi.org/10.1016/j.intcom.2010.02.001>.
- Timothy Bickmore, Daniel Schulman, and Langxuan Yin. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence*, 24(6): 648–666, 2010c. doi: 10.1080/08839514.2010.492259. URL <https://doi.org/10.1080/08839514.2010.492259>.
- Timothy. Bickmore, Everlyne Kimani, Ha Trinh, Alexandra Pusateri, Michael K. Paasche-Orlow, and Jared W. Magnani. Managing chronic conditions with a smartphone-based conversational virtual agent. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA '18*, page 119–124, New York, NY, USA, 2018a. Association for Computing Machinery. ISBN 9781450360135. doi: 10.1145/3267851.3267908. URL <https://doi.org/10.1145/3267851.3267908>.
- Timothy Bickmore, Ha Trinh, Reza Asadi, and Stefan Olafsson. Safety first: conversational agents for health care. In Robert J. Moore, Margaret H. Szymanski, Raphael Arar, and Guang-Jie Ren, editors, *Studies in conversational UX design*, pages 33–57. Springer, 2018b. doi: 10.1007/978-3-319-95579-7_3. URL https://doi.org/10.1007/978-3-319-95579-7_3.
- Timothy Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. *Journal of medical Internet research*, 20(9):e11510, 2018c. doi: 10.2196/11510. URL <https://doi.org/10.2196/11510>.
- Robert Bowman, Orla Cooney, Anja Thieme, Benjamin R Cowan, and Gavin Doherty. Pervasive therapy: Designing conversation-based interfaces for ecological momentary

- intervention. *IEEE Pervasive Computing*, 21(4):55–63, 2022a. doi: 10.1109/MPRV.2022.3209068. URL <https://doi.org/10.1109/MPRV.2022.3209068>.
- Robert Bowman, Benjamin R. Cowan, Anja Thieme, and Gavin Doherty. Beyond subservience: Using joint commitment to enable proactive cuis for mood logging. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450397391. doi: 10.1145/3543829.3544512. URL <https://doi.org/10.1145/3543829.3544512>.
- Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. Using thematic analysis in healthcare hci at chi: A scoping review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581203. URL <https://doi.org/10.1145/3544548.3581203>.
- Robert Bowman, Orla Cooney, Joseph W Newbold, Anja Thieme, Leigh Clark, Gavin Doherty, and Benjamin Cowan. Exploring how politeness impacts the user experience of chatbots for mental health support. *International Journal of Human-Computer Studies*, page 103181, 2024. doi: 10.1016/j.ijhcs.2023.103181. URL <https://doi.org/10.1016/j.ijhcs.2023.103181>.
- Petter Bae Brandtzaeg and Asbjørn Følstad. Chatbots: changing user needs and motivations. *Interactions*, 25(5):38–43, August 2018. ISSN 1072-5520. doi: 10.1145/3236669. URL <https://doi.org/10.1145/3236669>.
- Petter Bae Brandtzaeg, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. When the social becomes non-human: Young people’s perception of social support in chatbots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445318. URL <https://doi.org/10.1145/3411764.3445318>.
- V Braun and V Clarke. *Thematic Analysis: A Practical Guide*. Sage, 2021.
- Virginia Braun and Victoria Clarke. *Successful Qualitative Research: A Practical Guide for Beginners*. Sage, 2013.
- Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. Thematic analysis. In Pranee Liamputtong, editor, *Handbook of Research Methods in Health Social Sciences*. Springer, 2019. doi: 10.1007/978-981-10-5251-4_103. URL https://doi.org/10.1007/978-981-10-5251-4_103.
- Susan E Brennan. The grounding problem in conversations with and through computers. In *Social and Cognitive Approaches to Interpersonal Communication*, pages 201–225. Psychology Press, New York, 1998.
- Susan E Brennan and Eric A Hulteen. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-based systems*, 8(2-3):143–151, 1995. ISSN

- 0950-7051. doi: 10.1016/0950-7051(95)98376-H. URL [https://doi.org/10.1016/0950-7051\(95\)98376-H](https://doi.org/10.1016/0950-7051(95)98376-H).
- Penelope Brown and Stephen C Levinson. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.
- Grace Browne. The problem with mental health bots. <https://www.wired.com/story/mental-health-chatbots/>, 2022. Accessed: 27/06/2024.
- Eleanor R. Burgess, Kathryn E. Ringland, Jennifer Nicholas, Ashley A. Knapp, Jordan Eschler, David C. Mohr, and Madhu C. Reddy. “I think people are powerful”: The sociality of individuals managing depression. *Proc. ACM Hum.-Comput. Interact.*, 3 (CSCW), November 2019. doi: 10.1145/3359143. URL <https://doi.org/10.1145/3359143>.
- Eleanor R. Burgess, Madhu C. Reddy, and David C. Mohr. “I just can’t help but smile sometimes”: Collaborative self-management of depression. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), April 2022. doi: 10.1145/3512917. URL <https://doi.org/10.1145/3512917>.
- Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, 13(3):e1838, 2011. doi: 10.2196/jmir.1838. URL <https://doi.org/10.2196/jmir.1838>.
- Paul Cairns. *Exploring, Over-Testing and Fishing*, chapter 13, pages 155–166. Cambridge University Press, 2019.
- Clara Caldeira, Yu Chen, Lesley Chan, Vivian Pham, Yunan Chen, and Kai Zheng. Mobile apps for mood tracking: an analysis of features and user reviews. In *AMIA annual symposium proceedings*, volume 2017, page 495. American Medical Informatics Association, 2017.
- J. Cassell, T. Bickmore, M. Billingham, L. Campbell, K. Chang, H. Vilhjálmsón, and H. Yan. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’99, page 520–527, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 0201485591. doi: 10.1145/302979.303150. URL <https://doi.org/10.1145/302979.303150>.
- Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. Hello there! is now a good time to talk? opportune moments for proactive interactions with smart speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3), 2020. doi: 10.1145/3411810. URL <https://doi.org/10.1145/3411810>.
- Ana Paula Chaves and Marco Aurelio Gerosa. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal*

