공학박사 학위논문

# Designing Conversational Agents to Encourage User Narrative and Self-Reflection in Mental Wellbeing

정신건강에서 사용자 내러티브와 자아성찰을
지원하는 대화형 에이전트 디자인

**2020년 8월**

서울대학교 융합과학기술대학원
융합과학부 디지털정보융합전공

박소현

# Designing Conversational Agents to Encourage User Narrative and Self-Reflection in Mental Wellbeing

**Advisor: Bongwon Suh**

**Submitting a Ph.D. Dissertation of**
**Digital Contents and Information Studies**

**July 2020**

**Seoul National University**
**Graduate School of Convergence Science and Technology**

**SoHyun Park**

**Confirming the Ph.D. Dissertation written by**
**SoHyun Park**

**July 2020**

Chair 이 준 환 (Seal)

Vice Chair 서 봉 원 (Seal)

Examiner 이 원 종 (Seal)

Examiner 권 가 진 (Seal)

Examiner 이 선 희 (Seal)

# Abstract

In the advent of artificial intelligence (AI), we are surrounded by technological gadgets, devices and intelligent personal assistant (IPAs) that voluntarily take care of our home, work and social networks. They help us manage our life for the better, or at least that is what they are designed for. As a matter of fact, few are, however, designed to help us grapple with the thoughts and feelings that often construct our living. In other words, technologies hardly help us *think*. How can they be designed to help us reflect on ourselves for the better?

In the simplest terms, self-reflection refers to thinking deeply about oneself. When we think deeply about ourselves, there can be both positive and negative consequences. On the one hand, reflecting on ourselves can lead to a better self-understanding, helping us achieve life goals. On the other hand, we may fall into brooding and depression. The sad news is that the two are usually intertwined. The problem, then, is the irony that reflecting on oneself by oneself is not easy.

To tackle this problem, this work aims to design technology in the form of a conversational agent, or a chatbot, to encourage a positive self-reflection. Chatbots are natural language interfaces that interact with users in text. They work at the tip of our hands as if SMS or instant messaging, from flight reservation and online shopping to news service and healthcare. There are even chatbot therapists offering psychotherapy on mobile. That machines can now talk to us creates an opportunity for designing a natural interaction that used to be humans' own.

This work constructs a two-dimensional design space for translating self-reflection into a human-chatbot interaction, with user self-disclosure and chatbot guidance. Users confess their thoughts and feelings to the bot, and the bot is to guide them in the scaffolding process. Previous work has established an extensive line of research on the therapeutic effect of emotional disclosure. In HCI, reflection design has posited the need for guidance, e.g. scaffolding users' thoughts, rather than assuming their ability to reflect in a constructive manner.

The design space illustrates different reflection processes depending on the levels of user disclosure and bot guidance. Existing reflection technologies have most commonly provided minimal levels of disclosure and guidance, and healthcare technologies the opposite. It is the aim of this work to investigate the less explored space by designing chatbots called Bonobot and Diarybot. Bonobot differentiates itself from other bot interventions in that it only motivates the idea of change rather than direct engagement. Diarybot is designed in two chat versions, Basic and Responsive, which create novel interactions for reflecting on a difficult life experience by explaining it to and exploring it with a chatbot. These chatbots are set up for a user study with 30 participants, to investigate the user experiences of and responses to design strategies. Based on the findings, challenges and opportunities from designing for chatbot-guided reflection are explored.

The findings of this study are as follows. First, participants preferred Bonobot's questions that prompted the idea of change. Its responses were also appreciated, but only when they conveyed accurate empathy. Thus questions, coupled with empathetic responses, could serve as a catalyst for disclosure and even a possible change of behavior, a motivational boost. Yet the chatbot-led interaction led to surged user expectations for the bot. Participants demanded more than just the guidance, such as solutions and even superhuman intelligence. Potential tradeoff between user engagement and autonomy in designing human-AI partnership is discussed.

Unlike Bonobot, Diarybot was designed with less guidance to encourage users' own narrative making. In both Diarybot chats, the presence of a bot could make it easier for participants to share the most difficult life experiences, compared to a no-chatbot writing condition. Yet an increased interaction with the bot in Responsive chat could lead to a better user engagement. On the contrary, more emotional expressiveness and ease of writing were observed with little interaction in Basic chat. Coupled with qualitative findings that reveal user preference for varied interactions and tendency to adapt to bot patterns, predictability and transparency of designing chatbot interaction are discussed in terms of managing user expectations in human-AI interaction.

In sum, the findings of this study shed light on designing human-AI interaction. Chatbots can be a potential means of supporting guided disclosure on life's most difficult experiences. Yet the interaction between a machine algorithm and an innate human cognition bears interesting questions for the HCI community, especially in terms of user autonomy, interface predictability, and design transparency. Discussing the notion of algorithmic affordances in AI agents, this work proposes *meaning-making* as novel interaction design metaphor: In the symbolic interaction via language, AI nudges users, which inspires and engages users in their pursuit of making sense of life's agony. Not only does this metaphor respect user autonomy but also it maintains the veiled workings of AI from users for continued engagement.

This work makes the following contributions. First, it designed and implemented chatbots that can provide guidance to encourage user narratives in self-reflection. Next, it offers empirical evidence on chatbot-guided disclosure and discusses implications for tensions and challenges in design. Finally, this work proposes meaning-making as a novel design metaphor. It calls for the responsible design of intelligent interfaces for positive reflection in pursuit of psychological wellbeing, highlighting algorithmic affordances and interpretive process of human-AI interaction.

**Keywords:** Self-reflection; chatbot design; artificial intelligence; mental wellbeing; human-computer interaction

**Student Number: 2015-31344**

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1. Introduction

## 1.1. Background and Motivation

In May 2018, the world hailed at an “um-hmm” by a machine agent making a reservation for a woman’s haircut. Called Google Duplex, the system achieves phone call conversations for various human tasks [288]. This demo shows how close we have come to having real conversations with computers, a long-standing goal in human-computer interaction [288]. Natural language interfaces, e.g. Apple’s Siri, Microsoft Cortana and Google Assistant, enable interactions via talks-in-turn, to accomplish a number of everyday tasks either in voice or text, such as flight booking, online shopping and customer service. Moreover, Microsoft’s XiaoIce has recently made a chatbot phenomenon in China [289] for its engaging in social conversations with users, e.g. giving words of advice and pep talk.

Now conversational agents do our work, care for how we feel, and learn to be more human. What we still don’t have, however, is the agent that engages in what we think and how we feel, especially when it comes to ourselves. Neither Siri nor XiaoIce can refresh our memories and lead a meaning-making process. We make commands and feel attached to the agents, but we can’t learn and grow with them.

Self-reflection, or thinking deeply about oneself, has been widely studied and applied as a means of promoting self-awareness and self-understanding, as well as improving learning outcomes and achieving life goals. In education and learning, reflection is often regarded as a critical process to engage learners and trainees to review where they are and what they've learned, mull it over and evaluate it [27]. In business and leadership, leaders and entrepreneurs are encouraged to review their past successes and find room for improvement [210]. In healthcare, self-reflection is also important for clinicians and caretakers to examine their current clinical practice to review their actions, perceptions, motives and feelings toward the patient [98]. Finally, for many individuals, reflecting on past life events and thoughts can be beneficial for gaining self-insight and development [22].

Despite the positives, it is often difficult for individuals to take the *healthy* route of self-reflection on their own. In psychology, it has been pointed out that reflecting on oneself may lead to negative self-concepts for some, whilst others may benefit from the self-learning process [139]. Others discussed that people with brooding tendencies may not benefit from reflection [57]. It may turn into ruminating thoughts that lead to depression. The problem is that oftentimes reflection and rumination are a simultaneous process and it is difficult to discern between the two.

What are the ways in which technologies can support positive self-reflection? Reflection design has been one of the key research topics in human-computer interaction (HCI). With personal devices and gadgets, technologies record data and return it to the user for reflecting on their day, lifestyle, and health behavior (e.g. SenseCam [106], Affective Diary [251], and MirrorMirror [80]). They can also help retrieve past memories and rediscover new meanings (e.g. Pensieve [194] and Echo [116]). Most of these technologies wait to be picked up for serendipitous re-encounters of happy and triste memories.

Meanwhile, advanced natural language interfaces gave rise to the so-called "chatbot therapists," conversational agents that come in our way to help with mental health issues. Research suggests that these chatbots are effective in reducing depressive symptoms such as post-traumatic stress disorder [68,73,115]. While these agents can tap into an individual's negative emotional experiences, they are more focused on enhancing treatment efficacy as a cost-effective means of treating mental illnesses. In other words, there is a gap in technology that brings memory triggers and that gives a treatment.

This work aims to fill this gap by designing technology that engages users in their self-reflection as a "reflection companion," in the form of a conversational agent, or a chatbot. The bot is to help users talk about their negative life events or unresolved stress, and to provide guided prompts that can help scaffold their life stories, to gear the reflection toward a healthy route. Though there have been chatbots that talk to users about their problems, it is only a brief process toward treatment (e.g. Woebot [68], Wysa [115], Tess [73]), or an aimless conversation, as seen in commercial voice user interfaces (e.g. Amazon Alexa, Apple Siri, Google Assistant). The reflection assistant chatbot is to engage users in and lead a structured conversation in reflecting on stressful life events.

To design the chatbot, a design space needs to be constructed. This work proposes disclosure and guidance as key elements in constructing the design space. While self-reflection is an intrapersonal experience, a chatbot translates it into an interpersonal one, as in self-disclosure. The chatbot should be able to create a safe environment for telling stories, find deeper thoughts and confessing untold feelings. For such a process not to go astray, chatbot needs to provide appropriate guidance to scaffold the thinking process in a constructive manner. Put together, disclosure and guidance can construct a two-dimensional space with four different types of reflection processes, labeled as: revisiting, explaining, exploring and transformative. This space is illustrated in <Figure 1.1>.

In revisiting, both disclosure and guidance are low, and technologies prompt users often with artefacts of past memories. Greater disclosure and guidance lead to transformative reflection, where users are guided not only to tell a problem but also to actively seek ways to promote changes in behavior or lifestyle. Relatively less explored are explaining and exploring, where there are more disclosure and less guidance and vice versa. In explaining, users are more encouraged to tell their stories, as in what happened and how they felt. In contrast, in exploring, technologies can take it further to ask or challenge users to think about different aspects of the narrative.

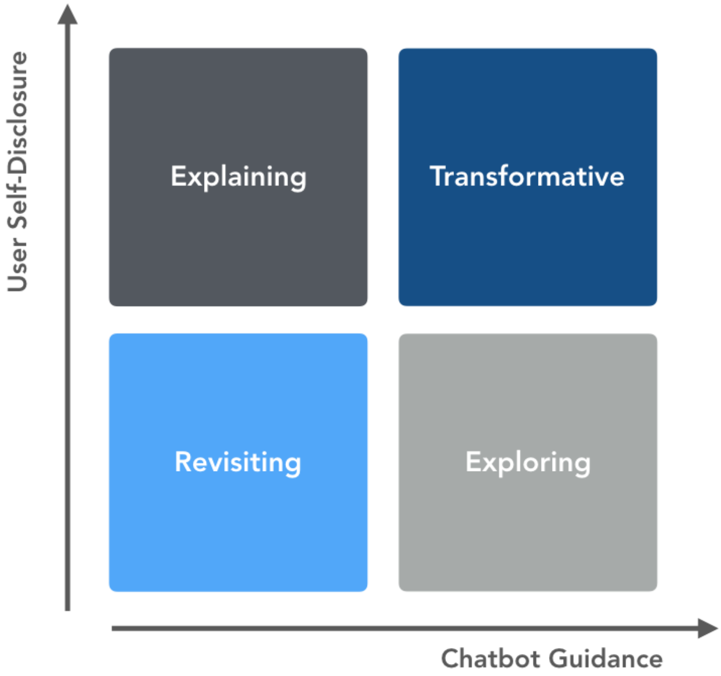


**Figure 1.1. The design space for reflection assistant chatbots.**

Many technologies that attempted to support user self-reflection are mostly found in the revisiting space (e.g. [80,106,155,193,209,251]). This work presents the design and implementation of two reflection assistant chatbots, Bonobot and Diarybot, which support the rest three reflection spaces: transformative, explaining, and exploring. Bonobot encourages users to talk about their stress and leads them to think about ways to cope with it. Diarybot offers two types of chats to encourage explaining and exploring reflections by either helping users write about their traumatic experiences in life or following up with it. In designing the two chatbots, this work is interested in investigating the user experience of reflection guided by the chatbot, their responses to the disclosure and guidance design, and their experience of chatbot-guided disclosure that may promote or challenge the existing notions in designing human-AI interaction.

## 1.2. Research Goal and Questions

### 1.2.1. Research Goal

The goal of this research is to design reflection assistant chatbots to support user disclosure and provide guidance to scaffold the process. The bots will be set up for an empirical investigation of: (a) how the chatbot design can encourage users' self-reflection; (b) how they respond to the design strategies; and (c) how the design may further inform AI-guided reflection. Each is discussed in more detail below.

### 1.2.2. Research Questions

#### Question 1. How do users experience the chatbot-guided reflection?

In this study, chatbots take the role of a "reflection partner" [169] that asks users to think and write about unresolved stress or difficult life experiences. It carries a conversation about them for a further reflection.

How would users respond to and engage in the chatbot-guided reflection process? In similar studies, Pensieve [194] users liked to reminisce and write about their past experiences. Echo [116] users who both recorded and reflected on the past experiences improved in their mental wellbeing by savoring positive emotions and drawing lessons from negative events. These systems allow users to engage in a dialogue with themselves [169], an *intrapersonal* experience. Similarly but differently, in this study users are encouraged to engage in a dialogue with chatbot assistants, now an *interpersonal* experience. Investigating how users experience this process will also lead to answer how they perceive the chatbots, how they form their narratives with them, and what their needs and expectations are in the chatbot-guided reflection.

### Question 2. How do users respond to the design strategies for disclosure and guidance?

In this work, chatbot-guided reflection incorporates two dimensions: disclosure and guidance. Chatbots are designed with differing levels of user self-disclosure and bot guidance, in order to support different types of reflection. The main interest is how such design would work, and what impact it would have on users. From literature, chatbots may very well support self-disclosure: Disclosure is a social exchange process [281], and chatbots are perceived as social actors [105,180]. Yet this work takes it further: If chatbots support the social sharing of emotions [219], can they support the cognitive processing of emotions as well? Rimé [219] states that a full recovery of negative emotional experiences often accompanies social sharing of emotions as well as of a meaning-making process. This work is interested in whether chatbots can take this role.

Moreover, this work is also interested in designing the guidance for scaffolding the reflection processes. Earlier work has argued that in a dialogue interaction, guided prompts can support deeper reflection [247] and even therapeutic effect [188]. However, less is explored on how the guides need to be designed and how users may respond to the design of the prompts. Specifically, unlike directive guided prompts that engage users in a treatment [68,115] (e.g. breathing, writing or thought exercises), non-directive guided prompts (e.g. open-ended questions) can invite a sharing of more spontaneous thoughts and feelings. Investigating user responses to the disclosure and guidance design will help garner practical implications and design guidelines for future reflection design.

### Question 3. How does the chatbot-guided reflection inform the design of AI-guided reflection?

In a broader context, engaging chatbots in self-reflection on life's most difficult experiences involves human-AI collaboration in a meaning-making process. Such an interaction concerns tensions and tradeoffs that may arise from the interdependent relationship, between bots supporting reflection and users reflecting on machine-generated guidance. How does involving chatbots change user expression, engagement and expectations? In a chatbot conversation, users are hidden from the exact workings of how the bot responds to them. Also, bots usually take the lead in order to prevent conversation failures and manage user expectations. However, we've seen from the previous research that users would like to take the initiative in human-AI collaboration [186]. It is also advised that rather than simply labeling the user to take the lead, communication should reach agreement to let the user literally take control of making decisions [186]. Now, questions arise from having chatbots assist self-reflection.

How do users feel control? How are user narratives affected? What are potential tradeoffs? How can chatbots ensure engagement and offer better assistance? Answering the questions will help define critical aspects of designing the future human-AI interaction.

## 1.3. Major Contributions

This work makes the following contributions. First, it successfully presents the design and implementation of reflection assistant chatbots and provides evidence for chatbot-guided reflection from user study. The findings indicate that users like to engage with chatbots for self-reflection, especially for social sharing of emotions and gaining new perspectives. They are willing to disclose the innermost thoughts and feelings about their significant life experiences to a nonhuman agent for its nonhuman- and human-like features. That chatbots are not human makes it easier for users to share some of the most private aspects of their life that had never been told. Also, the human-like qualities, such as asking questions and giving empathetic feedback, though programmed, were favored for discovering new insights and feeling understood.

In addition, this work offers practical implications for chatbot-guided disclosure. Both Bonobot and Diarybot provided guidance in the form of open-ended and directive questions, which effectively served to manage the flow of conversation and create a stepwise narrative to prompt users to think further. Moreover, maintaining contextual understanding by retrieving relevant user keywords in the bot responses was an effective strategy to encourage user engagement. For Bonobot, words of accurate empathy, rather than empty words of encouragement, could build trust and lead to sharing further. Instead of visual aids, i.e. images and videos, this study shows that designing chatbots requires a careful conversation design, which can be a powerful strategy to shape user narrative.

Finally, this work offers an empirical understanding of algorithmic affordance in human-AI interaction. As they actively engage in reflection, chatbots shape user perceptions and users form expectations around the process. The interaction constantly tests the boundaries: Knowing what the chatbot will say might wane user engagement, whereas unexpected chatbot behavior may stretch it too far and fail user expectations. In this study, Bonobot led users to think that it had some "intelligence" to solve their problems, an example of heightened user expectations when the bot workings are not revealed. They speculated and made assumptions about the bot, which was even more visible in Diarybot. Some actively adapted to Diarybot behavior, and how they perceived its workings influenced their engagement. This calls for an in-depth discussion on algorithmic affordance of AI-guided interaction. For designers and users alike, there can be tensions and tradeoffs as designers need to rethink transparency and interpretability of human-AI interaction, and what impact it may make on users on their autonomy and expectations in their engagement. It bears much importance for HCI researchers in the advent of general artificial intelligence, as both challenges and opportunities lie ahead. Taken together, this work calls for the responsible design of reflection technologies, particularly for intelligent agents that help us think, learn and grow from life's toughest lessons.