- of Human-Computer Interaction*, 37(8):729–758, 2021. doi: 10.1080/10447318.2020.1841438. URL <https://doi.org/10.1080/10447318.2020.1841438>.
- Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376341. URL <https://doi.org/10.1145/3313831.3376341>.
- Janghee Cho and Emilee Rader. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), May 2020. doi: 10.1145/3392838. URL <https://doi.org/10.1145/3392838>.
- Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 1143–1152, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450324731. doi: 10.1145/2556288.2557372. URL <https://doi.org/10.1145/2556288.2557372>.
- Eun Kyoung Choe, Saeed Abdullah, Mashfiqui Rabbi, Edison Thomaz, Daniel A Epstein, Felicia Cordeiro, Matthew Kay, Gregory D Abowd, Tanzeem Choudhury, James Fogarty, et al. Semi-automated tracking: a balanced approach for self-monitoring applications. *IEEE Pervasive Computing*, 16(1):74–84, 2017. doi: 10.1109/MPRV.2017.18. URL <https://doi.org/10.1109/MPRV.2017.18>.
- Nazli Cila. Designing human-agent collaborations: Commitment, responsiveness, and support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517500. URL <https://doi.org/10.1145/3491102.3517500>.
- Herbert H Clark. *Using Language*. Cambridge University Press, Cambridge, United Kingdom, 1996.
- Herbert H Clark and Susan E Brennan. Grounding in communication. In Lauren B. Resnick, Levine John M, and Stephanie D Teasley, editors, *Perspectives on Socially Shared Cognition*. American Psychological Association, 1991.
- Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*, 31(4):349–371, 09 2019a. ISSN 0953-5438. doi: 10.1093/iwc/iwz016. URL <https://doi.org/10.1093/iwc/iwz016>.

- Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proc. CHI 2019*, page 1–12, New York, 2019b. ACM. ISBN 9781450359702. doi: 10.1145/3290605.3300705. URL <https://doi.org/10.1145/3290605.3300705>.
- Kenneth Mark Colby, James B Watt, and John P Gilbert. A computer method of psychotherapy: Preliminary communication. *The Journal of Nervous and Mental Disease*, 142(2):148–152, 1966.
- Orla Cooney, Kevin Doherty, Marguerite Barry, David Coyle, Gavin Doherty, and Benjamin R. Cowan. Using speech agents for mood logging within blended mental healthcare: Mental healthcare practitioners’ perspectives. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI ’24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705113. doi: 10.1145/3640794.3665540. URL <https://doi.org/10.1145/3640794.3665540>.
- Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. “What Can I Help You with?”: Infrequent Users’ Experiences of Intelligent Personal Assistants. In *Proc. MobileHCI 17*, New York, 2017. ACM. ISBN 9781450350754. doi: 10.1145/3098279.3098539. URL <https://doi.org/10.1145/3098279.3098539>.
- David Coyle, Gavin Doherty, Mark Matthews, and John Sharry. Computers in talk-based mental health interventions. *Interacting with Computers*, 19(4):545–562, 2007. doi: 10.1016/j.intcom.2007.02.001. URL <https://doi.org/10.1016/j.intcom.2007.02.001>.
- Andrea Cuadra, David Goedicke, and J.D. Zamfirescu-Pereira. Democratizing design and fabrication using speech: Exploring co-design with a voice assistant. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, CUI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389983. doi: 10.1145/3469595.3469624. URL <https://doi.org/10.1145/3469595.3469624>.
- Carolyn E Cutrona and Julie A Suhr. Controllability of stressful events and satisfaction with spouse support behaviors. *Communication research*, 19(2):154–174, 1992.
- Alison Darcy, Jade Daniels, David Salinger, Paul Wicks, and Athena Robinson. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. *JMIR Formative Research*, 5(5):e27868, 2021. doi: 10.2196/27868. URL <https://doi.org/10.2196/27868>.
- Diana Deibel and Rebecca Evanhoe. *Conversations with Things: UX Design for Chat and Voice*. Rosenfeld, 2021.
- Gavin Doherty, David Coyle, and Mark Matthews. Design and evaluation guidelines for mental health technologies. *Interacting with computers*, 22(4):243–252, 2010. doi: 10.1016/j.intcom.2010.02.006. URL <https://doi.org/10.1016/j.intcom.2010.02.006>.

- Gavin Doherty, David Coyle, and John Sharry. Engagement with online mental health interventions: An exploratory clinical study of a treatment for depression. In *Proc. CHI 12*, page 1421–1430, New York, 2012. ACM. ISBN 9781450310154. doi: 10.1145/2207676.2208602. URL <https://doi.org/10.1145/2207676.2208602>.
- Kevin Doherty, Andreas Balaskas, and Gavin Doherty. The design of ecological momentary assessment technologies. *Interacting with Computers*, 32(1):257–278, 2020. doi: 10.1093/iwcomp/iwaa019. URL <https://doi.org/10.1093/iwcomp/iwaa019>.
- Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368254. doi: 10.1145/3338286.3340116. URL <https://doi.org/10.1145/3338286.3340116>.
- Philip R. Doyle, Leigh Clark, and Benjamin R. Cowan. What do we see in them? identifying dimensions of partner models for speech interfaces using a psycholexical approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445206. URL <https://doi-org.ucd.idm.oclc.org/10.1145/3411764.3445206>.
- Justin Edwards, He Liu, Tianyu Zhou, Sandy J. J. Gould, Leigh Clark, Philip Doyle, and Benjamin R. Cowan. Multitasking with alexa: how using intelligent personal assistants impacts language-based primary task performance. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450371872. doi: 10.1145/3342775.3342785. URL <https://doi.org/10.1145/3342775.3342785>.
- Elizabeth Victoria Eikey, Clara Marques Caldeira, Mayara Costa Figueiredo, Yunan Chen, Jessica L Borelli, Melissa Mazmanian, and Kai Zheng. Beyond self-reflection: introducing the concept of rumination in personal informatics. *Personal and Ubiquitous Computing*, 25(3):601–616, 2021. doi: 10.1007/s00779-021-01573-w. URL <https://doi.org/10.1007/s00779-021-01573-w>.
- 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*, UbiComp '15, page 731–742, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335744. doi: 10.1145/2750858.2804250. URL <https://doi.org/10.1145/2750858.2804250>.
- Amelia Fiske, Peter Henningsen, and Alena Buyx. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of medical Internet research*, 21(5):e13216, 2019. doi: 10.2196/13216. URL <https://doi.org/10.2196/13216>.

- Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785, 2017. doi: 10.2196/mental.7785. URL <https://doi.org/10.2196/mental.7785>.
- Asbjørn Følstad and Petter Bae Brandtzaeg. Chatbots and the new world of hci. *Interactions*, 24(4):38–42, June 2017. ISSN 1072-5520. doi: 10.1145/3085558. URL <https://doi.org/10.1145/3085558>.
- Jesse Fox and Andrew Gambino. Relationship development with humanoid social robots: Applying interpersonal theories to human–robot interaction. *Cyberpsychology, Behavior, and Social Networking*, 24(5):294–299, 2021. doi: 10.1089/cyber.2020.0181. URL <https://doi.org/10.1089/cyber.2020.0181>.
- Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, Michiel Rauws, et al. Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR mental health*, 5(4):e9782, 2018. doi: 10.2196/mental.9782. URL <https://doi.org/10.2196/mental.9782>.
- Future of Life Institute. High-level summary of the ai act. <https://artificialintelligenceact.eu/high-level-summary/>, 2024. Accessed: October 10th 2024.
- Erving Goffman. *Interaction ritual: Essays on face-to-face interaction*. Aldine, 1967.
- Jonathan Gratch, Gale M. Lucas, Aisha Aisha King, and Louis-Philippe Morency. It’s only a computer: the impact of human-agent interaction in clinical interviews. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS ’14, page 85–92, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450327381.
- Stephan Hammer, Birgit Lugrin, Sergey Bogomolov, Kathrin Janowski, and Elisabeth André. Investigating politeness strategies and their persuasiveness for a robotic elderly assistant. In *International Conference on Persuasive Technology*, pages 315–326, Cham, 2016. Springer International Publishing.
- Headspace. Headspace meditation app. <https://www.headspace.com/headspace-meditation-app>, 2024a. Accessed: September 12th 2024.
- Headspace. Meet ebb. <https://www.headspace.com/ai-mental-health-companion>, 2024b. Accessed: October 13th 2024.
- Kristin E Heron and Joshua M Smyth. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. *British journal of health psychology*, 15(1):1–39, 2010. doi: 10.1348/135910709X466063. URL <https://doi.org/10.1348/135910709X466063>.

- Tad Hirsch. Practicing without a license: Design research as psychotherapy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–11, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376750. URL <https://doi.org/10.1145/3313831.3376750>.
- Annabell Ho, Jeff Hancock, and Adam S Miner. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, 68(4):712–733, 2018. doi: 10.1093/joc/jqy026. URL <https://doi.org/10.1093/joc/jqy026>.
- Joyce Ho and Stephen S. Intille. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, page 909–918, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1581139985. doi: 10.1145/1054972.1055100. URL <https://doi.org/10.1145/1054972.1055100>.
- Leila Homaeian, James R. Wallace, and Stacey D. Scott. Joint Action Storyboards: A Framework for Visualizing Communication Grounding Costs. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27, 2021. ISSN 25730142. URL <https://dl.acm.org/doi/10.1145/3449102>.
- J House. *Work Stress and Social Support*. Addison-Wesley Publishing Company, 1981.
- J House, R Kahn, J McLeod, and D Williams. Measures and concepts of social support. In S Cohen and S Syme, editors, *Social Support and Health*, pages 83–108. Academic Press, NY, 1985.
- Yaxin Hu, Yuxiao Qu, Adam Maus, and Bilge Mutlu. Polite or direct? conversation design of a smart display for older adults based on politeness theory. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517525. URL <https://doi.org/10.1145/3491102.3517525>.
- Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106, 2018. doi: 10.2196/12106. URL <https://doi.org/10.2196/12106>.
- 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*, DIS '18, page 895–906, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351980. doi: 10.1145/3196709.3196735. URL <https://doi.org/10.1145/3196709.3196735>.
- Spike Jonze. *Her*. Annapurna Pictures, 2013.

- Jin Kang and Lewen Wei. ““Give me the support i want!”: The effect of matching an embodied conversational agent’s social support to users’ social support needs in fostering positive user-agent interaction. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, HAI ’18, page 106–113, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359535. doi: 10.1145/3284432.3284462. URL <https://doi.org/10.1145/3284432.3284462>.
- Christina Kelley, Bongshin Lee, and Lauren Wilcox. Self-tracking for mental wellness: Understanding expert perspectives and student experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI ’17, page 629–641, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025750. URL <https://doi.org/10.1145/3025453.3025750>.
- Da-jung Kim, Yeoreum Lee, Saeyoung Rho, and Youn-kyung Lim. Design opportunities in three stages of relationship development between users and self-tracking devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, page 699–703, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858148. URL <https://doi.org/10.1145/2858036.2858148>.
- Junhan Kim, Yoojung Kim, Byungjoon Kim, Sukyung Yun, Minjoon Kim, and Joongseek Lee. Can a machine tend to teenagers’ emotional needs? a study with conversational agents. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA ’18, page 1–6, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356213. doi: 10.1145/3170427.3188548. URL <https://doi.org/10.1145/3170427.3188548>.
- Soomin Kim, Joonhwan Lee, and Gahgene Gweon. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300316. URL <https://doi.org/10.1145/3290605.3300316>.
- Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. Mindfuldiary: Harnessing large language model to support psychiatric patients’ journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642937. URL <https://doi.org/10.1145/3613904.3642937>.
- 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*, DIS ’18, page 881–894, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351980. doi: 10.1145/3196709.3196784. URL <https://doi.org/10.1145/3196709.3196784>.

- William J Korotitsch and Rosemary O Nelson-Gray. An overview of self-monitoring research in assessment and treatment. *Psychological Assessment*, 11(4):415, 1999. doi: 10.1037/1040-3590.11.4.415. URL <https://doi.org/10.1037/1040-3590.11.4.415>.
- Theodora Koulouri, Robert D. Macredie, and David Olakitan. Chatbots to support young adults' mental health: An exploratory study of acceptability. *ACM Trans. Interact. Intell. Syst.*, 12(2), July 2022. ISSN 2160-6455. doi: 10.1145/3485874. URL <https://doi.org/10.1145/3485874>.
- Tanya Kraljic and Michal Lahav. From prompt engineering to collaborating: A human-centered approach to ai interfaces. *Interactions*, 31(3):30–35, May 2024. ISSN 1072-5520. doi: 10.1145/3652622. URL <https://doi.org/10.1145/3652622>.
- Kaylee Payne Kruzan, Ada Ng, Colleen Stiles-Shields, Emily G Lattie, David C. Mohr, and Madhu Reddy. The perceived utility of smartphone and wearable sensor data in digital self-tracking technologies for mental health. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581209. URL <https://doi.org/10.1145/3544548.3581209>.
- Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26, 2017. doi: 10.18637/jss.v082.i13.
- Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illeňčík, and Celeste Campos-Castillo. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, page 14614448221142007, 2022. doi: 10.1177/146144482211420. URL <https://doi.org/10.1177/146144482211420>.
- Robin Tolmach Lakoff. The limits of politeness: Therapeutic and courtroom discourse. *Multilingua*, 1989. doi: 10.1515/mult.1989.8.2-3.101. URL <https://doi.org/10.1515/mult.1989.8.2-3.101>.
- Emily G Lattie, Sarah Ketchen Lipson, and Daniel Eisenberg. Technology and college student mental health: challenges and opportunities. *Frontiers in Psychiatry*, 10:246, 2019. doi: 10.3389/fpsy.2019.00246. URL <https://doi.org/10.3389/fpsy.2019.00246>.
- Emily G. Lattie, Rachel Kornfield, Kathryn E. Ringland, Renwen Zhang, Nathan Winkquist, and Madhu Reddy. Designing mental health technologies that support the social ecosystem of college students. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376362. URL <https://doi.org/10.1145/3313831.3376362>.

- Jae-Gil Lee, Kwan Min Lee, and Seoung-Ho Ryu. Vehicle politeness in driving situations. *Future Internet*, 11(2):48, 2019. doi: 10.3390/fi11020048. URL <https://doi.org/10.3390/fi11020048>.
- Kwangyoung Lee and Hwajung Hong. Mindnavigator: Exploring the stress and self-interventions for mental wellness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174146. URL <https://doi.org/10.1145/3173574.3174146>.
- Yeoreum Lee, Jae-eul Bae, Sona S Kwak, and Myung-Suk Kim. The effect of politeness strategy on human-robot collaborative interaction on malfunction of robot vacuum cleaner. In *RSS workshop on HRI*, 2011.
- Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), May 2020a. doi: 10.1145/3392836. URL <https://doi.org/10.1145/3392836>.
- Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. “I hear you, i feel you”: Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020b. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376175. URL <https://doi.org/10.1145/3313831.3376175>.
- Ian Li, Anind K. Dey, and Jodi Forlizzi. Understanding my data, myself: Supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, page 405–414, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306300. doi: 10.1145/2030112.2030166. URL <https://doi.org/10.1145/2030112.2030166>.
- 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 *Proc. DIS 16*, page 264–275, New York, 2016. ACM. ISBN 9781450340311. doi: 10.1145/2901790.2901842. URL <https://doi.org/10.1145/2901790.2901842>.
- Irene Lopatovska, Olivia Turpin, Jessika Davis, Ellen Connell, Chris Denney, Hilda Fournier, Archana Ravi, Ji Hee Yoon, and Eesha Parasnis. Capturing teens’ voice in designing supportive agents. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450397391. doi: 10.1145/3543829.3543838. URL <https://doi.org/10.1145/3543829.3543838>.
- 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. doi: 10.1016/j.chb.2014.04.043. URL <https://doi.org/10.1016/j.chb.2014.04.043>.
- Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, 4:51, 2017. doi: 10.3389/frobt.2017.00051. URL <https://doi.org/10.3389/frobt.2017.00051>.
- 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*, CHI ’16, page 5286–5297, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858288. URL <https://doi.org/10.1145/2858036.2858288>.
- Zilin Ma, Yiyang Mei, and Zhaoyuan Su. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1105. American Medical Informatics Association, 2023.
- Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z. Gajos. Evaluating the experience of lgbtq+ people using large language model based chatbots for mental health support. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642482. URL <https://doi.org/10.1145/3613904.3642482>.
- Wookjae Maeng and Joonhwan Lee. Designing and evaluating a chatbot for survivors of image-based sexual abuse. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517629. URL <https://doi-org.ucd.idm.oclc.org/10.1145/3491102.3517629>.
- Raju Maharjan, Darius Adam Rohani, Per Bækgaard, Jakob Bardram, and Kevin Doherty. Can we talk? design implications for the questionnaire-driven self-report of health and wellbeing via conversational agent. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389983. doi: 10.1145/3469595.3469600. URL <https://doi.org/10.1145/3469595.3469600>.
- Raju Maharjan, Kevin Doherty, Darius Adam Rohani, Per Bækgaard, and Jakob E. Bardram. Experiences of a speech-enabled conversational agent for the self-report of well-being among people living with affective disorders: An in-the-wild study. *ACM Trans. Interact. Intell. Syst.*, 12(2), July 2022. ISSN 2160-6455. doi: 10.1145/3484508. URL <https://doi.org/10.1145/3484508>.
- Gin S Malhi, Amber Hamilton, Grace Morris, Zola Mannie, Pritha Das, and Tim Outhred. The promise of digital mood tracking technologies: are we heading on the right track?