## 1.4. Thesis Overview

The rest of the thesis is organized as follows. Chapter 2 first reviews literature to discuss what self-reflection is and what it means for mental wellbeing in order to illustrate the design space. A survey of reflection technologies follows, to find design opportunities. Both theoretical and technical background study yields strategies for the reflection assistant chatbots to be designed in this study. Chapter 3 describes Bonobot, a chatbot that encourages a transformative reflection. The goal and design decisions, along with a qualitative user study and discussion will follow. Chapter 4 introduces Diarybot, a chatbot that helps users explain and explore difficult life experiences. It describes goals and design decisions, as well as an experimental user study and its findings, with discussion on implications. Chapter 5 summarizes the findings gleaned from the previous two chapters and provides an in-depth discussion on the design of supporting guided disclosure with conversational agents. It also gives a general discussion on the broader implications of this study in human-AI interaction. Chapter 6 concludes this thesis, reviewing limitations and illustrating future work, along with some final remarks.

# Chapter 2. Literature Review

This section surveys previous research on self-reflection and related works on technologies designed for reflection. First, an overview of research on self-reflection and how is associated with self-rumination is presented. Then, the popular and long-established practice of expressive writing is discussed as a self-oriented reflection for mental wellbeing. Reviewing the literature, a design space for chatbot-guided reflection is illustrated with guidance and disclosure as necessary conditions. To design chatbots, an array of related technologies is reviewed find niche. Finally, this section concludes with theoretical and technical background as well as design strategies for chatbots as reflection assistants.

## 2.1. The Reflecting Self

Self-reflection has been a complex concept in psychology. It involves a variety of psychological and emotional processes that may or may not contribute to mental health. Reviewing related literature, emotional disclosure and guidance are suggested as two conditions for constructive reflection, in an effort to ensure consistent benefits of self-reflection.

### 2.1.1. Self-Reflection and Mental Wellbeing

Self-reflection has been widely studied with rather inconsistent results on health outcomes. Researchers have identified two broad paths of self-reflection: reflective and ruminative. While the former contributes to wellbeing, the latter undermines. Yet it has been difficult to dissociate the two processes as they often take place simultaneously.

#### 2.1.1.1. What is Self-Reflection?

Self-reflection commonly refers to thinking about oneself in order to achieve more conscious knowledge and control of oneself and one's actions [69]. Early thinkers have defined reflection as an acquisition of attitudes and skills in thinking [36], or a process of critical self-determination [92]. Alternatively, the more familiar approach to reflection was developed by Donald Schön on his *Reflective Practitioner* [236,237]. According to him, reflection is a spontaneous process of framing and reframing in one's professional practice [26,236]. It is a process of becoming aware of the influence of societal and ideological assumptions, especially ethical and moral beliefs, behind professional practice [284].

Then how does reflection occur? Rolfe [222] suggested three stages. First, the individual attends to the thoughts and feelings aroused by an event. Then he or she reevaluates his or her experience of this event. Finally, the individual may generate new insights or perspectives from his or her reflection [222]. Likewise, Moon [170] affirmed in reflection one draws on a past experience, reflects on it in the present and uses it to inform future practice. Atkins and Murphy [6] described how reflection may be triggered by an awareness of uncomfortable feelings and thoughts. Individuals' personal experience, feelings and cognition are intermingled in recalling the past events, resolving current difficulties, composing uncomfortable feelings, evaluating one's present and past and searching for new perspectives and solutions [284].

#### 2.1.1.2. Does Self-Reflection Work?

The broad coverage of self-reflection has made it malleable yet also difficult to define its boundaries, specifically in terms of its processes and outcomes. Earlier in the days, increased self-focused attention was believed to be positively associated with depression [161,182,279,280]. However, Hixon and Swann Jr [104] suggested otherwise. They conducted four experiments to test their hypotheses on the accuracy of social feedback on self; the agreement of self- and social-appraisals; the conditions on which sound self-insight can be promoted in reflection; and the duration of self-reflection to ensure benefit. The results indicate that self-reflection could lead to positive outcomes in self-insight, when one can accurately evaluate the social feedback from others; when their self-appraisals agree with those of others; when one focuses on what one is, not why one is; and when the opportunity to reflect contribute enough to enhance self-knowledge. One important takeaway from this work is that unlike previous studies that bear skepticism, self-reflection can actually be beneficial, when strong and unambiguous pieces of self-knowledge are reflected on with a focus on what, rather than why.

While Hixon and Swann Jr's work focused on when and how self-focused attention may contribute to greater self-knowledge, it was in a few years' time that more concrete routes of self-attention were identified. In their work, Trapnell and Campbell studied the association of private self-consciousness and five factor model of personality [267]. Building on an earlier work on private self-consciousness [67] that established a fundamental dichotomy in self-perception of public and private self, they suggest the private self-consciousness scale has confounded two distinct motivational dispositions in self-focus: rumination and reflection. They argue that rumination offers a conception of self-attentiveness motivated by perceived threats, losses or injustices to the self. On the other hand, reflection provides a conception of self-attentiveness from curiosity and epistemic interest in the self [267]. In terms of the five-factor model of personality, rumination is correlated with neuroticism; and reflection openness.

Since then, the dichotomy between fear and curiosity to differentiate rumination from reflection has reigned in self-reflection research. Given this view, the term "reflection" usually implies the positive route to greater self-insight. Theoretically, self-reflection can encourage self-knowledge and enhances mental health [261]. Still, various studies have presented rather confounding outcomes of reflective thinking. Some support the adaptive function in that it is related to forms of coping, such as problem solving or distraction [35], and less depression [268]. Others found that suicidal and non-suicidal groups differed in terms of the levels of reflection, indicating that less reflection is linked to suicidality [51]. However, those supporting maladaptive outcomes of reflection argue that it is positively correlated with depression [224,271]. Some even suggested that reflective thinking predicted depression level and suicidal ideation [131,167].

Takano and Tanno [261] modeled the relationship between reflection and rumination, in order to single out the unique effect of self-reflection on depression. To do this, they collected rumination and self-reflection assessments, along with self-rated depression symptoms from 111 college undergraduates. They measured rumination and self-reflection at two different time points. To test the bidirectional paths between reflection and rumination, they constructed a correlation matrix and conducted structural equation modeling. Contrary to the existing understanding of the relationship between rumination and reflection, their findings point to a unidirectional relation between self-reflection and self-rumination. While self-reflection significantly predicted self-rumination, the opposite did not hold the same. Also, while self-reflection was associated with a lower level of depression, self-rumination was highly associated. The total effect of self-reflection on depression was almost none. Takano and Tanno discuss that this is so because reflectors tend to reflect as well as ruminate; the adaptive aspects of the reflective thinking are canceled out by the maladaptive aspects. They also add that self-reflection may easily turn into self-rumination, when individuals attempting to understand their current problems fail to generate solutions during their problem-solving attempts. These point to the delicate boundary between reflection and rumination in the reflective thinking process, revealing the need for self-reflection to be carefully guided and taught as a learned skill.

When one peers inward, what happens? Previous work had predicted that introspection was usually associated with depressive symptoms and therefore self-focused attention might be maladaptive. However, we've gained more knowledge over time to find out that self-reflection may bear fruit when one has enough cognitive resources to establish sound self-concept. Moreover, there are two distinct but intertwined processes of reflection and rumination, which may be related to different personality correlates. Further research has suggested rumination and reflection may take place simultaneously, potentially having no gained effect. Still, many emphasize the practice of self-reflection [210], which now invites a further look on its best practices.

### 2.1.2. The Self in Reflective Practice

One popular and established practice of self-reflection for mental and psychological wellbeing may be expressive writing [196]. The simplicity and convenience of the writing task has attracted many to replicate the work to ensure a guaranteed health improvement. However, reports of inconsistent findings have motivated a group of researchers to introduce modifications and alternatives.

#### 2.1.2.1. Written Self-Disclosure

For many professionals and psychologists working to promote mental wellbeing, the aim is to heal the scars left from negative life experiences, traumas and other distress. The inhibition or avoidance of negative emotions [89] and the suppression of thoughts [276] lead to heightened physiological arousal, negative mood, and impaired cognition [214]. Moreover, individuals facing distress may be trapped in brooding, or rumination, repeatedly and passively focusing on the stressful event and its possible causes and consequences [246]. Brooding individuals may fall into the tendency that exacerbates further ruminative thinking, increases negative emotions, and interferes with problem solving [246]. In contrast, accessing, expressing, and processing inhibited emotions is thought to be adaptive [214]. Intervention techniques have included challenging negative thoughts, supporting the confrontation of painful images and emotions, and promoting active problem-solving strategies that may effectively ameliorate psychological and behavioral difficulties [246].

Many self-reflective demonstrations of this take place in a variety of forms. Personal journal writing can serve multiple purposes: a form of self-expression, a record of events, a form of therapy, or combinations of these and others. It is a form of reflective practice [26], as a device for working with events and experiences in order to extract meaning from them. Rainer [215] earlier suggested diary as the only form of writing that allows total freedom of expression. Stream-of-consciousness writing, in which words are poured out without pause for punctuation, spelling, or self-censorship, can also be of value [26]. In working with feelings, expressive writing has a particular role to play [26]. Written emotional disclosure as in expressive writing is a self-reflective practice in that it is to be tested without the presence of feedback of a listener or therapist [204]. It has been described as "solitary" and "anonymous" [201] and its parallels can be journal writing and diary-keeping [214]. Simple as it is, expressive writing has had many findings on improved health.

In the standard version of Pennebaker's expressive writing paradigm, participants are to write for 15 to 20 minutes daily for several days on either stressful experiences or non-emotional topics as control. This simple writing practice has been shown to positively affect the physical and psychological health of individuals diagnosed with cancer [253], asthma or arthritis [249], fibromyalgia [30], chronic pain [185], trauma (e.g. [245]) and anxiety (e.g., [63,190]). Further research suggested that it may also facilitate active problem solving by having writers analyze and process their experiences [246]. Lyubomirsky et al [149] found that writing about stressful experiences was more beneficial than merely thinking about them. They speculated that writing is associated with greater benefits because it allows people to organize the past experiences. In contrast, thinking about them is detrimental because it can rapidly transform into brooding or rumination. Likewise, Sloan and colleagues have demonstrated that expressive writing buffers against maladaptive rumination [246].

More recent reviews on expressive writing, however, have tempered these conclusions. They argue that inconsistent findings occur in non-clinical populations (e.g. [33,77,83,83,83,128,129,147,257]). It has even been suggested that some participants may experience negative long-term health effects after completion [77,83]. These concerns suggest the writing may not guarantee the positive effects of reflective pondering at all times.

One limitation of expressive writing is the non-directive instructions of the task. Participants write about their deepest thoughts and feelings about their most upsetting life experience [198], without any advice or instructions on how best to go about it. Thus, writing styles may be confounded by self-selection, and an unlimited number of writing methods may or may not turn out to be beneficial. This is illustrated in Pennebaker's process research [199,258], which suggests that individuals who write with ruminative, static patterns of thinking do not attain benefits.

In spite of the limitations, expressive writing provides opportunities. Guastella and Dadds [90] suggest that expressive writing can provide a valuable emotion-processing research tool that is an analogue for a therapeutic process, considered relatively free from therapist variables. Moreover, if refined and better focused, it could provide a cost-effective and easily disseminated intervention to assist the community in large following trauma exposure [90].

#### 2.1.2.2. The Self Conundrum: The Need for Guidance

The debatable aspects of expressive writing mirror the ruminative and reflective paths of self-attentional practices. Moreover, taken in a bigger picture of things, expressive writing as a self-reflective practice also necessitates an individual's continued engagement to ensure benefit. Indeed, Porter [210] points out that although many know the benefits of taking the time to pause and reflect on themselves, they are discouraged from doing so because of the following. First, many do not understand the reflection process. It is often vague to "reflect" on something, unless they are given specific and substantial cues. Moreover, many can avoid doing it because they do not want to fall into a shame spiral [32] or do not see a substantial outcome. The lack of clarity in direction and motivation may pose a barrier for any self-reflective practice to engage individuals.

Several follow-up research on expressive writing also support this view. Guastella and Dadds [90] suggested a more structured writing to complement the lack of instructions in Pennebaker's original format. They conducted an experiment with three writing conditions from the cognitive behavior models of trauma: exposure, devaluation, and benefit-finding. Their results provide evidence that participants engage in different emotional processes in each writing condition. Their findings also suggest that given the instructions, writers can engage in cognitive restructuring processes and therefore hint at the possibility of stepped-based procedures in writing.

In their later study, Guastella and Dadds [91] tried a growth writing paradigm, combining several emotional processes in a sequence in an expressive writing format. Here, they tried to shift a writer from writing a past event-focused narrative, to devaluation, and finally to finding benefits from the stressful experience. Their growth model assists the writer to progress through a sensory based processing strategy to more cognitive higher order reasoning-based processes [91]. Their findings show that a sequential model of specific emotion processes, where the writer shifts from sensory to more elaborate levels, can lead to a greater psychological benefit in the long run, compared to an unstructured writing group. Though a preliminary study, this study shows potential in designing a writing task that can engage a specific set of emotion processing models.

The earlier work suggests the possibility of integrating a step-wise approach in the unstructured, free-formed expressive writing to help individuals engage in specific cognitive restructuring processes for meaning making. In fact, this is not unlike what had already been suggested in expressive writing literature. Specifically, Smyth et al [250] compared two different forms of expressive writing in their study. They assigned more than hundred participants from a nonclinical population to write about control topics or about their thoughts and feelings regarding the most traumatic event of their life in either a fragmented, list-like format or a narrative format. While the fragmented writing group did not show any difference from the control group, the narrative group reported less restriction of activity and showed higher avoidant thinking than the others. Smyth and colleagues have concluded that the specific instruction to form a narrative of a trauma can invite different responses from others and further suggested that a narrative format may be required to achieve health benefits.

In a similar vein, Danoff-Burg and colleagues [56] conducted a study to compare a narrative form of expressive writing and the original format to a control writing condition. In their study, the narrative writing group showed higher levels of narrative structure than the expressive writing group. Greater narrative structure was associated with mental health gains, and self-rated emotionality of the essays was associated with less perceived stress at 1-month follow-up. In addition, both writing groups reported lowered perceived stress and depressive symptoms relative to controls but did not differ from each other with regard to these outcomes. Their findings suggest both emotional expression and narrative structure may be key factors underlying expressive writing's health benefits.

Despite the efforts to complement the non-directiveness of expressive writing instructions, research has shown that while narrative making may help, the effects cannot be guaranteed. Sales, Merrill and Fivush

[230] have studied the narrative meaning-making process of traumatized female adolescents. They found that narratives having a more external locus of control and more cognitive processing language about a highly negative past event were associated with increased depressive symptoms. The findings suggest certain types of narrative language reflect ongoing and unsuccessful efforts of meaning-making and outcomes may relate more to rumination than to resolution. The researchers also add that for narratives to produce beneficial results, a structured, scaffolded model of narrative meaning-making may be necessary. Taken together, previous research suggests expressive writing in a narrative format, with stepwise, scaffold fashion may work more consistently toward health benefit.

More recent studies in extended applications of expressive writing present an interesting suggestion of accompanying an audience. Lengelle, Luken and Meijers [139] investigated the factors that promote the benefit of self-reflection in career-identity development. They created a career-learning intervention as in a "career writing" method. It is a combination of creative writing, expressive writing and reflective writing. They summarized that to foster reflection for healthy life, designing requires a safe holding space facilitated by a compassionate and knowledgeable teacher or guide. Their findings indicate that a successful method will include engaging and observing feelings as well as having a mutually inspiring internal and external dialogue.

In this vein, Radcliffe et al [214] argued that in expressive writing, the researcher could essentially be an implicit audience. In other words, the fact that the writing is submitted to and read by the researchers means that there is an audience for the writing and therefore expressive writing is altogether a social experience that is not, in fact, private. Though the idea of an implicit audience or imagined reader [31] had been suggested, the "actual audience" had never been tested. Their findings point out that while both shared and private disclosures resulted in less cognitive intrusion and avoidance than the control, shared disclosure reduced depression and interpersonal sensitivity the most, and could only reduce physical symptoms amongst all conditions. They concluded that although truly private writing improves cognitive stress effects, shared writing has broader benefits, suggesting social disclosure in expressive writing may matter.

### 2.1.3. Design Space

Reviewing related works of research on self-reflection and expressive writing for mental health, this work proposes a design space for chatbot-guided reflection, with emotional expression and guidance to ensure best practices. Varying the levels of guidance and self-disclosure, reflection may take a different shape.

#### 2.1.3.1. Two Dimensions for Chatbot-Guided Reflection

So far, the review of literature shows that self-reflection may take a variety of forms and procedures, and self-focused attention may take either a reflective or ruminative path, or both in a simultaneous manner. A popular and established self-reflective practice for mental wellbeing is Pennebaker's expressive writing paradigm [196], which has shown for decades a continued line of research supporting that writing about one's trauma for about 20 minutes for three to four days may lead to improved health outcomes. Yet inconsistent findings exist, and efforts have been made to complement the lack of concrete instructions in the standard expressive writing format. It has been suggested that constructing a narrative or redesigning the writing instructions helps, so that it may encourage one or more specific emotional processes. More recently, a social procedure – with a guide or an audience – has been suggested to ensure broader benefits.

A consistent finding throughout the expressive writing hustle-bustle is that the written disclosure of emotions helps writers cope with the health consequences of negative life events [64,72,85,202]. Pennebaker [202] proposed that actively inhibiting thoughts and feelings about traumatic experiences requires effort. It is a cumulative stressor on the body and is associated with increased physiological activity, obsessive thinking or ruminating about the event, and a longer-term disease. Confronting a trauma through talking or writing and acknowledging the associated emotions is thought to mitigate inhibition, gradually lowering the overall stress on the body [8]. Such confrontation involves translating the event into words, enabling cognitive integration and understanding of it, which further contributes to the reduction in physiological activity associated with inhibition and rumination [8,202].

This theory has intuitive appeal but mixed empirical support [8]. Studies report that expressive writing mediates improved health outcomes [24,64,198,207]. However, this has not always been consistent. Participants writing about previously undisclosed traumas showed no differences in health from those writing about previously disclosed traumas [85]; and participants writing about imaginary traumas also demonstrated significant improvement in physical health [83]. Therefore, although inhibition may play a part, the observed benefits of writing are not entirely due to reductions in inhibition.

To tackle the inhibition problem, we can turn to another consistent finding within the expressive writing paradigm. It is that those who benefit from the writing process were more likely to increase the use of "cognitive mechanism" words (i.e. insight words such as "understand, realize" and causal words such as "because, reason") [204]. It is in this vein that the development of coherent narrative of trauma may yield a beneficial effect of expressive writing, reflecting an increased cognitive processing of the experience. Other studies have also addressed the linguistic features of the writing that session-to-session variations in pronoun use are related to health improvements, which may reflect a transformation in the way people think about themselves in relation to others and the world [23,201]. In addition, since it was suggested the more structured approach of the expressive writing paradigm can be more beneficial than simple diary-keeping [248], there has been an extended line of research on varied applications of expressive writing that incorporates more stepwise, structured approaches by adding more specificity and guidance in the instructions [90,91,139,214].

Hence this work proposes that to achieve a positive outcome from reflecting on past events, emotional disclosure that is scaffolded by appropriate guidance may be necessary. The guidance may be provided in such a manner to complement the lack of directions in the original expressive writing format, and to encourage the "cognitive processing" in recounting the event.

#### 2.1.3.2. Disclosure and Guidance

This work is motivated to suggest both disclosure and guidance as the necessary conditions for self-reflection to promote mental wellbeing. <Figure 2.1> is a two-dimensional design space where emotional disclosure and guidance are each put on a continuum. Depending on the levels of emotional disclosure and guidance provided to scaffold reflection, it presents four different reflection processes, which will be illustrated further with examples in the following subsection.

Each subspace in <Figure 2.1> is labeled, counter-clockwise from the bottom left corner: revisiting, exploring, transformative and explaining. These processes are illustrated to distinguish the different levels of user disclosure and bot guidance in design.

Revisiting space is the most common type of reflection on past events, experiences and memories. Not much self-disclosure, as in confession of innermost thought sand feelings, is needed here, nor much guidance or intervention. A simple memory trigger may suffice.

At the top right corner is the transformative space. It is suggested that transformative reflection is the ideal form of self-reflection, bringing about positive change in behavior as well as mental schema [69,247]. Here users need to make a bold transition from the past to the present, looking toward the future, changed self. The chatbot thus needs to be more engaged in providing instructions, directions, encouragements, or any other form of guidance to lead the user. The user, too, needs to form a narrative from revisiting to resolution.

Moving upwards from the revisiting space is where the self actively engages in disclosure, perhaps in the form of social exchanges and feedback with others. In the explaining space, still not much guidance is necessary, as one may simply ask the user to “tell more.” Moving towards right from the revisiting space is the exploring space, where the chatbot may actively intervene with thought processes by asking or challenging to reinterpret for example, a past memory, in a different light. Here users may actively interact with the bot their thoughts, feelings and ideas, within which process they can enlighten themselves with new meanings.

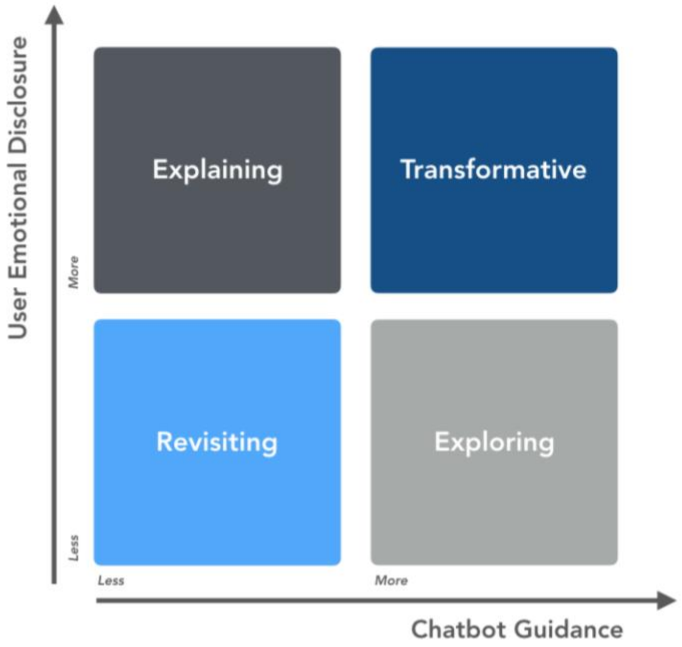


**Figure 2.1. The disclosure-guidance space for designing chatbot-guided reflection. Different levels of user disclosure and bot guidance can support different types of reflection.**

The goal of this study is to explore this space by designing chatbots that can provide a safe environment and assistance for self-reflection. According to the levels of disclosure and guidance, four types of reflection processes are illustrated. Next section will discuss each type and span a review of technologies designed to support each process.

## 2.2. Self-Reflection in HCI

How has the HCI community responded to the need for technologies to support reflection? A brief survey on technologies in related works is mapped onto the design space above according to their aims for design.

### 2.2.1. Reflection Design in HCI

The notion of reflection and reflective practice has been one of the central interests to the human-computer interaction (HCI) researchers and practitioners for quite long [69,232]. Reflection has been extensively studied in the context of learning and professional development [158,169,236,263], and health has been a focus in talking about self-reflection, promoting healthy behavior change (e.g. [4]) as well as promoting greater awareness and learning to self-manage chronic conditions such as diabetes (e.g. [153]) [69].

However, engaging in reflection is far from straightforward [247]. Designing technologies to support reflection is challenging, and what is even more daunting a task is to establish a shared understanding of what is to be designed when designing for reflection [69]. Many come from different perspectives and are working with different methods. Yet as discussed above, HCI shares an understanding of reflection as “reviewing a series of previous experiences, events, stories, etc., and putting them together in such a way as to come to a better understanding or to gain some sort of insight” [14], on which this work grounds its design space.

#### 2.2.1.1. Designing Technology for Reflection

What are the ways in which technology can be designed for reflection? Moon [170] illustrates many ways designers can use to create the time for, guide and encourage different levels of reflection: writing techniques, reflective questions, dialogues and discussions, nonverbal techniques, reviewing materials, self-assessments, using ill-structured material, and other methods for creating situations which require aspects of reflective thought. Building on Moon’s [170] notion of levels in reflection, Fleck and Fitzpatrick [69] discuss five levels of reflection: descriptive, reflective, dialogic, transformative and critical reflection. Here, in transformative and critical reflections reflectors engage in a fundamental change in understanding; that is, their self-insight can lead to transformation. In their view, a technology could engage multiple levels of reflections.

On the other hand, Mols et al [169] takes a memory perspective in defining reflection as reassessing the present to move toward perceiving, knowing, believing, feeling and acting. They specifically focus on specific design strategies to support reflection, e.g. dialogue-, information-, expression- and environment-driven, to establish a design space to support everyday life reflection. This view takes reflection as triggered by different modes of interaction or artefacts.

Building on earlier work, this work offers an overview of technologies for reflection based on the types of reflection. While earlier work has discussed reflection having hierarchical levels that vary in depth, this work argues that reflection can take different processes, and each is just as valuable in gain. Technologies can be designed to support each process in different ways. It has been suggested from literature that for effective self-reflection, one needs not only emotional disclosure but also safe and appropriate guidance. To explore this design space, it is necessary to survey the types of reflection processes according to different levels of disclosure and guidance. <Figure 2.1> has illustrated this design space. Below describes four types of reflections with examples of technologies that support them. It is not to achieve a comprehensive and exhaustive review of all technologies for reflection, but to illustrate most salient features of each.

#### 2.2.1.2. Four Types of Reflection

Technologies have increasingly become able to capture memories of the past. An earlier work by Stevens et al [256] has investigated how we should address the design of memory systems. They prototyped Living

Memory Box, a physical artefact with a computer and a translucent box that held mementos accompanied by user narratives. They advised the reflective systems to allow an archiving of practically anything and to support natural interactions, encourage storytelling and even create unique experiences from the memories. Their work shows the fluid and multifaceted nature of reflective thought; reflection design spans from the capturing of past memories to creating new life stories. This process may or may not be holistic or only a fragment can be perceived as a whole. It is the role of technology to embrace the different shapes of reflection, and its design needs to address them. Below, four types of reflections are illustrated in the aforementioned design space, depending on the level of user self-disclosure and the guidance provided.

### (1) Revisiting Process

Most reflection designs fall in this area. Often the technology invites the reflector to pause and ponder on a remnant or an artifact (e.g. photos, emails, texts) from the past. It also asks to provide some descriptions, to engage the reflector in the reflection process and discover new ways of interpreting the past. This requires minimal levels of disclosure and guidance, which suffices to revive the past memories.

With lifelogging tools and the quantified self, Li and colleagues [143] have conducted a qualitative study on user motivations in using personal informatics tools and thus their data on health and productivity. Their findings led them to identify two phases of reflection on personal data: discovery and maintenance. Users transition between the two phases to resolve unanswered questions about their data and set new goals. Their work testifies to the fact that people would like to pursue awareness about themselves, with personal gadgets and equipment that constantly record personal data. New information about themselves leads to new ideas and goals, to promote a better self.

As such, advances in technology have radically increased the access and ability of people to capture their lived moments “live.” Reflection technologies have focused on capturing a variety of personal data and bringing them back to the users for further thinking as a “memory aid” [106]. For example, SenseCam [106] is a sensor augmented with a wearable stills camera that is designed to capture a digital record of the wearer’s day by recording a series of images and sensor logs. The primary purpose of design is to help users recall the past memories for recollection. In a similar vein, Affective Diary [251] tried to capture not only a personal digital record but also “bodily memorabilia” with mobile body sensors. In their experimental user study, they found that users were able to recall the past moments and learned something new about themselves. Later, AffectAura [155] allowed for a continued recording of emotional states over a long period of time, by putting together a multimodal sensor set-up for logging of audio, visual, physiological and contextual data, with a classification scheme for predicting user affective state and an interface for user reflection. What these have in common is that they tried to capture embodied moments that often go unnoticed and even forgotten, and bring them back to the users for discovery. Understanding such a design has also been attempted in Life Tree [193], where users play a game of breathing exercises to grow a tree. The “bringing back” aspect of these technologies could successfully engage users into the rediscovery of self and their desire to enhance self-knowledge.

Recording and revisiting personal data are not limited to sensors but visual archives. Storytellr [132] is an authoring tool for narratives, which integrates aspects of storytelling with photo activities such as annotation, search and construction. This is to help users' recollection of past events with photos as memory triggers. Taken further, MirrorMirror [80] is a hearing aid, a Speechreading Acquisition Tool (SAT) that allows users to practice their speechreading by recording and watching videos of people they frequently speak with. Here, photos and videos captured by digital technologies are used as a tool to help users face what they have been.

Technologies have already allowed us to reflect on ourselves by recording and retrieving the past. Abovementioned technologies are designed to incorporate a variety of data as memory triggers for people to aid recall and sensemaking, and perhaps serendipitous reinterpretation. However, those that focus on revisiting the past do not necessarily focus on taking the recollection further.

**(2) Transformative Process**

Technologies for transformative reflection are in a similar vein, yet they aim for leading the user to a positive change in behavior. Hence these are often found in persuasive design (e.g. [70]) and healthcare (e.g. [48]). This involves higher levels of both disclosure and guidance, for it requires a close examination of the reflector's *as-is* to move onto *to-be*. It often provides steps or guidelines for the reflector to follow and engage within the process.

Persuasive technologies often concern changing problem behavior for health. According to Consolvo et al [48], design strategies for persuasive technologies also incorporate a reflection component, to encourage users to reflect on their behavior by showing them what they have done and how the behavior relates to their goal. Examples include MAHI [153], where users diagnosed with diabetes enroll in an education program with getting feedback on their key measurements. Also, Community Mosaic

[191] helps underprivileged communities to eat healthy food by asking users to take photos of food they eat to inspire others in the community to eat healthier. What these have in common is the strong scaffolding element to engage users in the process, such as getting others to comment and feedback.

In reflection design, Slovák et al [247] define transformative reflection as "eliciting change in behavior or mental schemas." Taking Schon's concept of reflective practicum into two social-emotional learning (SEL) studies, they suggest a two-step process: The first step offers a set of questions aimed to help understand characteristics of the "right" experiences that are likely to be conducive for transformative reflection. Second, they propose explicit, social, and personal components for technology design in scaffolding the selected experiences. Based on their findings, they argue that transformative reflection needs a careful scaffolding of guidance as well as a safe interpersonal element for sharing experiences, which aligns with the design of persuasive and healthcare technologies.

A relatively less explored domain is the transformative reflection for emotional experiences, particularly negative ones. As a matter of fact, this usually takes the route of designing technologies for mental health and wellbeing. These technologies usually focus on emulating counsellor or therapist behavior via real or virtual interpersonal communication design, which will be discussed in the later section. Taken together, transformative reflection is more explored in terms of behavior change and therefore concerns various healthcare technologies. Little has been found how one would voluntarily go about the process and how they discover self-insight.

**(3) Explaining Process**

Technologies for explaining the past provide users with memory aids, cues, or triggers to recall the past and invite them to actively engage in them by answering further prompts. These technologies usually require users to illustrate what they are thinking or feeling, in addition to capturing their affective or cognitive state. They invite the reflectors to provide their reactions or interpretations on the past events and perhaps find new meaning. This type of technology needs higher levels of self-disclosure from the reflector's part; it often involves a narrative, and in the process of making a narrative, new understandings may emerge.

Social technologies are a good candidate for explaining reflections. PosiPost Me [121] follows positive psychology tradition and leads users to elicit positive thoughts and share with friends. Instead of capturing the past memories, PosiPost Me prompts users to complete an unfinished sentence about themselves to others, thus allowing for self-expression and social awareness. In a similar vein, MobiMood [44] enables groups of friends to share their moods with one another via a mobile app. Rather than capturing and recording moods by oneself, explicit sharing of moods in-situ has triggered further conversations and communications among users, allowing for their own interpretation.

Besides social technologies, memory triggers can also ask for further explanation. Pensieve [194] supports everyday reminiscence by emailing users memory triggers that contain their previous social media posts or text prompts about common life experiences. The Pensieve system allows for explaining reflection since the system takes the proactive role and asks people to answer a set of questions about their past memories. The researchers find that people value spontaneous reminders, as well as the ability to write about them. Their findings point to an important factor in self-reflection that people would like to express themselves, preferably to an audience, even a hypothetical one.

Echo [116] is most similar to Pensieve and most relevant to the scope and purpose of this study. It is a smartphone-based app for recording everyday experiences for reflection. The researchers explore the concept of technology-mediated reflection (TMR) with Echo, and find that TMR can improve mental wellbeing. More specifically, Echo encourages users to reflect on prior social media posts. Users view the post and record their current happiness ratings and are asked to enter their current reactions. In two deployments of Echo, researchers found Recorders and Reflectors engage in different emotional processes. Unlike Recorders who only kept a digital record of the day, Reflectors reviewed their past memories and reevaluated their happiness, also writing about and analyzing them. While this work leverages the Echo system in the way that it had users to "reflect" on past events, we dig deeper into the reflection process by targeting different levels of guidance into writing about them.

Thus far, technologies for explaining reflection mostly provide past memory cues and ask users to find meaning. Pennebaker's expressive writing [196] could fit in this category, with paper and pencil as technological medium. These technologies focus on what happened, and what they might mean. The potential downside of this reflection could be that people can be self-immersed; people see things in the way they'd like to see. In other words, the systems do little to challenge the boundaries or test conflicting thoughts and emotions.

### (4) Exploring Process

In exploring reflection, the technology actively asks the reflector to provide more than explanations, but reinterpretations or perhaps think outside the box. Here the reflector can engage in a cognitive process to answer and respond to the technologies to account for past actions and thoughts, as well as challenge themselves for new insights. This design involves higher levels of guidance, with the technology potentially leading the reflection process in surprising or unexpected ways for finding new patterns and meanings.

Again, Echo [116] is relevant in this category, as the reflection activity was rather broadly defined for the study participants. The Reflectors who engaged in active journaling and analyzed their thoughts and feelings could learn new lessons. However, there was also a downside when unpleasant events came back, they would not lend themselves to personal growth. The first author of the study who actually took part in the experiment said she preferred to forget the details of a negative life event. This is quite contrary to Pennebaker's research on expressive writing; writing about negative life experiences could lead to self-insight and self-knowledge [204]. Still, it is not a pleasant experience to invest such a time and effort to think about negative life events.

Reflections for exploring life events, especially negative ones, for resolving past trauma and stress can address this. Here technologies intervene to make inquiries about the event to the user, not only in *how-you-felt* way, but also *how-about-this* way. In other words, it expands the scope of the event to a bigger picture by distancing users from the event and challenging them to think from a new perspective. The key design challenge here would be making such creative yet contextually relevant cues for the users.

### 2.2.2. HCI for Mental Wellbeing

In recent years, there has been an increase in research exploring the role of technology and interaction design in supporting mental health and therapy [264]. Systems in therapy are often designed to facilitate communication between therapist and client, to provide therapy-specific contents or to support a patient's self-monitoring activities and therapy compliance [154]. Outside therapy, technologies help patients become co-creators of their care [60]. Now they can have greater access to health-related information than before. There are online services for self-care, health advice or counseling. Here the focus is to review an array of recent technologies designed for patients to review negative life experiences from the past for emotional wellbeing.

Communicating emotions is inherently social. In sharing our feelings we invite empathic responses, allowing others to meet our needs and enable the building or maintenance of social relationships, an element that is of fundamental importance to maintaining wellbeing [225]. Thus, a lot of work in HCI has been invested in the design of technologies that would perhaps emulate the role of a counsellor or a therapist who would help a client communicate his or her feelings. Recently, this has taken the form of computer-supported peer-to-peer dialogues, or conversational agents commonly called chatbots. The design of these technologies enables the reciprocal exchange of feelings, consolation and empathy, which enables in a virtual space an interpersonal relationship, which is a powerful determinant of health and wellbeing [225].

#### 2.2.2.1. Engaging Peers, Social Networks and Bots

Thus far HCI researchers and designers have taken, broadly, three approaches: engaging peer support, leveraging social media and creating a virtual therapist. For peer support, technologies are designed and used for getting together online peer support groups and communities to learn about therapeutic techniques to support one another. For example, Moderated Online Social Therapy [2,138] is an online peer support program that encourages people with schizophrenia to learn about cognitive and behavioral strategies via a social network, moderated by clinicians. Panopoly [175] is a crowdsourced mental health intervention for peers to help reframe each other's thoughts using therapeutic techniques. Others include Spheres of Wellbeing [265] and Self Harmony [20]. Spheres of Wellbeing [265] are interactive objects that engage people with mental illnesses to participate in a co-design process for empathetic interventions. Self Harmony [20] engages participants to engage in design processes to reduce self-harm.

More recently, O'Leary et al [188] designed guided and unguided chats between peers for emotional support. They conducted a two-week experiment with 40 participants with mental health conditions. Their findings show that anxiety was significantly reduced from pre-test to post-test; participant experiences testify to that guided chats provided solutions to problems and new perspectives, and were perceived as "deep," while unguided chats offered personal connection on shared experiences and were experienced as "smooth" [188]. This sheds much light on this study in that it incorporated the idea of designing guides for chat among peers. The guided prompts were based on a problem-solving framework, similar to problem-solving therapy and cognitive behavioral therapy [188]. Broadly, the prompts included open-ended questions to invite explorations on client concerns as well as suggestions/advice for solving problems, and reflective listening skills that are often used by therapists to show empathy.