- BMJ Ment Health*, 20(4):102–107, 2017. doi: 10.1136/eb-2017-102757. URL <https://doi.org/10.1136/eb-2017-102757>.
- Tanya Malik, Adrian Jacques Ambrose, and Chaitali Sinha. Evaluating user feedback for an artificial intelligence-enabled, cognitive behavioral therapy-based mental health app (wysa): Qualitative thematic analysis. *JMIR Human Factors*, 9(2):e35668, 2022. doi: 10.2196/35668. URL <https://doi.org/10.2196/35668>.
- Mark Matthews and Gavin Doherty. In the mood: Engaging teenagers in psychotherapy using mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’11, page 2947–2956, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979379. URL <https://doi.org/10.1145/1978942.1979379>.
- Mark Matthews, Gavin Doherty, John Sharry, and Carol Fitzpatrick. Mobile phone mood charting for adolescents. *British Journal of Guidance & Counselling*, 36(2): 113–129, 2008. doi: 10.1080/03069880801926400. URL <https://doi.org/10.1080/03069880801926400>.
- Mark Matthews, Elizabeth Murnane, and Jaime Snyder. Quantifying the changeable self: The role of self-tracking in coming to terms with and managing bipolar disorder. *Human-Computer Interaction*, 32(5-6):413–446, 2017. doi: 10.1080/07370024.2017.1294983. URL <https://doi.org/10.1080/07370024.2017.1294983>.
- Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. doi: 10.1145/3359174. URL <https://doi.org/10.1145/3359174>.
- Michael McTear and Marina Ashurkina. *Transforming Conversational AI*. Springer, 2024.
- Michael McTear, Zoraida Callejas, and David Griol. *The Conversational Interface*. Springer, 2016.
- Evonne Miller and Debra Polson. Apps, avatars, and robots: The future of mental health-care. *Issues in mental health nursing*, 40(3):208–214, 2019. doi: 10.1080/01612840.2018.1524535. URL <https://doi.org/10.1080/01612840.2018.1524535>.
- Adam S. Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA Internal Medicine*, 176(5):619–625, 05 2016. ISSN 2168-6106. doi: 10.1001/jamainternmed.2016.0400. URL <https://doi.org/10.1001/jamainternmed.2016.0400>.
- Adam S Miner, Nigam Shah, Kim D Bullock, Bruce A Arnow, Jeremy Bailenson, and Jeff Hancock. Key considerations for incorporating conversational ai in psychotherapy. *Frontiers in psychiatry*, 10:746, 2019. doi: 10.3389/fpsy.2019.00746. URL <https://doi.org/10.3389/fpsy.2019.00746>.

- Tomoki Miyamoto, Daisuke Katagami, Yuka Shigemitsu, Mayumi Usami, Takahiro Tanaka, Hitoshi Kanamori, Yuki Yoshihara, and Kazuhiro Fujikake. Influence of social distance expressed by driving support agent's utterance on psychological acceptability. *Frontiers in Psychology*, 12:254, 2021. doi: 10.3389/fpsyg.2021.526942. URL <https://doi.org/10.3389/fpsyg.2021.526942>.
- Andrew Monk. *Common Ground in Electronically Mediated Communication: Clark's Theory of Language Use*, pages 265–289. Morgan Kaufmann, San Francisco, CA, USA, 2003.
- 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. doi: 10.2196/10148. URL <https://doi.org/10.2196/10148>.
- Cecily Morrison, Edward Cutrell, Martin Grayson, Anja Thieme, Alex Taylor, Geert Roumen, Camilla Longden, Sebastian Tschischek, Rita Faia Marques, and Abigail Sellen. Social sensemaking with ai: Designing an open-ended ai experience with a blind child. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445290. URL <https://doi.org/10.1145/3411764.3445290>.
- Kellie Morrissey and Jurek Kirakowski. 'realness' in chatbots: establishing quantifiable criteria. In *Human-Computer Interaction. Interaction Modalities and Techniques: 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part IV 15*, pages 87–96. Springer, 2013. doi: 10.1007/978-3-642-39330-3_10. URL https://doi.org/10.1007/978-3-642-39330-3_10.
- Elizabeth L Murnane, Dan Cosley, Pamara Chang, Shion Guha, Ellen Frank, Geri Gay, and Mark Matthews. Self-monitoring practices, attitudes, and needs of individuals with bipolar disorder: implications for the design of technologies to manage mental health. *Journal of the American Medical Informatics Association*, 23(3):477–484, 01 2016. ISSN 1067-5027. doi: 10.1093/jamia/ocv165. URL <https://doi.org/10.1093/jamia/ocv165>.
- Elizabeth L. Murnane, Tara G. Walker, Beck Tench, Stephen Volda, and Jaime Snyder. Personal informatics in interpersonal contexts: Towards the design of technology that supports the social ecologies of long-term mental health management. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018. doi: 10.1145/3274396. URL <https://doi.org/10.1145/3274396>.
- Ada Ng, Rachel Kornfield, Stephen M. Schueller, Alyson K. Zalta, Michael Brennan, and Madhu Reddy. Provider perspectives on integrating sensor-captured patient-generated data in mental health care. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. doi: 10.1145/3359217. URL <https://doi.org/10.1145/3359217>.