Social media have also contributed to supporting mental health online. Social media can provide an interesting glimpse into people's mental health [42,192]. Peers with depression and other conditions seek information, emotional support, and advice [10,55,66]. It's been suggested that peer support platforms can glean people's mental health needs such as when, why, and how people seek out help [187]. Research findings have reported that people with mental health issues prefer to go online for support for the benefits of anonymity, empowerment, and access [111,142,157,208,212]. Nevertheless, it is not always guaranteed that online support groups can be effective. Participating in online communities for mental health can be distressing and exacerbate symptoms, even when people report having positive experiences [122,238,260]. Evidence of online interactions between peers with depression show that people have negative experiences with unsupportive members, negative content, and conflict of beliefs [142]. Training peers and providing scaffolding could help, but considerable moderation may be advised in seeking emotional support online.

Alternatively, with advances in chatbot technologies there have been attempts to build conversational agents that can engage in virtual psychotherapy. Most widely known is Woebot [68], a text-based conversational agent that delivers cognitive behavior therapy (CBT) principles in a conversational format. Researchers set up a randomized control study with 70 individuals with depressive symptoms. Compared to a control group that referred to depression guidelines by the National Institute of Health, the treatment group that talked with Woebot significantly showed significantly reduced symptoms of depression, measured by the PHQ-9, over the study period. Others take a similar approach, replicating various behavior-based therapeutic techniques to

Woebot. Shim [148] and Vivibot [86] also incorporate positive psychology and CBT interventions in a chatbot form, resulting in participants showing improvements in psychological wellbeing such as lowered anxiety and perceived stress, as well as higher engagement. Tess [73] is also a behavioral coaching chatbot that addresses different facets of behavioral health including depression and anxiety. Deployed in an adolescent pre-diabetes patient group, it testified to a promising potential to accompany clinicians. Finally, Wysa [115] is an AI-based emotionally intelligent mobile chatbot app that is aimed at building emotional resilience and thereby promoting mental wellbeing. In fact, the chatbot uses a combination of self-help practices such as CBT, dialectical behavior therapy (DBT), motivational interviewing (MI), positive behavior support, behavioral reinforcement, and mindfulness. What these chatbots have in common is that they rely on the widely established, or evidence-based practices for chatbots to emulate real-life psychotherapists, for both resource-effectiveness and efficacy.

Some are taking a slightly different approach, focusing on the empathy side of the chatbot-mediated therapy. Koko [174] is a chatbot app that uses a corpus-based machine learning approach to simulate expressed empathy. The system generates chatbot responses from an existing pool of online peer support data. While the majority of the user evaluations on Koko's empathetic responses were deemed acceptable, users would prefer those from their peers. The findings point to an interesting tension in designing for chatbot therapists. Although empathy is a significant factor in determining a therapy outcome, machine-generated empathy would not be perceived "genuine" per se.

### 2.2.3. Design Opportunities

So far, the reflection design in HCI has mainly focused on inviting the users to revisit and reinterpret their past experiences by providing a variety of cues and nudges prescribed through design. For health and wellbeing, most HCI approaches have been invested in designing the technology to best emulate a helper – in the form of a therapist of a coach – to correct or “prescribe” the right treatment path. Marrying the two together, this work proposes to design a social experience that can help users transform, explain and explore their understanding of past life experiences for wellbeing. Most reflection technologies in HCI have provided gentle reminders that perhaps trigger tristful reminiscences of one’s past. That is, while the technology engages the user for the re-discovery of the past, it is entirely left for the user to be responsible for the reinterpretation of the event.

The opportunity lies in-between. This work takes a novel approach by engaging users in building a spontaneous conversational narrative. The user is guided by technology that may nudge him or her to explain, explore or transform their understanding of the event in the past. Ideally, it can take the form of a conversational agent or a chatbot. Whilst chatbot technologies have already been widely studied for mental wellbeing in HCI, there is little transparency in how they are designed to lead and communicate in the conversation. When it comes to sharing emotional experiences, the feeling of being understood, mutual respect and empathy are some of the most important determinants of how the outcome may turn out. For therapists it takes years of training to master how to talk to their clients [101]. Since machines cannot talk like humans but can only be programmed to talk in certain ways, designing the talks-in-interaction, including turns, sequences, pauses, questions, or any other devices that make a conversation is a crucial task in design [171–173].

Nonetheless, in the survey of related research so far it is relatively less explored how to make machines talk and in what ways. In fact, most reflection design technologies in HCI only communicate with users with visual or audio cues and textual nudges, which does not require continued engagement even for a small talk. Though mental health chatbots do engage in conversations with users, they have been more invested in how to implement the therapeutic techniques in action, or the therapeutic impact of talking machine that is essentially an amalgam of different counselling methods (e.g. [68,115]). Therefore design transparency in conversational UX [173] is strongly needed for chatbot technologies for self-reflection.

In the storytelling process with a chatbot, it is important not only that the user responds to the chatbot but also that its guidance is relevant. In this context, relevance refers to contextual understanding and appropriateness of the chatbot responses. Contextual understanding means that the chatbot stays in the conversation and follows up with the user within the flow of the conversation. Appropriateness of the guidance, however, takes it further. The chatbot response must fit in the context but also encourage, expand or challenge the context in a way that may contribute towards the user's reflection process. The subtlety of the message delivered by the chatbot can entail multiple interpretations by the user. To maintain the minimal level of contextual understanding and appropriateness of the guidance delivered by the chatbot response, a key portion of the user's original message will be extracted and incorporated in the return response in the design process. In this manner, the chatbot response conveys the semantic sense to the user that it maintains the contextual flow as well as provides the hermeneutic space in which the user can re-explore what has already been said by the self.

It is one of the aims of this work to explore this unique design space with chatbot technology and present empirical evidence via experimental user study. Next section will discuss strategies to build chatbots for user-driven reflection narratives.

## 2.3. Conversational Agent Design

Having reviewed related works on self-reflection and its technologies, the goal of this study is restated: to design and implement a reflection assistant chatbot for guided disclosure for transformative, explaining, and exploring reflection processes. This section describes the theoretical background and techniques with which the bot is to be designed.

### 2.3.1. Theoretical Background

This section examines the subject of the interaction to be designed: conversation. It illustrates a formal understanding of what conversation is and what it consists of.

#### 2.3.1.1. What is Conversation?

Conversation is inherently a face-to-face interaction [172]. In discourse analysis (DA), spoken conversation is defined as "any interactive spoken exchange between two or more people," referring to the broad social phenomenon [34]. On the other hand, in conversation analysis (CA), conversation is a particular kind of social activity, a speech-exchange system that displays certain features including speaker exchanges, turn-taking, talk continuity, turn allocation, repairs and so on, in and of which presents some extent of machinery and patterns [229]. Since this work closely concerns the design of conversation for machine agents, it follows CA conventions and examines the three principles of conversation. These principles refer to the generic patterns of human conversations, the tendencies that people show when they engage in a conversation. Conversational agents also follow these general patterns since not doing so would be an awkwardness leading to conversation failure. The principles are: recipient design, minimization and repair.

### (1) Recipient Design

Recipient refers to the subject of what we say and how we say it in a naturally occurring conversation. Depending on the recipient, what we say and how we say it may take a number of forms and shapes. Earlier research has suggested that speakers tend to design their talk for their recipients in various ways, such as adapting to their perceived level of knowledge [172,228,229]. According to Sacks et al [229], recipient design is "a multitude of respects in which the talk by a party in a conversation is constructed or designed in ways which display an orientation and sensitivity to the particular other(s) who are the co-participants." Recipient design generally concerns the speaker's word selection, topic selection, ordering of sequences, options and conventions for starting and terminating conversations, etc. [229]. Naturally, it is imperative that when a teacher talks to a student, he or she needs to adapt to the student's knowledge level and choose words and phrases accordingly, taking steps to make sure the student follows. For conversational agents, they need to consider the user or audience with whom they engage and tailor their responses accordingly. Thus the principle of recipient design requires a comprehensive understanding of target users, their needs and behavior [172]. In this work, the first and foremost consideration when it comes to users is that they are bringing an emotional subject matter to the conversation with the agent, on which they reflect and scaffold their thinking processes thereof. Hence the type of conversation the agent is to deliver should assume the emotional and cognitive needs of processing such information.

### (2) Minimization

Another general rule of thumb is often referred to as minimization [141,172,228]. This principle essentially has to do with efficiency [141]. When speakers engage in a conversation, they design their turns and use words in a way that would help their recipient understand in the most efficient manner. Sacks and Schegloff [228] gives an example of using names when referring to a particular individual. When we try to describe a common acquaintance, we'd rather use the name, instead of trying to give a series of descriptions to refer to him or her. Yet for conversational agents, the minimization principle rather applies to making the agent's response as terse and cogent as possible, using the fewest words as possible [172]. It is recommended to design conversational agents so that they give minimized utterances without sacrificing understandability [172].

### (3) Repair

Repair principle is an essential element of any human conversation in cases of misunderstanding and failure. In times of interactional troubles in a conversation, we use various ways to remedy it. In CA, it's referred to as "repair," the range of practices that we have for managing troubles in speaking, hearing or understanding [235]. Since the necessity of a repair means that the trouble occurred at the previous turn, repair includes methods for repeating or paraphrasing all or parts of a prior turn

[172]. Repairs can take place at any point of conversational sequence, i.e. the flow of turns in a speaker-recipient exchange, and they are a basic component of conversational competence that are used to manage local troubles in the production and design of natural language utterances [172]. This principle in fact may lessen the burden of an agent to give the "perfect" answer all the time; as long as repairs are in store, agents can try to repair the conversation to make due adjustments.

The three principles of conversation briefly survey the mechanics of a natural human conversation. Agents as speakers should consider the needs and interests of the user, the recipient in the conversation. Moreover, they should engage in the conversation in an as efficient manner as possible. Finally, the agents should be ready to make repairs in the conversation in case the user demands clarification or signals misunderstanding.

#### 2.3.1.2. Types of Conversation

When we refer to a conversation, we usually mean the ordinary conversation which may consist of the broadest range of activities from delivering news, seeking help or advice, learning to much more, the kind of interaction we may have with our family, friends and even strangers [172]. In Conversation Analytic theory, an ordinary conversation is considered the most flexible type of conversation from which other types are adapted for particular purposes by adding special constraints [79]. In this work, we classify types of conversation according to its purpose [45]: transactional and interactional, and discuss a few examples.

**(1) Transactional Conversations**

Transactional conversation pursues a practical goal, often fulfilled during the course of one interaction [45]. In this type of conversational exchanges, both parties engaging in the conversation clearly know their roles, expectations and goals of the conversation. An example of this type is service conversations [172]. It is the kind of interaction we have with a sales or an organizational representative. Here, the roles are fixed, usually the customer asking for service, and the salesperson trying to answer questions. For such, transactional conversations usually have distinctive openings, with the conversation being terminated within one sequence or only a few more, when repairs are needed.

### (2) Interactional Conversations

Interactional conversations are social conversations [45]. The aim is not to complete a task, but to build, maintain and strengthen positive relations with one or more interlocutors [45]. Social conversations range from small talk to longer interactions such as talk between friends, colleagues and strangers. Often it can help develop common ground and build rapport [39]. Though it serves a different purpose, an interactional conversation can share and overlap with transactional conversations in natural conversations [39]. An example is counseling conversations. In counseling conversations, often one seeks advice to a therapist, counselor or advisor. In psychotherapy, rapport building between a therapist and a patient is an important factor toward outcome. Thus, though counseling conversations do happen for a purpose, like transactional conversations, they are inherently social like interactional conversations.

In this work, a conversational agent is designed to primarily support the user's self-reflection, a transactional conversation where speaker roles are clearly defined and a goal is to be achieved. Yet the nature of the conversation is social in that to help self-disclosure, the agent needs to help the user feel safe and trustworthy to engage deeper.

## 2.3.2. Technical Background

Having examined what conversation means in the field of CA, its core principles and types of conversation, the technical understanding of what constitutes a conversational agent is discussed.

### 2.3.2.1. Natural Language Interfaces

Natural language interfaces are user interfaces that use human language, i.e. natural language, to interact with the user. Conversational interfaces are very different from graphical user interfaces (GUI), in that graphic elements are generally minimal [172]. The interaction metaphor for these interfaces is the natural human conversation, rather than direct manipulation [242]. Since the very first chatbot, ELIZA [275], appeared in history, many natural language interfaces have appeared: Apple's Siri, Amazon's Alexa, Google's Assistant, Microsoft's Cortana and IBM's Watson are just a few examples as of now, and we are expecting many more. While most of these systems accept voice input from users (voice user interfaces; VUI), many accept text input (text-based conversational agent; chatbot), sometimes from standard applications like SMS and instant messaging [172]. Users readily engage in interactions with natural language interfaces to check the weather, set reminders, call and send messages, play music, launch apps, search for information, and interact with other connected devices [45]. Nonetheless, natural language interfaces, or agents that communicate with human users in natural language, are still awkward, confusing, or limited and fraught with troubles [172]. Though many of them are modeled after the natural human conversation, it is a complex system in its own right [229,233], which requires works of machinery [227]. Though a perfectly natural conversation is impossible at present, thanks to the wondrous advances in the natural language processing (NLP) methods, some formalities and conventions of natural language conversation in CA can be applied in this work, so as to mechanically design a conversational agent.

#### 2.3.2.2. Conversational Agent Models

One of the challenges in artificial intelligence (AI) has been endowing the machine with the ability to converse with humans using natural language [269]. Early conversational systems, such as ELIZA [275], Parry [47], and A.L.I.C.E. [274], were designed to mimic human behavior in a text-based conversation in order to pass the Turing Test [240,269]. These systems, precursors to today's chatbots, were mostly based on hand-crafted rules. As a result, they worked well only in constrained environments [244].

Since the 1990s, a lot of research has been conducted on task-based conversational agents. Examples include the DARPA Airline Travel Information System (ATIS), the DARPA Communicator program, and the ATIS and Communicator systems (e.g. [54,97,213]). The task-based chatbots showed an excellent performance only within domains with well-defined schemas. In the past several years, a tremendous amount of investment has been made to developing intelligent personal assistants such as Apple's Siri, Microsoft's Cortana, Google's Assistant, Facebook's Messenger, and Amazon's Alexa. These assistants are not only designed to answer user questions but also proactively anticipate user needs and provide in-time assistance like reminders or recommendations [231]. The challenge remains that they must work well in many open domains as users expect them to manage their work and lives efficiently.

More recently appeared are social chatbots, e.g. Microsoft's XiaoIce.

The primary goal of a social chatbot is to be a virtual companion to users. By establishing an emotional connection with users, social chatbots can better understand them and therefore help them over a long period of time [244]. These social chatbots and intelligent assistants have become popular due to progress in many relevant perceptual and cognitive AI technologies, e.g., natural language understanding (e.g. [7,17,160,259]), speech recognition and synthesis (e.g. [58,103,282]), computer vision (e.g. [130]), information retrieval (e.g. [62,112]), multimodal intelligence (e.g. [95,123,272]), and empathic conversational systems (e.g. [74]).

### 2.3.3. Design Strategies

Having examined the theoretical and technical background, the conversational agent in this work adopts an interactional conversation mediated by a social chatbot. Drawing from the works of the renowned humanistic psychologist, Carl R. Rogers, two client-centered methods, expressive writing and motivational interviewing, are explored for design strategies for the chatbot.

#### 2.3.3.1. Chatbot Persona

To effectively guide the recipient of the conversation in this study, it is important that the bot takes on an appropriate speaker model. Because the primary purpose of designing to support self-reflection is to encourage user self-disclosure, the agent is to take after a Rogerian psychologist, as did ELIZA [275], and his successors as individual persona.

**(1) Client-Centeredness in Rogerian Psychology**

Carl R. Rogers (1902-1987) was one of America's most influential counselors, psychotherapists, and most prominent psychologists [126]. He is best known for the establishment of client-centered therapy that is later renamed as person-centered therapy. Unlike the popularized ideas of unresolved sexual conflicts derived from the psychoanalytic tradition at the time, Rogers was deeply inspired by and led his career with the ideas of client self-insight and self-acceptance in his therapy.

In his time, Carl Rogers challenged the field of psychotherapy in two ways. First, though Rogers was not the first to use the term “client” for a therapy recipient, he popularized its use. The word implies a departure from the medical model of illness, in that a person seeking help should be not treated as a helpless patient but as a responsible client [126]. Rogers believed the growth-producing process of counseling could help all individuals and professionals could be trained to provide such help. Thus, counselors, social workers, clergymen, medical workers, youth and family workers, and others could use his counseling methods regardless of their profession.

Second, Rogers introduced the “nondirective” method. Though other therapies might profess a similar belief, Rogers’ method of creating the therapeutic atmosphere was drastically different from other approaches [126]. His initial method avoided questions, interpretation, suggestions, advice, or other directive techniques. Rather, it relied exclusively on a process of carefully listening to the client, accepting the client for who he or she was, and reflecting back the client’s feelings. The acceptance and reflection of feelings would create a level of safety for deeper exploration and a mirror in which to further understand and reflect on the client’s own experience, which would lead the individual to further insight and positive action.

The essence of Rogers’ client-centeredness in therapy includes three conditions. When a counselor communicates congruence, unconditional positive regard, and empathic understanding so that the client perceives them at least to a minimal degree, then the "necessary and sufficient conditions for therapeutic personality change" are present [221]. He argued and demonstrated that the client has within himself the ability and tendency to understand his needs and problems, to gain insight, to reorganize his personality, and to take constructive action. What clients need, said Rogers, is not the judgment, interpretation, advice or direction, but supportive counselors and therapists to help them rediscover and trust their inner experiencing, achieve their own insights, and set their own direction [126].

What Rogers pursued throughout his nearly six-decade career is radically different from psychoanalysis and behaviorism, the two other schools of thought at the time. First, he put much more emphasis on the individual's phenomenal being. This is done by the therapist's empathy with the client's frame of reference, or the therapist's helping the client find meaning in life as perceived by the client himself. Second, his method focused not on remediation of problems but on psychological health, well-being, self-actualization, or what he called "the fully functioning person" [221]. The goal was to help people experience their full human potential. Finally, he was deeply interested in what distinguishes human beings from other species, such as choice, will, freedom, values, feelings, goals and others human concerns, which remained as key subjects of his study.

In this work, Rogers' client-centeredness sets the backdrop of the design of chatbots for self-reflection. In other words, the chatbots pursue the role of a humanistic psychologist that nudges and waits for the user to share his or her stories in the narrative-making process. Its existence solely serves the role of a "supporter" [126], instead of giving advice. The following subsection will discuss the specific methods within Rogers' humanistic tradition in which the chatbots will deliver conversations.