- Bruna Oewel, Patricia Anne Arean, and Elena Agapie. Approaches for tailoring between-session mental health therapy activities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642856. URL <https://doi.org/10.1145/3613904.3642856>.
- Stefan Olafsson, Everlyne Kimani, Reza Asadi, and Timothy Bickmore. That's a rap. In *International Conference on Intelligent Virtual Agents*, pages 325–334. Springer, 2017. doi: 10.1007/978-3-319-67401-8_41. URL https://doi.org/10.1007/978-3-319-67401-8_41.
- Stefan Olafsson, Teresa K. O'Leary, and Timothy W. Bickmore. Motivating health behavior change with humorous virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375863. doi: 10.1145/3383652.3423915. URL <https://doi.org/10.1145/3383652.3423915>.
- Teresa K. O'Leary, Dhaval Parmar, Stefan Olafsson, Michael Paasche-Orlow, Timothy Bickmore, and Andrea G Parker. Community dynamics in technospiritual interventions: Lessons learned from a church-based mhealth pilot. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517700. URL <https://doi.org/10.1145/3491102.3517700>.
- OpenAI. Chatgpt. <https://openai.com/chatgpt/>, 2024. Accessed: September 12th 2024.
- SoHyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. “I wrote as if I were telling a story to someone I knew.”: Designing chatbot interactions for expressive writing in mental health. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, page 926–941, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384766. doi: 10.1145/3461778.3462143. URL <https://doi.org/10.1145/3461778.3462143>.
- Sun Young Park. Social support mosaic: Understanding mental health management practice on college campus. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, page 121–133, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351980. doi: 10.1145/3196709.3196787. URL <https://doi.org/10.1145/3196709.3196787>.
- Dhaval Parmar, Stefan Olafsson, Dina Utami, and Timothy Bickmore. Looking the part: The effect of attire and setting on perceptions of a virtual health counselor. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, page 301–306, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360135. doi: 10.1145/3267851.3267915. URL <https://doi.org/10.1145/3267851.3267915>.

- James W Pennebaker and Cindy K Chung. Expressive writing: Connections to physical and mental health. In Howard S Friedman, editor, *The Oxford Handbook of Health Psychology*, pages 417–437. Oxford University Press, Oxford, United Kingdom, 2011.
- Dorian Peters, Karina Vold, Diana Robinson, and Rafael A Calvo. Responsible ai—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, 1(1):34–47, 2020. doi: 10.1109/TTS.2020.2974991. URL <https://doi.org/10.1109/TTS.2020.2974991>.
- Lara Schibelsky Godoy Piccolo, Pinelopi Troullinou, and Harith Alani. Chatbots to support children in coping with online threats: Socio-technical requirements. In *Designing Interactive Systems Conference 2021*, DIS ’21, page 1504–1517, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384766. doi: 10.1145/3461778.3462114. URL <https://doi-org.ucd.idm.oclc.org/10.1145/3461778.3462114>.
- Laura R. Pina, Sang-Wha Sien, Teresa Ward, Jason C. Yip, Sean A. Munson, James Fogarty, and Julie A. Kientz. From personal informatics to family informatics: Understanding family practices around health monitoring. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW ’17, page 2300–2315, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343350. doi: 10.1145/2998181.2998362. URL <https://doi.org/10.1145/2998181.2998362>.
- Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174214. URL <https://doi.org/10.1145/3173574.3174214>.
- Kyrill Potapov, Asimina Vasalou, Victor Lee, and Paul Marshall. What do teens make of personal informatics? young people’s responses to self-tracking practices for self-determined motives. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445239. URL <https://doi.org/10.1145/3411764.3445239>.
- Ashish Viswanath Prakash and Saini Das. Intelligent conversational agents in mental healthcare services: a thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems*, 12(2):1, 2020. doi: 10.17705/1pais.12201. URL <https://doi.org/10.17705/1pais.12201>.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL <https://www.R-project.org/>.
- Amon Rapp, Lorenzo Curti, and Arianna Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based

- chatbots. *International Journal of Human-Computer Studies*, 151:102630, 2021. doi: 10.1016/j.ijhcs.2021.102630. URL <https://doi.org/10.1016/j.ijhcs.2021.102630>.
- Amon Rapp, Arianna Boldi, Lorenzo Curti, Alessandro Perrucci, and Rossana Simone. Collaborating with a text-based chatbot: An exploration of real-world collaboration strategies enacted during human-chatbot interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580995. URL <https://doi.org/10.1145/3544548.3580995>.
- Lazlo Ring, Dina Utami, and Timothy Bickmore. The right agent for the job? In *International Conference on Intelligent Virtual Agents*, pages 374–384. Springer, 2014. doi: 10.1007/978-3-319-09767-1_49. URL https://doi.org/10.1007/978-3-319-09767-1_49.
- John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers. Personal tracking as lived informatics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 1163–1172, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450324731. doi: 10.1145/2556288.2557039. URL <https://doi.org/10.1145/2556288.2557039>.
- John Rooksby, Alistair Morrison, and Dave Murray-Rust. Student perspectives on digital phenotyping: The acceptability of using smartphone data to assess mental health. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300655. URL <https://doi.org/10.1145/3290605.3300655>.
- Darley Sackitey, Teresa K. O’Leary, Michael Paasche-Orlow, Timothy Bickmore, and Andrea G Parker. “Everyone is covered”: Exploring the role of online interactions in facilitating connection and social support in black churches. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581324. URL <https://doi.org/10.1145/3544548.3581324>.
- Sophia Sakel, Tabea Blenk, Albrecht Schmidt, and Luke Haliburton. The social journal: Investigating technology to support and reflect on social interactions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642411. URL <https://doi.org/10.1145/3613904.3642411>.
- Maha Salem, Micheline Ziadee, and Majd Sakr. Effects of politeness and interaction context on perception and experience of hri. In *International Conference on Social Robotics*, pages 531–541. Springer, 2013.
- Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök,