In fact, the legacy of Carl Rogers transcends the boundaries of the humanistic tradition. His core conditions for therapeutic relationship serve the basis of all training and professions in clinical psychology. There has been a wide array of branching methods from person-centered therapy, among which two of them concern the purpose and scope of this study: motivational interviewing and expressive writing.

### (2) Two Descendants

#### A. Motivational Interviewing

The clinical method of motivational interviewing (MI) evolved from the person-centered approach of Carl Rogers, maintaining his pioneering commitment to the scientific study of therapeutic processes and outcomes [162]. What MI sets forth mirrors much of what Rogers himself already had in his pioneering article on the necessary and sufficient conditions for personality change [221]. MI counselors accept their clients in an unconditional manner and have a collaborative relationship with them. Counselors' goal in this approach is to accompany and help clients in the process of change, which is in agreement with clients' aspirations and values. In addition, counselors seek to evoke clients' intrinsic motivation to change and to make it emerge, rather than imposing it. Clients are considered to be the main persons responsible for their behavior change and counsellors support the client's autonomy. It is aligned with Rogers' belief in client self-actualization.

MI was developed as a method of communication, rather than a set of techniques, and the MI style overrides the techniques used [53]. Here it diverges from the traditional Rogerian methods of open questions and reflective listening. MI was motivated to target behavior change of a problematic drinker, and its focus is on how to impact the client in a way that is not assertive or imperative. The key idea is that low motivation is not just a client problem, but a shifting state that is very sensitive to the behavior of the counsellor. Progress in counselling is more likely to occur if motivation to change is not imposed from without, but elicited from within in an atmosphere free of conflict [223].

At the heart MI is an attempt to have a constructive discussion about change in which the client drives the process as much as possible [223]. In an MI conversation, the counsellor will actively look for opportunities to explore ambivalence about, for example, drinking, and will try to understand what broader values and issues are important to the client. How the client's aspirations coexist or conflict with the drinking problem will often provide the fuel for decision-making and change. Its central principle, that motivation to change should be elicited from people, not somehow imposed on them, but gradually concretized from within. Upon this foundation of respectful collaboration, strategies and techniques are used to explore the person's values and goals and their relation to the addictive problem, and to elicit motivation for change from the client.

Yet this method is confrontational [223]. While in traditional alcohol counselling the confrontation often was overt, in MI the confrontation is intended to arise within the client. In fact, the probability of change increases with such discomfiture. Here, counsellors need to provide clear structure to the session, with their having a clear view about what direction they would like the client to take. This typically involves gently coaching the client to explore the conflicts. By summarizing these for the client, and giving the person room to reflect, the motivation to change is more likely to be enhanced.

It is imperative in MI that ahead of all skills and practice, the core concepts and principles that serve the spirit of MI are carefully observed firsthand. In short, MI embraces three concepts: the fluctuating nature of client readiness to change; the acceptance of change as a process and therefore viewing client ambivalence as a normal state and response; and the observable conflict toward change, i.e. client resistance. The methods that serve to practice these concepts include: empathic listening skills, eliciting self-motivating statements, and responding to resistance [223].

Even a quick and brief overview of MI gives a good description of its client-centeredness and its goal-directedness toward behavior change as a distinctive style of communicating with a client. Invariably, MI is all about supporting client autonomy throughout the interview process, and therefore clients are invited to a process of revisiting problem behavior and pondering on their desires, abilities, reasons and need for change [176]. MI introduces a variety of techniques, such as expressing empathy through reflective listening, communicating respect for and acceptance of clients and their feelings, and using open-ended questions, to allow clients the opportunity for self-reflection and exploration of their problem behavior [15]. Others include reflective listening and summarizing, within a nonjudgmental, collaborative relationship. MI practitioners also emphasize sincere affirmations, complimenting rather than denigrating, and listening rather than telling [15]. Even more additional strategies include having clients discuss a typical day or week related to problem behavior [16]. Rapport is built through reflective listening, enhancing the therapeutic environment. Though feedback to the client is allowed, such as information and advice, it is advised to ask for the client’s permission before doing so. A final technique involves exploring concerns that the client may have as a result of problem behavior. By discussing these concerns in detail and allowing time for self-reflection, practitioners may help clients progress through stages of change.

This work adopts MI as an effective means of communication for a reflection assistant chatbot. With MI skills and techniques, the chatbot is able to emulate a novice MI practitioner that tries to follow the client's life path toward positive change. While Rogers' person-centered approach gives abstraction of the chatbot persona in the self-reflection process, MI can define and provide concrete phases, skills and practice that may construct an interview session with a client. In this way, the client can revisit the heart of a problem and may consider possibilities and potential outcomes of change, the type of "transformative" reflection that may lead to change in behavior [247].

### B. Expressive Writing

Expressive writing was pioneered by James Pennebaker and was replicated in a number of studies. While it may seem distant from the Rogerian psychology at first, it has to do with it at the most fundamental level. In the 10th year of expressive writing, James Pennebaker said in an interview with Dennis King and Janice Holden, that the health benefits of expressive writing speak to the fact that "just being able to put together a coherent and meaningful story about the trauma is therapeutic if there is a caring or interested person to read it" [125]. Pennebaker also added that expressive writing hints at the role of a therapist "to create an environment where a person feels completely free to reveal what they are thinking and feeling, and allow them to put things together" [125]. Also in his words, "Carl Rogers was onto something in the development of his technique of letting clients come to some kind of understanding of the event on their own" [125]. In this manner, expressive writing shares the fundamental spirit of Rogers' client-centeredness in that therapeutic effect may already begin in the very unfolding of a client's trauma, in its process. Clients themselves may help in this process. The therapist just has to make sure to create a safe enough environment to tell the story. Pennebaker's stance has not changed since; in a personal conversation two decades later, he still holds expressive writing is a Rogerian method.

In this work, expressive writing is again actively pursued as a means of reflecting on self to promote mental wellbeing. As a matter of fact, it has long been a way of coping with trauma, "with or without audience" [205]. The fact that it is simple and does not necessarily require feedback makes it convenient for anybody to practice, because in its original form, the disclosure of a once inhibited traumatic event can be therapeutic. This work takes a step further to argue that expressive writing can accompany a virtual audience that is a chatbot. In fact, Radcliffe et al [214] have argued that an expressive writing activity is already a social one as it involves an implicit audience that is the researcher himself.

This work takes it further. It engages a virtual audience and an interlocutor in the process of disclosure, to lead and support the further scaffolding process of reflecting on unresolved stress. In this process, the disclosure becomes a narrative, from an "account" of what happened to an "anecdote" of what it means to the writer. In other words, it aims at the three conditions of expressive writing to ensure benefit. First, it supports the narrative-making by turning the solo writing activity into a conversation. Naturally it is a story-telling activity, rather than a formal written composition where one iterates rounds of revision for clarity and conciseness. Second, it makes disclosure a social activity, having a virtual audience as a chatbot. Therefore it is not assumed that the writing will be read, but it is read and told to a chatbot who seemingly understands or claims to understand what is to be told. To designate a reader in this way conveniently makes it easier for the writer to tell stories. Finally, in this social sharing of emotions the risk of disclosure is minimized, as the chatbot, ironically, cannot think. To disclose to a nonhuman agent makes it a distinctively different experience from sharing with family, friends, close acquaintances or therapists. In this manner, expressive writing transcends its original format and becomes an expressive conversation, where a user freely writes his or her stories of foregone misfortunes to a nonhuman companion. Furthermore, the process of reflecting on the trauma may take a different route, depending on the bot's guidance. The final subsection will come back to designing different reflection processes with the bot guidance.

In sum, the legacy of Carl Rogers' person-centered approach toward self-reflection has been inherited in motivational interviewing and expressive writing. While these have widely been used in palliative care, counselling and psychotherapy [286], this work proposes them as a means to support disclosure of emotional problems in reflection, with guidance provided by a chatbot. Toward this goal, a brief overview of natural language processing techniques for chatbot implementation is as follows.

#### 2.3.3.2. Chatbot Intelligence

The personal assistant chatbot to support individuals' self-reflection processes mostly match the descriptions of a social chatbot. It aims to be a "virtual companion" [244] to users by building an emotional connection and relationship. Moreover, it concerns not only the relational component but also a procedural component, to help users engage in the process of scaffolding their thoughts and feelings [247]. Thus this work introduces two design elements: emotional intelligence and procedural intelligence. The following will discuss what they are and how they can be achieved in the current state of technology.

### (1) Emotional Intelligence

Reflection assistant chatbots, as a type of social chatbots, aim to be a virtual companion for users in their very personal moment of reflective thinking. Therefore it is their primary goal to meet the user’s emotional needs [244]. Given the sensitivity and delicacy of the subject of reflection in the context of this work, it is also important that the chatbot ensures user safety and emotional security. Hence the emotional intelligence of a reflection assistant chatbot entails the following capabilities: empathy, social skills, and safeguarding.

#### A. Empathy

A social chatbot must have empathy [244]. It needs to be able to identify the user’s emotions and detect how they flow and change over time. This may include query understanding, user profiling, emotion detection, sentiment recognition, and dynamic tracking of user mood of [244]. Understanding of contextual information as well as commonsense knowledge is also critical.

Many therapy chatbots concern empathic responses. For example, Woebot [68] incorporated a “therapeutic process-oriented feature” that is empathic listening, and Wysa [115] included empathetic listening in their engagement efficiency criteria. Koko [174] aimed for an artificially simulated empathy. However, except for Koko, therapy chatbots show limited transparency in their design for empathetic listening skills. While such skills are included in the therapy techniques that they used, how the responses are put together within the bot remains unknown. As for Koko, it trained a machine learning algorithm on large-scale peer-support chat data. Yet in a naturally flowing conversation the machine-generated empathetic responses fail to catch up with human responses.

Carl Rogers emphasized on the concept of “accurate empathy” [162]. This is a therapeutic skill that includes a commitment to understanding the client’s personal frame of reference and the ability to convey the meaning back to the client via reflective listening [162]. This perspective-taking process encompasses an accurate understanding of both cognitive and emotional aspects of the client’s experience as well as attunement to the client’s unfolded experience [82], a feat practically unachievable by a machine agent. In this work, chatbot empathy skills will be adapted from established therapist behavior to ensure the conversation does not lead the user astray or interrupt the reflection process.

### B. Social Skills

Every user comes from a different background, interests, and needs. A social chatbot needs to have the ability to personalize the responses for different users [244]. It needs to generate responses that are appropriate, encouraging and motivating, and most importantly, fit the interests of a user. It needs to guide the topics of conversation and promote a connected relationship in which the user feels well understood. It should be aware of inappropriate information and avoid generating biased responses.

Woebot [68] prides itself on being a chatbot that speaks like the way humans do. The bot’s conversational style was modeled on human clinical decision making and the dynamics of social discourse [68]. The friendly way of speaking is almost a must for therapy chatbots, including Wysa [115]. However, most therapy-delivering chatbots focus on replicating a therapeutic session and earning measurable outcomes (e.g. [73,86]), which leaves the question of designing a chatbot that is sociable and amiable enough to convey genuineness, congruence and unconditional positive regard to users [221].

In this work, the social skills of a reflection assistant as a chatbot will emulate those of a psychotherapist taken from a practitioner's manual (e.g. [178]). Since therapist behavior has been studied as an important construct playing a significant part in therapy success [101,177], instead of generating fully-automated responses, this work maintains an adaptation of established therapist behavior to ensure consistent agent persona [144] and enough sociability for user engagement.

### C. Safeguarding users

Ensuring user safety and privacy is absolutely necessary for chatbots especially in healthcare services. Many chatbots give an initial session in which they direct the user to read and understand how they are going to keep their data secure and how, as machine agents, their services may be limited compared to natural human capabilities. It is also necessary that users are taught how to reach for help in cases of emergency. Hence it is important for chatbots that concern any aspect of physical and human health to closely abide by principles of ethical design.

Mulvenna et al [179] presented an ethics design manifesto to guide systems development. Their manifesto includes 12 principles including providing enough information for people to make informed decisions at every stage, and respecting people's right to choose how they engage with the product or service. Moreover, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [290] proposes a set of principles to guide "ethically aligned design." It includes ensuring the design does not violate human rights laws, prioritizes well-being in design and use, holds the designer accountable and responsible, operates in a transparent manner, and minimizes the risk of misuse [37].

The planned research conducts experimental user studies that deal with quite sensitive data. It is concerned with personal life histories and experiences that may have had impact on the user's health and wellbeing. Hence, this work will closely follow abovementioned guidelines to ensure user safety, and to respect user's right in choosing to engage and interact with the chatbot system. For any case of emergency, risk or potential danger, the lead researcher will closely maintain contact with health professionals at her institution for immediate help and care. In fact, the very design of chatbot utterances for empathy and social skills will be adapted from model therapist behavior, to minimize the risks of failures in such a sensitive context of emotional disclosure.

### (2) Procedural Intelligence

While social chatbots are mostly concerned with building a safe and connected relationship with users via chitchat, a reflection assistant chatbot should be able to lead and engage the user into the reflection process and help scaffold complicated thoughts and feelings on a deeply emotional experience from the past. Other roles also include envisioning the future once the unresolved past is taken care of. These processes do not occur at once; rather they occur in a linear, or sometimes in back and forth manner when a user is in an ambivalent state of mind in pondering on a change of action. The procedural intelligence concerns the following capabilities: contextual understanding, technical skills, and managing turns in talk.

#### A. Contextual understanding

To move from one step to another, it is important for a chatbot to catch the context of the conversation and respond appropriately. A perfect contextual understanding, however, is still a challenge. However, to capture the minimal context of the user input, the chatbot can capture the key words and phrases of the user input and retrieve its next response. Though it is natural for humans to identify and understand the key subject matter in a naturally occurring conversation, it is a difficult task for a machine agent. Over the course of natural language processing research, many have tackled this problem by preparing a predefined set of keywords and scripting matching rules.

The pioneering vision of Joseph Weizenbaum to create a talking machine in fact started out with a simple script of writing ELIZA. The first chatbot in history worked on the following technical solutions [275]: (1) the identification of key words, (2) the discovery of minimal context, (3) the choice of appropriate transformations, (4) generation of responses in the absence of key words, and (5) the provision of an editing capability for ELIZA scripts. The discovery of minimal context and maintaining it worked with extracting the pre-defined keywords in the script and giving them weight according to contextual importance. ELIZA responses were basically a reassemble of keywords and pre-defined sentences written in the script. Such a simple technique yet produced a powerful impact and changed history forever; following ELIZA, there were many other chatbots that followed suit (e.g. A.L.I.C.E., Parry).

With advances in natural language processing, the problem of catching up with conversational context and maintaining history is gradually being conquered [244]. However, it is still a grandiose mission for a machine agent to naturally follow up with a human conversation that transcends boundaries and is bound with complex nuances and homonyms. For the bot to stay on track and prevent contextual failures, this work inherits the legacy of ELIZA, with respect to its editing capability of the "script" [275] to build an intelligent enough chatbot assistant to maintain a minimal context of the conversation and manage its flow.

### B. Technical skills

To construct a positive reflection experience for a user, a reflection assistant chatbot can not only reiterate words of empathy, but carefully guide the user into a deeper thinking of the problem at hand. The chatbot can do this in various ways: it may ask users questions, show images and videos, or order specific actions to follow. Usually these skills come from theory and practice. Many psychotherapy chatbots emulate therapist behavior, such as giving directions and asking questions.

Many therapy chatbots, as mentioned above, follow behavior-based therapy techniques, e.g. cognitive behavior therapy (CBT), dialectical behavior therapy (DBT), positive psychology, mindfulness practices, and many else. This is so because giving users directions and instructions are much easier than further engaging in a naturally occurring conversation. Moreover, many research findings have reported promising outcomes of doing so (e.g. [68,86,115]). Nonetheless, many lack a design understanding to illustrate how they have selected, incorporated, modified, or arranged such techniques for the chatbot.

This study aims to present clear design strategies and processes of designing technical skills in a chatbot. It will provide design rationales and choices in implementing the skills, adapted in a chatbot conversation. In doing so, conversational strategies such as managing the beginning and the end, pause and turns will also be addressed.

### C. Managing turns in talk

Reflection may occur at an uncertain point in time, and its process may not always be linear [210]. Often the fluid and repetitive nature of reflection is what shuns users away from practicing it. Therefore, the reflection assistant chatbot should be able to lead a conversation that has a clear beginning and an end. In addition, a conversation that unfolds in a stepwise fashion can help a user follow a clear path in the reflection process. Many chatbots in fact have this "routine" in conversation, and managing the turns in talk is an important factor in designing for conversational user experience [173].

Recently, there has been an interdisciplinary endeavor in developing natural conversation framework (NCF) for bot services. It is concerned with the mechanics of how humans take turns and sequentially organize conversations, especially borrowing the findings in the field of conversation analysis (CA) [172]. In fact, with the proliferation of natural language processing (NLP) technologies, Moore et al [173] suggest there is an increasing demand for a discipline of conversational UX design. For better conversational UX, the NCF offers generic conversational UX patterns that are platform-independent and are inspired by natural human conversation patterns from CA, such as those of turn-taking or sequence organization [229,234]. The NCF so far has been implemented on both the IBM Watson Assistant and Dialog services.

This work follows the conversational UX patterns outlined in the NCF, including the design of turn-taking and repairs. Turn-taking is important in natural conversations to ensure both parties stay involved and engaged in the conversation. To enable an interaction where both the user and the bot actively participate in a conversation, turn-taking will be allocated in such a manner that both parties equally take turns. Turn-taking exceptions will apply when the bot initializes the chatting session where it has to give instructions for the users. Also, repairs are especially necessary for a chatbot that cannot inherently understand what a user really means. It is also possible that the user repeats himself or herself for any purpose such as clarification [234], hence the chatbot should be ready to recognize repeated turns by the user and respond accordingly. Usually, for general patterns of user questions and requests some repair responses can be in store to run the conversation. However, the repairs cannot be long, as the bot is incapable of engaging in a naturally occurring conversation. Turn-taking is again important to make the conversation stay back on track; even if the conversation fails at a moment, having the bot come back to the predesignated turn may remedy it. Other UX considerations, such as the length of conversation and number of speaker turns, will be informed by evidence from NCF practice and pilot iterations.

#### 2.3.3.3. Chatbot Skills

An examination of emotional and procedural intelligence for a reflection assistant chatbot outlines how it should engage users for constructing a narrative of an emotional experience. First of all, it should lead a stepwise conversation with phases in which users proceed with the agent. Within the conversation, it should be able to use technical skills and generate empathic responses to lead them into further reflection. Additionally, the chatbot should construct a safe holding environment for users to truly express their feelings and thoughts without a sense of inhibition, and the interaction should not be taxing or demanding.

These features require a set of carefully defined rules for a chatbot-directed conversation, where the bot follows stages to manage turns in talk and generates keyword-based responses to inspire further reflective thinking. This section will address strategies to design the rules for chatbot behavior for the three previously discussed reflection processes: transformative, explaining and exploring.

### (1) Transformative

Transformative reflections lead to change or understanding of what happened and why [13,69,170]. Here the interest lies in reflecting on past negative life experiences and gaining self-knowledge and self-insight that would positively lead toward wellbeing. Such a transformative effect, i.e. leading to a change in behavior or an insight, is always guaranteed, yet the reflection assistant chatbot can lead and aid users in such a path.

***Motivational interviewing.*** Here the chatbot can actively pursue a motivational interview. Motivational interviewing is not only helpful when users experience ambivalence toward change [286], but it can also help users proactively think about change and reorient themselves toward a better future.

***Shifting gears.*** Since chatbots do not naturally understand the flow of human conversations, it is more convenient for chatbots to take the role of asking questions and lead turns in a conversation to minimize the risks of failure. Asking questions, however, should not be done in a bombarding manner which may be overwhelming. The questions need to be organized for a specific purpose, such as reflecting on the past or thinking about future actions, so that users are not confused in the conversation.

***Building connections.*** It is also important that the chatbot builds a relationship with users. Reflection assistants are social chatbots, which are machine companions that care for users' emotional needs. In addition, motivational interviewing counsellors make it a crucial component that the counselor convey accurate empathy toward the client. Therefore the chatbot needs to provide emotional feedback and responses that would build empathic connections with users.

## (2) Explain and Explore

Reflections for explaining the past events and exploring untapped ramifications of the event involves equivalent emotional and procedural intelligence for a chatbot, but its core difference lies in that reflection takes place in user disclosure, rather than bot guidance. Here, the main responsibility of the chatbot is to nudge the user with a set of appropriate cues and wait for the user to unfold his or her thinking.

### A. Explain

Here the chatbot is mainly to ask the user what happened. The conversation needs not be long, yet the interaction can allow as much time as the user needs to revisit and recount the event. It is more important for the chatbot here to ensure the user at first for safety and privacy, and give clear instructions as to how to begin the process.

***Expressive writing.*** Expressive writing can create an environment where users are free to write about what happened in whichever form of narrative. It is a perfect vehicle for users to freely choose what to write and how to write about it. Moreover, expressive writing is not limited to any specific subject, so it provides flexibility in application.

***Engaging users.*** Expressive writing needs to happen in a safe environment where users feel relaxed and safe from any risk of breach in privacy. Ideally, the chatbot can introduce itself before users begin to write, engaging them in the process by relieving them of any anxiety and nervousness of disclosure.

***Giving instructions.*** The chatbot needs to provide instructions in a way that it is easy enough for users to understand what they are, but not so simple to make the whole process superficial. Also, it is important that the instructions are not too specific to restrain their own perspectives and insights in the thinking process.

### B. Explore

Exploring can take the explaining process further into thinking about what had passed without conscious awareness. In addition to the three strategies in explaining, exploring reflection employs two more strategies that can help users take a step back and reconsider what happened from a different perspective.

***Distancing.*** Once users write about a negative life experience, they can purposefully distance themselves from it by reviewing it. Here, the chatbot's role is to help them catch the unresolved thoughts and feelings surfaced in the writing process. It can do this by detecting positive or negative emotional words and phrases that describe the self.

***Switching perspectives.*** The chatbot can help users find a distance from and relive the past moment by switching perspectives. It can do this by analyzing what the user writes and extract the key persons and/or objects related to the event. As most traumas involve interpersonal or intrapersonal conflicts, the chatbot can look for textual cues and prompt the user for exploring on these untapped terrains of the event.

## 2.4. Summary

In sum, a review of literature on self-reflection shows that reflection can take a complex interplay of reflection and rumination, where the former can lead to self-insight and positive health outcomes while the latter to increased depressive symptoms. Although expressive writing has been widely established as a self-administered practice of reflecting on negative emotional experiences, its lack of concrete instructions could lead the reflector astray. Moreover, recent research has suggested that the social nature of disclosure can lead to better health outcomes. Putting these together, a disclosure-guidance space for chatbot-assisted reflection was suggested. The key idea is that guided disclosure could effectively prevent ruminative thoughts and lead the reflector toward constructive self-reflection. Reviewing reflection design and technologies that support reflection in HCI shows that reflection design has been much invested in inviting users to "revisit" the past, while others more interested in learning. Meanwhile, healthcare technologies have been designed toward delivering guidance to treat problem behavior. This study is, then, motivated to explore the less explored design space of self-reflection that explains, explores, and transforms the understanding of past emotional experiences for mental wellbeing, by designing chatbots that can be a virtual social companion for users in narrating the journey. In doing so, this work pursues Carl Rogers' client-centered thought and employs two descendant methods: motivational interviewing and expressive writing. Both methods have the client make their own narratives for their problems, with the counsellor providing support and indirect guidance. In other words, client-centeredness respects the client having their own right and strength to shift gears towards healthy personality change.

Since machine agents are inherently incapable of such intelligence, client-centered attitude suits the role. In this work, two reflection assistants that each adopts these methods will be illustrated in the following chapters, Bonobot and Diarybot. Technical background and skills required for the implementation of these bots are also illustrated in the current review of literature. The key design aspects in building reflection assistants are emotional and procedural intelligence. For emotional intelligence, it is imperative for the bot to maintain a minimal contextual understanding of the user narrative to lead and continue on within the conversation. Procedural intelligence is also necessary for the bot to develop the conversation in the right manner in order to help scaffold the user self-reflection.

# Chapter 3. Designing Chatbot for Transformative Reflection

This chapter begins with the design of *Bonobot*, a web-based chatbot application that encourages users' transformative reflection. It aims to achieve the goal by designing the conversational flow and bot utterances that would entrust the bot to *lead* the conversation on user's subject of perceived stress.

## 3.1. Design Goal and Decisions

Bonobot intends to encourage users' self-reflection in a way that invites to talk about their problems and any possibilities of changing their behavior. It uses motivational interviewing (MI) counsellor skills to help users consider the idea of behavior change for stress management. It helps users look at the problem at hand and prompts questions for them to ponder the idea of change. It was implemented as a web-based text messaging application that generates an automated motivational interview with graduate students for coping with stress at school. The topic of conversation was set up at the beginning in order to maintain minimum level of contextual awareness. This section describes the design decisions made to implement the bot and its conversation.

### (1) MI technical and relational components

MI entails technical and relational components [241]. The technical component includes counsellor verbal techniques (e.g. open questions, reflections) to facilitate change talk, where client argument for behavior change is formulated [152,166]. Also important to the efficacy of MI is the relational component [166], an empathic understanding experienced by clients in counsellor's helping them verbalize change. In this work, the technical component is translated in a series of MI skills to represent MI counsellor behavior that may evoke change talk. As for the relational component, since Bonobot is a nonhuman agent that cannot communicate empathy, such a feat is achieved by designing the interaction as follows: (a) contextualizing the chatbot responses to the graduate school context [177]; (b) not bombarding questions at the user [5,165]; and (c) using different combinations of MI skills [110] in the progress.

**(2) Chatbot responses**

To ensure that Bonobot provides responses and communicates them in a proper manner to qualify for both MI components, its responses took the following steps in preparation. First, model counsellor statements were collected from MI literature, such as: [40,49,59,137,150,163,164]. Second, the list was reviewed to gather more generic statements. For example, statements that are narrow-focused (e.g. "You've been homeless since April … what happened that made your anger reach a breaking point last night?" [59]) were removed, and portions of statements were blanked to be replaced with fillers from user input (e.g. "What was helpful when you feel (*client_input_emotion*)?" [49]). Third, to help the agent be more expressive of empathy with respect to the life of a graduate student, some statements were modified and replaced with more contextualized statements (e.g. What were your initial goals when you first planned for a graduate degree?).

A total of 220 prepared statements were later reviewed by certified therapists. They first referred to the Motivational Interviewing Skills Code (MISC) [110] to evaluate the responses. However, because this work primarily concerns the chatbot responses as MI counsellor language, they used the Motivational Interviewing Treatment Integrity (MITI) that refers only to the therapist behavior [166] from MISC. They coded each statement with the following MITI categories. Examples of the coded responses are provided in <Table 3.1>.

**Table 3.1. Examples of Bonobot responses by MI skill. Questions are classified into two different types to be served in the focusing and evoking stages of the conversation.**

| MI Skill | Q Type | Example Response |
|---|---|---|
| Giving Information (GI) | | • I am a chatbot that listens to your stories. |
| Questions (Q) | Focusing Questions (FQ) | • In what way does this bother you? |
| | | • How would you feel about that? |
| | Evoking Questions (EQ) | • How have you coped with difficult times in the past? |
| | | • What were your initial goals when you first planned for a graduate degree? |
| Reflections (R) | | • It's tough being a grad student. |
| | | • You certainly have lot on your mind. |
| MI-Adherent Statements (MIA) | | • Sometimes you show a determination that surprises even you. |
| | | • It seems like you are a really spirited and strong-willed person in a way. |

- ***Giving Information (GI).*** MI counsellor gives information to educate or provide feedback. As for Bonobot, it provides templated responses to address its role, privacy rules, and the beginning and closing of the session.

- ***Questions (Q).*** In MI, the counsellor is expected to ask questions that invite elaboration on the problem as well as questions that may evoke change talk. Bonobot uses both types according to the stage of the conversation: *focusing questions* (FQ) and *evoking questions* (EQ).
- ***Reflections (R).*** Reflections convey understanding, facilitate exchanges, or further add substantial meaning to what clients say. Bonobot uses simple reflections to acknowledge client remarks and lead the conversation.
- ***MI-Adherent Statements (MIA).*** MI-adherent statements include any counsellor behavior that is aligned with the MI approach. These are designed as affirmations, statements that encourage client positive traits in their articulation of change.

**(3) Conversational flow**

Though a human counsellor would spontaneously use MI-consistent skills, a fully natural language conversation is beyond current state of the art. Hence Bonobot is to deliver a summons-answer sequence, which can facilitate an exchange of user volleys [110] between the summoner and the summoned. Here, the summoning agent leads the conversation by asking questions, to which the summoned user answers. The agent, in turn, gives feedback. Such an orderliness continues with alternations of volleys between the two parties, as in an *a-b-a-b* formula [234].

For the conversation as a whole, Bonobot leads the four processes of MI [165]: Engaging, Focusing, Evoking, and Planning, as in <Figure 3.1>. In MI, Engaging builds a relational foundation with the client. The client's target behavior is determined in Focusing. In Evoking, change is explored, ideally with resolution of ambivalence. Planning consolidates client commitment and actions. For Bonobot, a set of operational aims are defined to reflect the four processes within the technical boundaries. In Engaging, Bonobot shares brief introductions with the user and gives instructions to use the chatbot. In Focusing, and Bonobot asks the user to detail their problem, possibly having them identify an inner struggle. This leads to Evoking, where Bonobot explores future goals with the user, affirming their own ideas for change. Lastly, Bonobot invites the user to ponder the overall session in Planning.

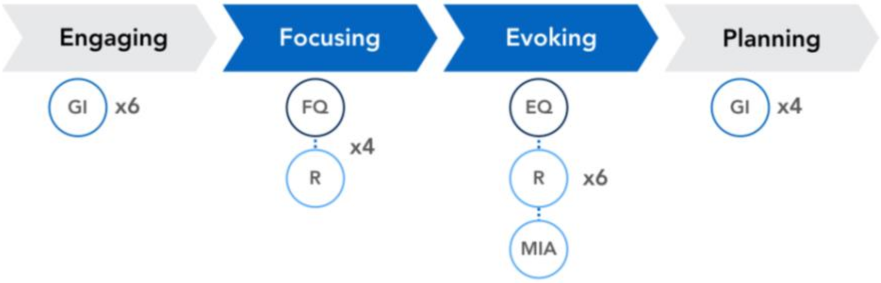
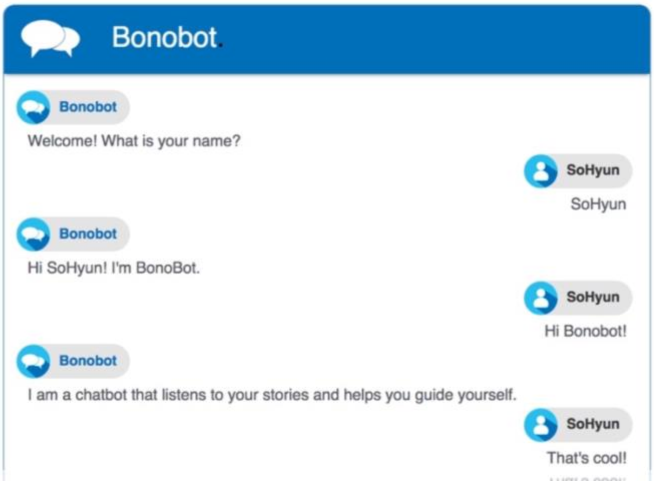


**Figure 3.1. The stages and sequence of Bonobot conversation. The circles below the arrows represent MI skills and numbers indicate the number of repetitions of the subsequence allocated for Bonobot in each stage.**

To reflect the aim of each process, the bot uses different combinations of MI skills in each stage. For the first and last stages, Engaging and Planning, Bonobot interacts with pre-defined GI's, to properly manage the beginning and ending of the conversation. In Focusing, FQ's are followed with R's to reveal and reflect on any struggle about the problem. In Evoking, EQ's are prompted to encourage change talk, and are followed by R's and MIA's to explore and affirm the idea of change. As advised by literature [5,165], no more than two questions are asked in a row. R’s and MIA’s are primarily placed after FQ’s and EQ’s as feedback. A series of pilot study sessions informed the final sequencing and turns.

## 3.2. Chatbot Implementation

Bonobot begins a conversation by introducing itself and informing the user of the conversation to be held. <Figure 3.2> shows the initializing screen of the bot. It then runs the conversation by generating responses based on keywords. Extending the framework of ELIZA [275], Bonobot identifies user keywords but generates responses in the form of an MI skill. Two modules, Flow Manager and Response Generator, run the system by executing the sequence and generating responses.



Bonobot
Welcome! What is your name?

SoHyun
SoHyun

Bonobot
Hi SoHyun! I'm BonoBot.

SoHyun
Hi Bonobot!

Bonobot
I am a chatbot that listens to your stories and helps you guide yourself.

SoHyun
That's cool!

**Figure 3.2. Bonobot’s initializing screen. Bonobot first asks for user’s preferred screenname to proceed. Here it welcomes a mock user named SoHyun.**

### 3.2.1. Emotional Intelligence

A pool of keywords and responses was prepared for Bonobot to run a context-aware, as well as empathetic conversation.

***Keywords:*** Most keywords in a reproduced ELIZA script [133] were replaced with ones extracted from online graduate student communities, r/PhD and r/GradSchool on Reddit, a social media platform. 1,000 posts from each were crawled and categorized by open coding for topics based on their title and content [29]. Any disagreement was resolved via discussion among coders. A word frequency analysis using tf-idf [217] yielded keywords by topic. The keywords were given weights from zero to five by an iterative process, so that ones with higher relevance would be weighted higher. Finally, a total of 70 keyword categories were prepared.

***Responses:*** Responses are generated from the pool of prepared MI statements, triggered by keywords from user input. For each keyword, a designated set of MI skills was allotted. Altogether, with repetition, a total of 209 FQ's, 188 EQ's, 166 R's, and 140 MIA's were prepared in the chatbot script. There were 8 GI templates to be used in the beginning and end of the conversation. In cases of zero identified keyword, extra responses were prepared to resume the conversation.

### 3.2.2. Procedural Intelligence

Bonobot's two modules, Flow Manager and Response Generator, were programmed using JavaScript. Python's Flask framework was used as the Web application server. The modules work together to run the four-staged conversation.

***Flow Manager:*** Flow Manager runs the conversation from one stage to another. At the beginning and end of the conversation, it assigns templated responses to lead the user into and wrap up the conversation. In between, Flow Manager counts the steps in the sequence so that the conversation follows the sequence. If a user does not respond in 10 seconds, it prompts an additional question from Response Generator.

***Response Generator:*** Response Generator identifies keywords and assemble responses <Figure 3.3>. For instance, suppose a user types in “I don’t know if I can graduate.” in Evoking stage. Flow Manager alerts the MI skill to be printed next (“EQ”), and Response Generator extracts keywords from user input (“I”, “know”, “if”, “graduate”). It prints the reassembled response (“EQ”; “What changes do you wish to make, if any?”) under the highest weighted keyword (“know (5)”). It never repeats the same response twice.

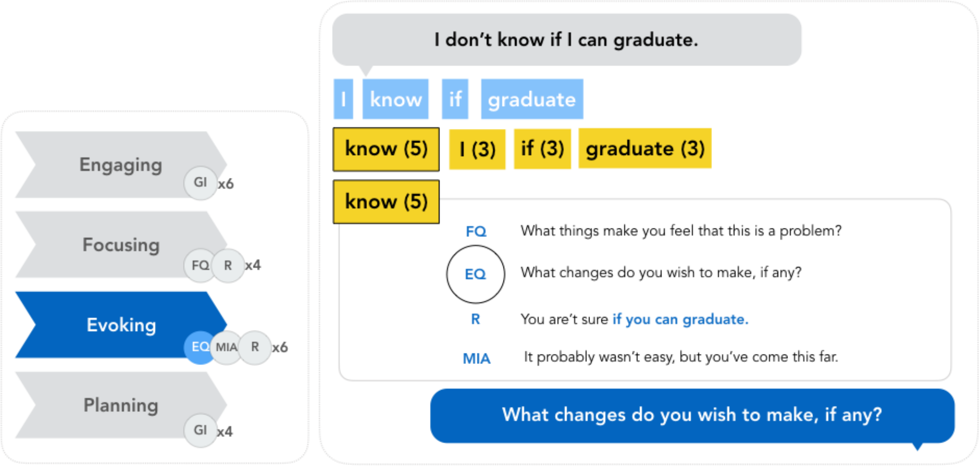
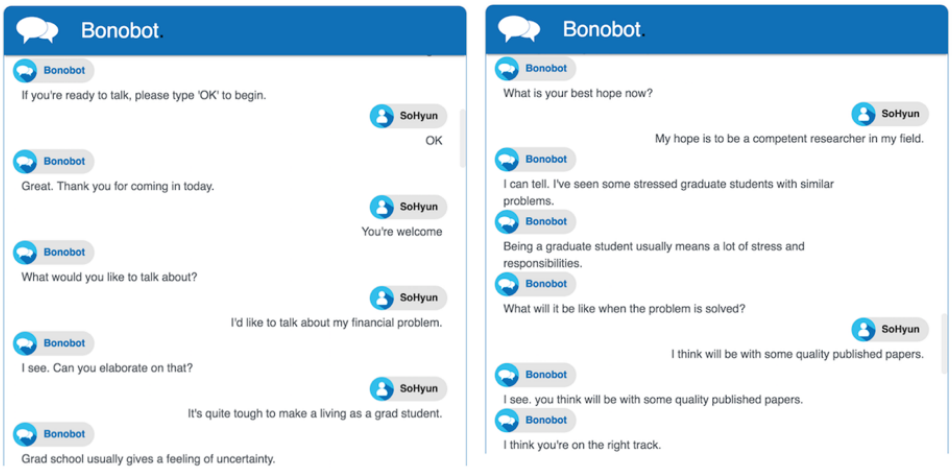

Engaging
GI x6
Focusing
FQ R x4
Evoking
EQ MIA R x6
Planning
GI x4
I don't know if I can graduate.
I
know
if
graduate
know (5)
I (3)
if (3)
graduate (3)
know (5)
FQ
What things make you feel that this is a problem?
EQ
What changes do you wish to make, if any?
R
You are't sure if you can graduate.
MIA
It probably wasn't easy, but you've come this far.
What changes do you wish to make, if any?