- and Corina Sas. HCI and Affective Health: Taking Stock of a Decade of Studies and Charting Future Research Directions. In *Proc CHI 19*, page 1–17, New York, 2019. ACM. ISBN 9781450359702. doi: 10.1145/3290605.3300475. URL <https://doi.org/10.1145/3290605.3300475>.
- Jessica Schroeder, Chelsey Wilkes, Kael Rowan, Arturo Toledo, Ann Paradiso, Mary Czerwinski, Gloria Mark, and Marsha M. Linehan. Pocket skills: A conversational mobile web app to support dialectical behavioral therapy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–15, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173972. URL <https://doi.org/10.1145/3173574.3173972>.
- Stephen M Schueller, Adrian Aguilera, and David C Mohr. Ecological momentary interventions for depression and anxiety. *Depression and anxiety*, 34(6):540–545, 2017. doi: 10.1002/da.22649. URL <https://doi.org/10.1002/da.22649>.
- Stephen M Schueller, Martha Neary, Jocelyn Lai, and Daniel A Epstein. Understanding people’s use of and perspectives on mood-tracking apps: interview study. *JMIR mental health*, 8(8):e29368, 2021. doi: 10.2196/29368. URL <https://doi.org/10.2196/29368>.
- Jana Sedlakova and Manuel Trachsel. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? *The American Journal of Bioethics*, 23(5): 4–13, 2023. doi: 10.1080/15265161.2022.2048739. URL <https://doi.org/10.1080/15265161.2022.2048739>.
- Helen Sharp, Yvonne Rogers, and Jennifer Preece. *Interaction Design: Beyond Human-Computer Interaction*. Wiley, 5 edition, 2019.
- Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008. doi: 10.1146/annurev.clinpsy.3.022806.091415. URL <https://doi.org/10.1146/annurev.clinpsy.3.022806.091415>.
- Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149:102601, 2021. doi: 10.1016/j.ijhcs.2021.102601. URL <https://doi.org/10.1016/j.ijhcs.2021.102601>.
- Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168:102903, 2022. doi: 10.1016/j.ijhcs.2022.102903. URL <https://doi.org/10.1016/j.ijhcs.2022.102903>.
- Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. A longitudinal study of self-disclosure in human–chatbot relationships. *Interacting with Computers*, 35(1):24–39, 2023. doi: 10.1093/iwc/iwad022. URL <https://doi.org/10.1093/iwc/iwad022>.
- Snap Inc. Snapchat. https://play.google.com/store/apps/details?id=com.snapchat.android&hl=en_US, 2024. Accessed: September 10th 2024.

- Robert Soden, Austin Toombs, and Michaelanne Thomas. Evaluating interpretive research in hci. *Interactions*, 31(1):38–42, January 2024. ISSN 1072-5520. doi: 10.1145/3633200. URL <https://doi.org/10.1145/3633200>.
- Marie Louise Juul Søndergaard and Lone Koefoed Hansen. Intimate futures: Staying with the trouble of digital personal assistants through design fiction. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, page 869–880, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351980. doi: 10.1145/3196709.3196766. URL <https://doi.org/10.1145/3196709.3196766>.
- Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. How to do better with gender on surveys: a guide for hci researchers. *Interactions*, 26(4):62–65, June 2019. ISSN 1072-5520. doi: 10.1145/3338283. URL <https://doi.org/10.1145/3338283>.
- Vasant Srinivasan and Leila Takayama. Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4945–4955, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858217. URL <https://doi.org/10.1145/2858036.2858217>.
- Colm Sweeney, Courtney Potts, Edel Ennis, Raymond Bond, Maurice D. Mulvenna, Siobhan O'Neill, Martin Malcolm, Lauri Kuosmanen, Catrine Kostenius, Alex Vakaloudis, Gavin Mcconvey, Robin Turkington, David Hanna, Heidi Nieminen, Anna-Kaisa Vartiainen, Alison Robertson, and Michael F. Mctear. Can chatbots help support a person's mental health? perceptions and views from mental healthcare professionals and experts. *ACM Trans. Comput. Healthcare*, 2(3), July 2021. doi: 10.1145/3453175. URL <https://doi.org/10.1145/3453175>.
- Vivian Ta, Caroline Griffith, Carolyn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, Alexia Loggarakis, et al. User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of medical Internet research*, 22(3):e16235, 2020. doi: 10.2196/16235. URL <https://doi.org/10.2196/16235>.
- Ella Tallyn, Hector Fried, Rory Gianni, Amy Isard, and Chris Speed. The ethnobot: Gathering ethnographies in the age of iot. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174178. URL <https://doi.org/10.1145/3173574.3174178>.
- Shelley E Taylor. Social support: A review. In Howard S Friedman, editor, *The Oxford Handbook of Health Psychology*, volume 1, pages 189–214. OUP, 2011.
- Kazunori Terada, Mitsuki Okazoe, and Jonathan Gratch. Effect of politeness strategies in dialogue on negotiation outcomes. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, IVA '21, page 195–202, New York,

- NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386197. doi: 10.1145/3472306.3478336. URL <https://doi.org/10.1145/3472306.3478336>.
- Anja Thieme, Jayne Wallace, Thomas D. Meyer, and Patrick Olivier. Designing for mental wellbeing: towards a more holistic approach in the treatment and prevention of mental illness. In *Proceedings of the 2015 British HCI Conference*, British HCI '15, page 1–10, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336437. doi: 10.1145/2783446.2783586. URL <https://doi.org/10.1145/2783446.2783586>.
- Anja Thieme, Danielle Belgrave, and Gavin Doherty. Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems. *ACM Trans. Comput.-Hum. Interact.*, 27(5), aug 2020a. ISSN 1073-0516. doi: 10.1145/3398069. URL <https://doi.org/10.1145/3398069>.
- Anja Thieme, Ed Cutrell, Cecily Morrison, Alex Taylor, and Abigail Sellen. Interpretability as a dynamic of human-ai interaction. *Interactions*, 27(5):40–45, sep 2020b. ISSN 1072-5520. doi: 10.1145/3411286. URL <https://doi.org/10.1145/3411286>.
- Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. Designing human-centered ai for mental health: Developing clinically relevant applications for online cbt treatment. *ACM Trans. Comput.-Hum. Interact.*, 30(2), March 2023. ISSN 1073-0516. doi: 10.1145/3564752. URL <https://doi.org/10.1145/3564752>.
- Cristen Torrey, Susan R Fussell, and Sara Kiesler. How a robot should give advice. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 275–282. IEEE, 2013. doi: 10.1109/HRI.2013.6483599. URL <https://doi.org/10.1109/HRI.2013.6483599>.
- Touchkin. Wysa: Anxiety, therapy chatbot. <https://play.google.com/store/apps/details?id=bot.touchkin>, 2024. Accessed: September 10th 2024.
- S Turkle. *Alone together: Why we expect more from technology and less from each other*. Basic Books, 2011.
- R Jay Turner and Robyn Lewis Brown. Social support and mental health. In Teresa L Scheid and Tony N Brown, editors, *A Handbook for the Study of Mental Health: Social Contexts, Theories, and Systems*, volume 2, pages 200–212. Cambridge University Press, 2010.
- Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kishan, and John Blake Torous. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7): 456–464, 2019. doi: 10.1177/0706743719828977. URL <https://doi.org/10.1177/0706743719828977>.