**Figure 3.3. A graphic illustration of automated response generation in Bonobot. Here, in Evoking stage, the Flow Manager designates the forthcoming “EQ” response to be retrieved next. Upon the user response, Response Generator analyzes the keywords by weight and selects the response in store.**

Pilot sessions with 10 graduate students (7 male) aged between 24 and 32 determined two distinct subsequences for Focusing and Evoking stages: (1) to encourage the user to share the problem, an FQ is followed by an R; and (2) to affirm the user's consideration of change, an EQ-R pair is followed by an MIA. In each stage, Bonobot is to repeat the subsequence 4 and 6 times, respectively (see Figure 3.1). This will make up a total of 8 and 18 Bonobot turns in each, with possible extra ones due to the 10-second inactivity rule. Finally, the conversation takes place on a text messaging app in an Internet browser, as shown in <Figure 3.1>. Excerpts from a mock conversation with Bonobot are illustrated in <Figure 3.4>.



Bonobot
If you're ready to talk, please type 'OK' to begin.

SoHyun
OK

Bonobot
Great. Thank you for coming in today.

SoHyun
You're welcome

Bonobot
What would you like to talk about?

SoHyun
I'd like to talk about my financial problem.

Bonobot
I see. Can you elaborate on that?

SoHyun
It's quite tough to make a living as a grad student.

Bonobot
Grad school usually gives a feeling of uncertainty.

Bonobot
What is your best hope now?

SoHyun
My hope is to be a competent researcher in my field.

Bonobot
I can tell. I've seen some stressed graduate students with similar problems.

Bonobot
Being a graduate student usually means a lot of stress and responsibilities.

Bonobot
What will it be like when the problem is solved?

SoHyun
I think will be with some quality published papers.

Bonobot
I see. you think will be with some quality published papers.

Bonobot
I think you're on the right track.

**Figure 3.4. Excerpts from an example conversation with Bonobot. The bot is having a mock conversation with a user named SoHyun. The conversation on the left shows the Focusing stage, where the user shares the problem. On the right is the Evoking stage, where the bot invites the user to consider making a change.**

## 3.3. Experimental User Study

An experimental user study was designed to investigate (1) the conversational user experience in terms of self-reflection; (2) the impact of design strategies on their experience; and (3) their needs for better support.

### 3.3.1. Participants

A recruitment ad for volunteers was posted on a Seoul National University online bulletin. A total of 30 full-time graduate students were recruited. The inclusion criteria were that they could (1) communicate with the chatbot in English, (2) share their concerns about being a graduate student, and (3) participate in an interview about the chatting experience.

### 3.3.2. Task

An online chatting session was prepared to invite users to talk with Bonobot. To capture the participants' reflection within the conversation with Bonobot in-depth and with detail, the participants were invited for a post-hoc semi-structured interview upon completing the conversation with Bonobot.

### 3.3.3. Procedure

Participants were invited into a room with a comfortable chair, a big table and a laptop computer. A laptop was used instead of a user's mobile phone for consistency and screen convenience. They answered a survey of demographic information and the Perceived Stress Scale (PSS-10) [46]. The experimenters left the room while the participants chatted with Bonobot. They returned on the participant's notice and conducted semi-structured interviews, reviewing the conversation on the laptop screen. The entire process was designed for an hour, and participants received a $10 beverage coupon as a reward upon completion.

### 3.3.4. Ethics Approval

Before they gave consent, all participants were informed of the purpose and procedure of the study and that they could resign from it at any point if they felt uncomfortable. The study conformed to the principles of scientific research with human subjects. All procedures including the surveys and interview guidelines were submitted to and approved by Seoul National University Institutional Review Board (IRB No. 1708/001-018).

### 3.3.5. Surveys and Interview

A qualitative study was conducted to gain a detailed understanding of the users' reflection experience with chatbot-guided MI skills. This included a brief demographic survey to measure users' perceived stress, semi-structured interviews to collect their evaluations of the experience as well as their perceptions of the chatbot design. The surveys were taken at the beginning of the study, asking participants' age group, gender, and perceived stress. Perceived stress was measured by PSS-10 on a 4-point Likert scale. It is one of the most widely used instruments to assess one's perception of stress in the course of the previous month, and higher PSS scores are associated with higher risks to negative health conditions [46]. The collected scores were computed for mean and standard deviation values.

Upon completion of the task, participants took part in interviews with their consent for audio-recording. The interviews were anonymized and transcribed for analysis. The interview questions included themes on user-chatbot conversational encounters for reflection, user perceptions of chatbot interactions, and further engagement. Detailed questions were asked for an in-depth elaboration on the conversational experience, and notes were taken on participant-indicated conversational happenings. The interview was designed for 30 minutes and did not exceed 40 minutes at most.

Interviews were analyzed via a six-phase process of thematic method by Braun and Clarke [29]. First, all of them were transcribed verbatim. The transcripts were reviewed and segmented by each anonymized participant, using Optimal Workshop's Reframer online [287]. Second, a process of open coding was conducted to generate initial codes, with free-format labels. Two coders generated 10 free-labeled categories for 387 segments via discussion, and the Cohen's Kappa statistic (*k*) was 0.9201. The disagreed labels were resolved via further discussion. Labels were then again reviewed and renamed for initial codes. Codes were once again reviewed to search and define themes. As advised by Braun and Clarke [29], themes that merely reiterate the interview questions were avoided, but those that can reveal the depth of the data were reviewed and redefined in iteration. The final themes are: boosting motivation; wanting accurate empathy; and needing superhuman intelligence.

## 3.4. Results

In the conversation, participants preferred Bonobot's questions to its feedback. EQs were a good means of reflecting on themselves and for some, an instrument for motivational boost. Some, not all, of Bonobot's responses for affirmation and empathy were appreciated. Better design strategies for long-term engagement for transformative reflection were also suggested.

### 3.4.1. Survey Findings

Participants were in their twenties (n=20) and thirties (n=10), and a half of them were male (n=15). The average PSS score was 22.5 (SD=5.0), higher than the norm in the region [46]. Conversation topics included lack of confidence in research (n=12, 40%), psychological burden of writing theses (n=5, 17%), financial constraints (n=3, 10%), uncertainty about the future (n=3, 10%), work-and-life balance (n=2, 7%), people skills (n=2, 7%), and other (n=3, 10%). Topics of concern were mostly in agreement (90%) with the themes discovered from the content analysis from the Reddit posts.

### 3.4.2. Qualitative Findings

**(1) Evoking questions are a motivational booster**

Participants mostly favored the way Bonobot kept asking them questions. It felt like they were being heard (n=18). In particular, they preferred the EQs in the third stage as they were "something new and interesting" (P2) and "triggered inspiring ideas" (P13). P1 said he liked them as "the questions were profound [...] I had to think deep down and discover the answers inside." Questions such as "What can be some of the good things about making a change?", "What do you wish to be different?", "How would you like things to turn out for you?", and "What could be the next step now?" triggered to think "who you really are and what you really want" (P12) and "what needs to be done to achieve your goals" (P13). P11 said that "it was really the third stage" that "felt quite convenient to draw something out" from him.

However, they did not like questions that made them reiterate their answers. Though Bonobot never repeated any questions, participants felt that some questions were essentially the same and repetitive, which was a bit annoying to some participants (n=7). In addition, some questions did not feel productive when they were not relevant to the context of their problems and spanned too grand a scheme of things. P4 pointed out an example: "Bonobot asked me what I would have chosen to do if I did not pursue a graduate degree. But I've never thought of such an idea—something other than grad school. I'd say that wasn't quite helpful." P30 added that:

> *"The questions aimed too broad a range that each question could entail a whole lot different story by itself. I think the conversation was too short for that."*

The Evoking stage offered a chance to reflect on themselves (n=19). P4 said "I think it was a time to reflect on me and my situation. [...] I liked I had the chance to rediscover myself with my own words." For P22, it was:

> *"It's a Socratic questioning, isn't it? In the end, you answer for yourself. Bonobot asks me questions, and by answering them, I get a better understanding of myself."*

P23 also said it was like "a catalyst" that kept nudging her to think about herself and her life. This self-reflection spurred a sense of motivation (n=11) that had been "sort of buried in" (P27). P30 said she could gain a motivational boost:

> *"You know, I'm always like, 'what am I to do now?,' 'this is too hard,' and 'I can't do it.' Now, I have this question inside, 'so how do I want this to be resolved?' This moves me forward. I feel like I need to do something about it."*

However, evoking ideas about change was not to everyone's liking (n=6). Some participants expressed their distaste about the idea of having a directive conversation, especially hypothesizing that the user should have something to "change about." P18 said he disliked the idea of "having a conversation with a purpose." P17 said:

> *"Bonobot clearly had an idea about what it wanted to hear from me—something positive—and it wanted me to say it, which made me feel like Bonobot had the lead over the conversation, not me."*

In addition, some participants had trouble facing themselves in such a conversation. To them, the problem at hand felt so great that they could not attempt to think about making a change about it. Maybe they knew it, but they rather wanted emotional support when they decided to share their problem. P20, in particular, had a feeling that she did not fit in the evoking stage of the conversation, while she like the focusing stage:

> *"I guess I am not exactly sure* [about the evoking stage]. *I know I need to change, and I know what I need to do to make that change happen. But that's causing the stress! But Bonobot's questions felt like it was trying to remind me of that, instead of letting me vent."*

**(2) Accurate empathy can lead to disclosure**

In between questions, Bonobot gave feedback as in reflections and affirmations. Most participants (n=21) liked Bonobot saying "such sweet words" (P21). P9 said, "I thought Bonobot used words of empathy really well, you know, even if some felt like templates, they were good." However, it was not quite up to their expectations (n=13) due to timing and contextual awkwardness. P29 said, "here, Bonobot said the right thing, but it doesn't fit into what I said. I had to doubt whether it really understood me." P2 also said that "I know it tried to encourage me, but sometimes it did at the wrong time, which made me wonder if Bonobot was to encourage me no matter what." P10, P16, and P25 said they anticipated something more than Bonobot simply repeating what they have said. P27 recalled, "It was not bad, but it can be weird… you don't really recite word for word when you talk."

Some feedback, such as "I hear your struggle.", "You certainly have a lot on your mind.", "That's understandable.", and "You're not the only one in this." also felt rather banal to feel fully understood. P4 said "Some felt like they were just there because they had to." P21 and P24 found it odd that Bonobot repeated similar expressions in the conversation. In addition, they were just “too nice” (P2). P1 added that “you know, if you were to talk with a human being, you wouldn’t really say the nicest things throughout.” P4 still “appreciated the niceness” as he rarely has a chance for those words. However, for P17, P20, P24, and P28, words of empathy only echoed what they could expect from anybody around them.

Still, they appreciated Bonobot’s taking the role of a nonjudgmental listener (n=11). P3 thought she made “a virtual friend who listens to [her] and tries to understand [her].” It was essentially a private conversation where they could talk about things that they cannot usually open up to their family or friends (n=8). P19 said he could feel more relaxed talking to Bonobot “for [he] did not have to worry about what Bonobot would think of [him].” P24 said he shared the same subject that he did with a colleague, which ended up in an argument. He felt better talking with Bonobot “for [Bonobot] does not have any interests that may conflict with mine.”

In particular, participants preferred Bonobot’s words of empathy that concerned the life in graduate school (n=13). P3 said that Bonobot seemed to know “what it is to be a graduate student.” For P4, Bonobot’s empathetic responses were not only an encouragement but also an instrument to build on his story: “I liked that it [Bonobot] seemed to understand what I said about my advisor.” He wanted to continue on, yet Bonobot went on to the next question. P6 felt touched when Bonobot said, “Don’t let it discourage you,” to his disconsolation with his progress. When Bonobot asked P7 about the past achievements to which he had none, it replied “That’s okay,” with which he felt touched and thought that he could tell something more to it. P23, in particular, was pleased with Bonobot saying “A lot of graduate students suffer from variations of the same problem.” For others, the graduate school-related responses felt clicked with P15, P9, P19 and P30, such as: “That’s exactly how many students feel during their graduate program.”, “Grad school usually gives a feeling of uncertainty.” and “It’s tough being a grad student.”

**(3) User expectations are heightened**

Participants expressed the need for both nonhuman and human intelligence from Bonobot. The nonhuman feature concerned machine intelligence surpassing human capabilities of processing and searching for information. For example, talking about making a change brought about a need for solutions (n=9). P7 said “I need more information I guess, about the problem I talked about.” P8 said the conversation would have helped “if the chatbot told me how to write a paper.” P26 made a specific suggestion:

> *“A chatbot can deliver news articles or life tips. You know, say I have a sleeping problem. It can give me various suggestions, such as music recommendations, health information, or other tips found online.”*

A few participants also indicated a need for making personal agenda (n=5). P13 recommended that Bonobot ask more detailed, branching questions such as “How much financial aid do you get?” or “What are the current career options?” that are tailored to his situation. P5 said “I would appreciate it much more if it organized for me a list of reminders and agenda from what we talked about in the conversation.” P14 and P22 also suggested that offering specific action items would be helpful. P26 said that planning an agenda with Bonobot would potentially inspire a sense of partnership. This indicates that the conversation with Bonobot triggers an idea for taking an action, and users attribute the inception of the idea to the bot and ask for more help.

They also demanded Bonobot to be more human-like, helping them feel more like they are heard, such as more personalized questions about their life, “as if [they were] talking to a human being” (P20). Addressing such an elaborated context of their problems would signal “a continued relationship” (P5) with the chatbot (n=7). They would like to be able to use Bonobot if they felt the need in the future, for they thought it was “quite useful” (P13) to organize their thoughts and review their current motivations. Moreover, they wanted more emotional responses that are appropriately contextualized to their input (n=13). P12 said, “This chatbot says some sweet words, but I would prefer more emotional expressions like, ‘I can’t believe that happened to you!’ or ‘That must have been very hard on you,’ things like that.” In other words, they wanted Bonobot’s responses to be more natural that they can feel supported by a real human being. P27 put it this way: “You know, I’d like words that are more for me and me only, not like the mundane ones that anybody can say to everyone else.” For P25, more personalized responses would have helped her feel more empathized:

> *“What if it said something more concrete, like, “You must have had a hard time communicating with your advisor all this while,” instead of just a simple expression of empathy? Then I would think that it really understands my feelings.”*

## 3.5. Implications

In Bonobot study, motivational interviewing (MI) skills were used to design a chatbot conversation for encouraging a transformative reflection. Qualitative findings show that participants were able to engage in a reflective process in which they could look back on sources of unresolved stress in graduate school, how they would like it to be solved, and what actions they might take. More specifically, participants preferred the evocative questions that prompted participants to think about desires, abilities, and need for change. Reflections and affirming statements were appreciated mostly when they conveyed an accurate understanding of the participant's situation. Finally, participants requested more emotional intelligence as well as both sensemaking and decision-making support, some even entrusting the problem to the bot. Based on these findings, this section discusses the implications of designing transformative reflection with chatbot guidance, and how it may create tensions in user control and autonomy in human-AI interaction.

### 3.5.1. Articulating Hopes and Fears

The essence of MI is to invite the client to resolve ambivalence for behavior change and elicit hopes and fears about it [165]. In a way, it is aligned with transformative reflection in that it also pursues eliciting change in behavior or mental schemas [247]. In fact, an earlier work by Slovak et al [247] has already established the need for scaffolding the reflection process, just like Bonobot's structured conversation. However, in their work, such process accompanied an interplay of curricular components in a learning environment, which takes time and can only be conditioned on the "right sort of experiences" [247]. On the other hand, with Bonobot, participants naturally engaged in a conversation where they could talk about their problems and further delve into articulating their hopes and fears toward making a change, with the "right sort of guidance."

Participants regarded evocative questions as constructive means to revisit their source of stress, leading to the idea of change. In the interview, participants who were able to consider change were willing to share their immediate plans to follow through. However, for a few, the distaste and even resistance to problem solving actions was also observed. We find both types of reactions in alignment with previous work [166], and highlight the potential influence of change talk, especially in terms of transformative reflection. While earlier work [69,247] argue that self-reflections need to aim for transformative reflection, the findings of the Bonobot study indicate that not everyone is ready for it, or needs further support to overcome fears. In their articulating problems and distress associated, individuals are at different stages of coping, as well as have different ideas to approach them. This indicates that reflection unfolds in many possibly different ways. Though the bot offered only one structured conversation to the participants, the fact that participants could bring up their own problems and the bot responded to different keywords helped. Moreover, that the conversation took steps within the four MI processes was helpful for participants to think why the problem is stressful and how they want it to be resolved. Though the conversation did not fit everyone's preference, it did achieve a feat: it invited everyone to think about what can be done about the problem. In terms of Lazarus and Folkman's transactional model in coping with stress [136], this process is referred to as a cognitive reappraisal. Positive reinterpretation is not only a means to reduce emotional distress but also a form of active, problem-focused coping [38]. In reflecting on the problems, not only the emotional side but also the cognitive side of dealing with them can help conceive the idea for change, in a more concrete and action-oriented manner, as testified by the participants in the interviews.

### 3.5.2. Designing for Guidance

The Bonobot conversation consisted of two different types of dialogue acts: questions and feedback. The questions were a convenient means to disclose problems and encouraged to reflect on change for most and for some, inspired a motivational boost to take an action. However, feedback received mixed opinions. Some participants were pleasantly surprised that the bot could imitate human empathy and give them encouragement, while others were rather disappointed that it did not feel quite genuine. Due to its technical constraints as a chatbot, some Bonobot responses were ill-assembled and could not correctly fit to user responses. However, even when Bonobot gave well-suited feedback, some participants were suspicious of auto-completion, or a templated response that would be retrieved no matter what. An accurate empathy requires a profound contextual understanding, which is hardly achieved by chatting robots with limited natural language capabilities.