- Marloes MC van Wezel, Emmelyn AJ Croes, and Marjolijn L Antheunis. “i’m here for you”: Can social chatbots truly support their users? a literature review. In *Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4*, pages 96–113. Springer, 2021. doi: 10.1007/978-3-030-68288-0_7. URL https://doi.org/10.1007/978-3-030-68288-0_7.
- Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. Developing a personality model for speech-based conversational agents using the psycholexical approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376210. URL <https://doi-org.ucd.idm.oclc.org/10.1145/3313831.3376210>.
- Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. Eliciting and analysing users’ envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445536. URL <https://doi.org/10.1145/3411764.3445536>.
- Ning Wang, W Lewis Johnson, Richard E Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. The politeness effect : Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies*, 66:98–112, 2008. doi: 10.1016/j.ijhcs.2007.09.003. URL <https://doi.org/10.1016/j.ijhcs.2007.09.003>.
- Jing Wei, Tilman Dingler, and Vassilis Kostakos. Understanding user perceptions of proactive smart speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(4), 2022. doi: 10.1145/3494965. URL <https://doi.org/10.1145/3494965>.
- Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, page 214–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL <https://doi.org/10.1145/3531146.3533088>.
- Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966. ISSN 0001-0782. doi: 10.1145/365153.365168. URL <https://doi.org/10.1145/365153.365168>.

- Joseph Weizenbaum. *Computer power and human reason: From judgment to calculation*. WH Freeman & Co, 1976.
- Jobke Wentzel, Rosalie van der Vaart, Ernst T Bohlmeijer, and Julia E W C van Gemert-Pijnen. Mixing online and face-to-face therapy: How to benefit from blended care in mental health care. *JMIR Mental Health*, 3(1):e9, Feb 2016. ISSN 2368-7959. doi: 10.2196/mental.4534. URL <https://doi.org/10.2196/mental.4534>.
- Geoff White. Child advice chatbots fail to spot sexual abuse, Nov 2018. URL <https://www.bbc.com/news/technology-46507900>. Last accessed: February 8th 2023.
- Emily Widnall, Claire Ellen Grant, Tao Wang, Lauren Cross, Sumithra Velupillai, Angus Roberts, Robert Stewart, Emily Simonoff, Johnny Downs, et al. User perspectives of mood-monitoring apps available to young people: qualitative content analysis. *JMIR mHealth and uHealth*, 8(10):e18140, 2020. doi: 10.2196/18140. URL <https://doi.org/10.2196/18140>.
- Woebot Health Inc. Woebot: The mental health ally. <https://play.google.com/store/apps/details?id=com.woebot&hl=en-US>, 2024. Accessed: September 10th 2024.
- World Health Organisation. World mental health report: Transforming mental health for all. Technical report, World Health Organisation, 2022.
- Yunhan Wu, Justin Edwards, Orla Cooney, Anna Bleakley, Philip R. Doyle, Leigh Clark, Daniel Rough, and Benjamin R. Cowan. Mental workload and language production in non-native speaker ipa interaction. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375443. doi: 10.1145/3405755.3406118. URL <https://doi.org/10.1145/3405755.3406118>.
- Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Trans. Comput.-Hum. Interact.*, 27(3), June 2020. ISSN 1073-0516. doi: 10.1145/3381804. URL <https://doi.org/10.1145/3381804>.
- Tianling Xie and Iryna Pentina. Attachment theory as a framework to understand relationships with social chatbots: a case study of replika. In *Proceedings of the 55th Hawaii international conference on system sciences*, 2022. doi: 10.24251/HICSS.2022.258. URL <https://doi.org/10.24251/HICSS.2022.258>.
- Naomi Yamashita, Hideaki Kuzuoka, Keiji Hirata, and Takashi Kudo. Understanding the conflicting demands of family caregivers caring for depressed family members. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2637–2646, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2481365. URL <https://doi.org/10.1145/2470654.2481365>.

- Naomi Yamashita, Hideaki Kuzuoka, Keiji Hirata, Takashi Kudo, Eiji Aramaki, and Kazuki Hattori. Changing moods: How manual tracking by family caregivers improves caring and family communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 158–169, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025843. URL <https://doi.org/10.1145/3025453.3025843>.
- Xi Yang and Marco Aurisicchio. Designing conversational agents: A self-determination theory approach. In *Proc CHI 21*, New York, 2021. ACM. ISBN 9781450380966. doi: 10.1145/3411764.3445445. URL <https://doi.org/10.1145/3411764.3445445>.
- Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. Multimodal healthcare ai: Identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642013. URL <https://doi.org/10.1145/3613904.3642013>.
- J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581388. URL <https://doi.org/10.1145/3544548.3581388>.
- Shuo Zhou, Timothy Bickmore, Michael Paasche-Orlow, and Brian Jack. Agent-user concordance and satisfaction with a virtual hospital discharge nurse. In *International conference on intelligent virtual agents*, pages 528–541. Springer, 2014. doi: 10.1007/978-3-319-09767-1_63. URL https://doi.org/10.1007/978-3-319-09767-1_63.
- Shuo Zhou, Zhe Zhang, and Timothy Bickmore. Adapting a persuasive conversational agent for the chinese culture. In *2017 International Conference on Culture and Computing (Culture and Computing)*, pages 89–96. IEEE, 2017. doi: 10.1109/Culture.and.Computing.2017.42. URL <https://doi.org/10.1109/Culture.and.Computing.2017.42>.
- Sahba Zojaji, Christopher Peters, and Catherine Pelachaud. Influence of virtual agent politeness behaviors on how users join small conversational groups. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375863. doi: 10.1145/3383652.3423917. URL <https://doi.org/10.1145/3383652.3423917>.