Nonetheless, when Bonobot used graduate school-related responses, they clicked with the participants. Those responses made some feel like "the bot knows what it's talking about" (P28) and assume its contextual intelligence. It shows that not only questions but also carefully designed responses may help users feel understood and trust the bot, increasing compliance to the bot guidance. Those who experienced this with Bonobot tried to take its questions seriously, thinking that there must be a reason why it asks such questions and trying to elaborate on their point. This is most evident with P13, who was deeply moved by the bot's motivational questions. This finding illustrates that for the chatbot guidance to work toward the most benefit, the users need to trust the human purpose in the algorithmically generated guidance. Showing accurate empathy is no different matter; the users need to be assured of the bot's emotional and machine intelligence to feel a safe enough environment to share personal stories and pursue change. Though this is an even tougher challenge in technology, Bonobot could achieve it to a certain extent with catching the keyword and setting the scope of the topic in the conversation. For a more flexible and scalable conversation, advanced natural language processing techniques may achieve more sophisticated expressions of empathy and evocations.

### 3.5.3. Rethinking Autonomy

One interesting aspect of participants' perceptions and evaluations of their conversations with Bonobot was that they would demand more intelligent support: statistical information for job seekers with graduate degrees; recommended sleep habits; stress diagnosis; solutions for personal problems; tips in writing theses and research. What's more interesting is that they wanted them personalized and tailored to them, assuming the bot should somehow be able to "catch it" (P12) and make instantaneous amends for them. Or, some participants just wanted it to "tell [them] what to do" (P20) in the face of such a stressful situation.

This observation makes a grave implication in designing intelligent agents that engage in transformative reflection. So far, intelligent agents are mostly targeted on supporting task-based inquiries: booking flights, setting an alarm; online shopping orders; and many others alike. These are rather simple and repetitive tasks that do not violate users' self-determination but are only help increase productivity and efficiency of time. However, for users to ask the agent to analyze their problems and provide appropriate support hints at their dependence on the agent for decision-making on important life matters [1]. It poses critical questions to be answered by both interaction designers and HCI researchers. When agents like Bonobot guide a user into transformative self-reflection, how much should they be held responsible for user decisions in making change? How much support can be automated and in what ways? In human-AI interaction, under-reliance represents inefficiency, while over-reliance represents risk [1]. While users would like to take control in human-AI collaboration [186], in the face of distress users may decide to submit to machine intelligence out of helplessness. In therapy, the same is also observed in individuals experiencing ambivalence toward change [15]. The role of a human counsellor is to provide information but not giving solutions or answers [5]. Perhaps the real challenge is to manage users' expectations toward AI, while still encouraging continued engagement. It is a critical tension in designing intelligent agents that are to engage in the thought processes in reflection. As more and more agents deliver natural language conversations and nonverbal interactions, expectations may surge beyond control, eventually losing user interest. The Bonobot conversation was designed in a way that the bot assumes the role of a hypothetical MI counsellor, and so did all participants. The professional persona for the bot, therefore, risks users compromising control or failing their expectations. Thus it may be advised to design the bot guidance in a way that it entitles the user to be the one best knowledgeable of his or her problem. For example, instead of designing for speculative empathy, the chatbot would rather ask more detailed questions of past actions or make references of user input in the history of conversations.

The lessons learned in the Bonobot study are that while detailed guidance can inspire a transformative reflection to inspire an idea of change, it can only deliver a less flexible interaction and perhaps may fail user expectations when the user demands are not met in a long run. It is suggested, in designing for chatbot-guided reflection on life's most difficult events, to design the chatbot interactions in a way that may encourage perceived control of the situation by promoting more disclosure. The following Diarybot study will investigate this idea further.

## 3.6. Summary

In this chapter, the design, implementation as well as experimental user study findings of *Bonobot*, a chatbot that encourages transformative reflection, were presented. Bonobot delivers a structured conversation of carefully sequenced motivational interviewing (MI) skills. The findings indicate that a delicate, well-organized guidance can lead an increased self-disclosure and inspire self-insight. Some gained a motivational boost for their graduate career from talking with Bonobot. Participants could also gain moral support, and the evoking questions used to invoke the idea of change worked for most of them. They subsequently demanded the bot to be able to offer solutions and more intelligent functionality such as information search and agenda-making.

The findings carry grave implications for the HCI community. First, unlike many conversational agents that serve task-based queries, this work shows potential in designing the chat interactions that can bring about a certain type of thinking process, perhaps a difficult one, too, on life's pressing stressors. The study also shows promise in designing the bot guidance to play an effective role in leading the participants into the purpose-driven conversation such as a motivational interview. However, that the agent led the conversation could inflate user reliance on the machine, which may risk self-determination in the long run. Moreover, the bot's resemblance of a counsellor increases user expectations for its capabilities, resulting in a demand for even super-human intelligence.

That Bonobot supported only a limited conversation on a graduate school life matter did not contribute to meeting the expectations, either. Life's stress is often multi-faceted and does not have a clear-cut answer, even from a human expert. It is therefore needed to design a chatbot that can support a moderate level of guidance yet encourages more user narrative, for users to take charge of their decision-making and to allow themselves an opportunity to explore and embrace the hard times.

# Chapter 4. Designing Chatbots for Explaining and Exploring Reflections

This chapter introduces *Diarybot*, a chatbot that serves two different chat interactions for explaining and exploring reflections. The primary purpose of both Diarybot chats is to encourage users' own narrative of the life's most difficult event, with different levels of bot interaction. In other words, Diarybot invites the users into their own making of narrative and *follows* up with it.

## 4.1. Design Goal and Decisions

Diarybot supports the user's self-reflection in a way that the user can recall a negative life experience and explore undiscovered meanings from it. It aims to help users reflect by rethinking the event from a different point of view, with two different types of chat: Basic and Responsive. In Basic chat, Diarybot asks users to recall a past life trauma. In Responsive chat, it invites users to recount the trauma, and walks them through a series of follow-up prompts that are put together with algorithmically selected keywords taken from the user's writing.

Diarybot delivers a chatbot-adapted, Korean-translated expressive writing instructions [196,201,204], which is the most appropriate writing procedure [105] that invites a user to describe one of the most difficult life events from the past. The expressive writing prompt in Pennebaker's words [200] is as follows:

> *For the next four days, I would like for you to write about your very deepest thoughts and feelings about the most traumatic experience of your entire life. In your writing, I'd like you to really let go and explore your very deepest emotions and thoughts. You might tie your topic to your relationships with others, including parents, lovers, friends, or relatives, to your past, your present, or your future, or to who you have been, who you would like to be, or who you are now. You may write about the same general issues or experiences on all days of writing or on different traumas each day. All of your writing will be completely confidential.*

Diarybot's two chats share the same expressive writing instructions, but the chat interactions following the writing are different. Responsive chat expands the whole conversation with a series of follow up prompts. <Figure 4.1> illustrates a diagram of two chats in the Diarybot system. Next describes design considerations and decisions made for each chat.

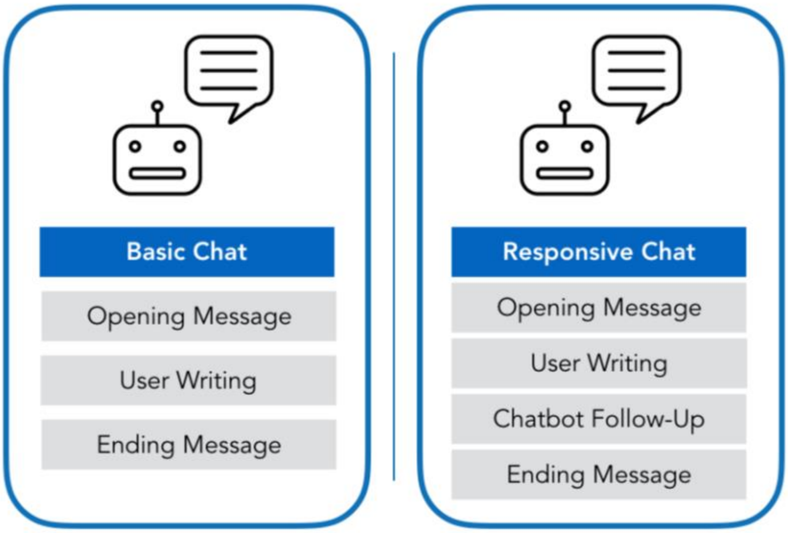


**Figure 4.1. The interaction procedures for Basic and Responsive chats of Diarybot. While the Basic chat only allows writing about a trauma, the Responsive chat adds a follow-up process to it.**

### 4.1.1. Design Decisions for Basic Chat

In Basic chat, the expressive writing prompt is the major interaction to bring about an explaining reflection, for putting trauma into words is already a rediscovery process [248]. By telling a story to a chatbot, the user explains what happened and/or how he or she felt at the time, in whatever desired manner to write. Unlike expressive writing that is a solitary process, Basic chat involves an audience that is Diarybot. The presence of Diarybot creates an effective environment for constructing a "narrative," which is an essential element in expressive writing's benefit [250].

### 4.1.2. Design Decisions for Responsive Chat

The Responsive chat provides a follow-up interaction upon user's writing about a trauma to invite an exploring reflection. The interaction consists of five prompts in open question format, either in templated form or responsive form depending on whether the bot can retrieve algorithmically selected keywords from user's writing. Below describes the design decisions and processes.

**(1) Open question prompts**

One important aim of designing the Responsive chat was to help an exploring reflection on a possibly traumatic experience. In psychotherapy, therapists usually respond to a mental health client by using a variety of techniques such as open questions [102]. In Responsive chat, a total of five open questions were designed to invite users to explore their feelings and find alternate meanings of the event written in the chat. Open questions are a useful technique not only to lead users into further thinking [100] but also lead a conversation by calling for an immediate response from a user [229,234]. The questions are designed to inspire the recognition and interpretation of an emotional event [140,181], which involves emotional, social and self-awareness.

### (2) Data-driven prompts

The follow-up questions are guided prompts that both reflect the context of the user's writing and promote his or her emotional, social/situational, and self-awareness. Diarybot selects sentiment and relationship keywords from the user's writing for a set of responsive questions. In case it does not find any keywords, the bot uses a set of template questions. Design intentions for each set of guided prompts are illustrated in <Figure 4.2>.

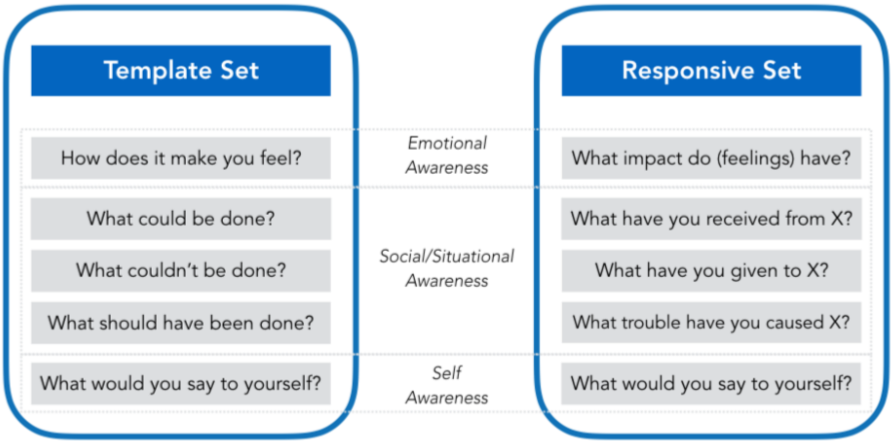

Template Set
Responsive Set
How does it make you feel?
Emotional Awareness
What impact do (feelings) have?
What could be done?
What couldn't be done?
What should have been done?
Social/Situational Awareness
What have you received from X?
What have you given to X?
What trouble have you caused X?
What would you say to yourself?
Self Awareness
What would you say to yourself?

**Figure 4.2. Design intentions for guided prompts in Diarybot's Responsive chat. Responsive set on the right uses keywords for emotional and social awareness. Each box contains design intention for the prompt. Template set on the left indicate that no keywords can be retrieved from user writing. Instead of social awareness, it provides prompts for situational awareness.**

***Emotional awareness:*** In psychotherapy, a recognition of feelings needs to precede any interpretive actions to be taken [273]. To support users to be more emotionally aware, the first follow-up prompt was designed to help users recognize the feelings explicit in the writing and consider their bodily and psychological impact.

***Social/Situational awareness:*** To assist users to take a step back from, and be able to better interpret, their emotional experiences, three questions are borrowed from a Japanese meditation practice called Naikan. It promotes self-understanding by asking three simple questions [239] that invite the person to reflect on the relationships with key person or any being in the situation at hand. The questions are as follows:

- What have you received from X, if anything?
- What have you given to X?
- What troubles, if any, have you caused X?

In Naikan, "X" is the subject of a trainee's choice. Most start with their own mother [239]. In Diarybot, X is replaced with a key person or relation identified from the user's writing. This is sought to maintain a minimal context and support a continued thought process, and to seek potential switching of perspectives to promote any potential health benefit [23]. If Diarybot does not find a relationship keyword, it uses an alternative set of template questions that do not require "X" but focus on the self for situational awareness:

- What could be done better?
- What couldn't be changed?
- What would you like to have done?

***Self awareness:*** Finally, Diarybot asks the user to leave a message for self. This is intended to invoke a summarization of the chat so as to facilitate a re-construction of meaning from the interaction [78].

### (3) Varied prompts

The prompts were reproduced five times with synonymous phrases to prevent repetitiveness but maintain consistency for user engagement. As a result, there are five sets each for responsive and template prompts. <Table 4.1> illustrates an example of a responsive set.

### (4) Length requirement

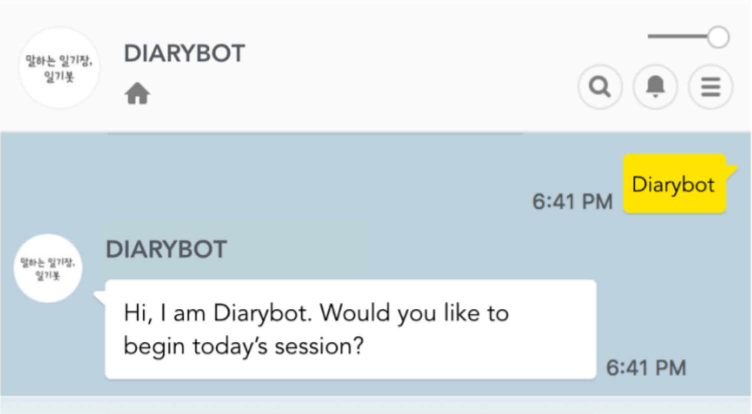
Finally, Diarybot required the user to write no less than 100 words in Korean. This was to make the sentiment analysis process run smoothly. This requirement is from Pennebaker's expressive writing. However, participants did not have much difficulty with this requirement as the event to be written often needed to exceed the requirement. Users were informed of this requirement at the beginning of the conversation.

**Table 4.1. Responsive prompts in Diarybot's Responsive chat. There is a total of 5 questions in the follow-up conversation. Each placeholder (A, B, C, D and X) is to be replaced with words retrieved from the user's writing content with Diarybot's skills.**

| Target Awareness | Example Prompt |
|---|---|
| Emotional | • In your writing today, feelings of A, B, C, D were found. What impact have they had on your body and mind? |
| Social | • Now let's think more about X. What have you received from your relationship with X, if anything?<br>• Then what have you given to X? Even tiniest things are welcome.<br>• Finally, what troubles, if any, have you caused to X? Most people find this question hard, but please take your time. |
| Self | • Before we wrap up our writing today, what would you like to say to yourself? |

## 4.2. Chatbot Implementation

Diarybot is a chatbot called “Plus Friend” on a messenger app called KakaoTalk,[1] the most popular messenger app in Korea. Users can simply add the chatbot in the same way that they add a friend on the app. The choice of KakaoTalk was for two reasons: First, it is easy to access, and services are available on PC as well as on mobile. Also, it offers Kakao Developers platform,[2] on which chatbots can be added, built and customized. One can register a Plus Friend instance and add customized skills for the chatbot to serve task-specific inquiries.



Diarybot

6:41 PM

DIARYBOT

Hi, I am Diarybot. Would you like to begin today’s session?

6:41 PM

**Figure 4.3. Diarybot’s welcome screen. Users can summon Diarybot whenever they want to, and Diarybot responds by asking the user whether he or she wants to proceed. The example is reproduced in English for language consistency in this thesis.**

Two Diarybot instances were registered and bot-specific skills were added for Basic and Responsive chats. At first, both proceed the same

[1] KakaoTalk is a mobile messenger app that was launched in 2010. As of 2020, it has 50 million monthly users, and the average number of message exchanges reaches 11 billion. Source: *Chosun Ilbo*. Retrieved on May 28, 2020: <<https://biz.chosun.com/site/data/html_dir/2020/05/01/2020050100799.html>>

[2] Kakao offers a developer’s platform for registering, building and customizing chatbots as Kakao Plus Friends. Source: *Kakao I Open Builder*. <<https://i.kakao.com/login>>

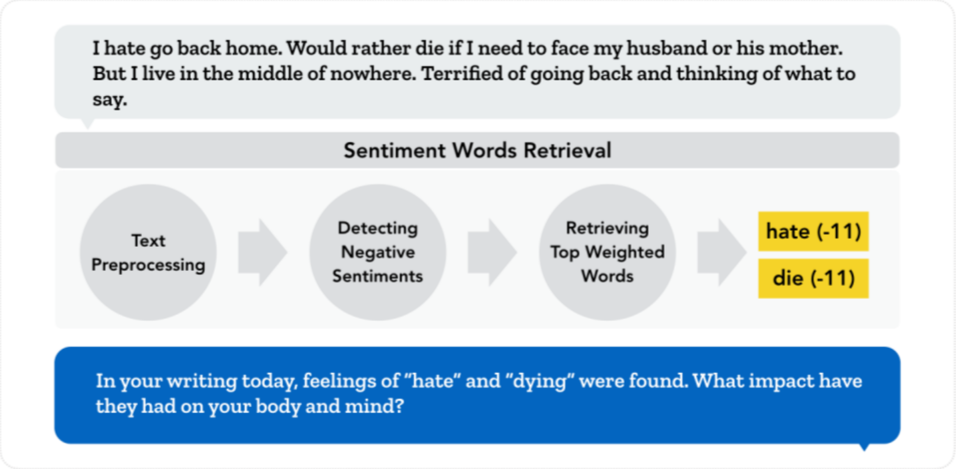
way. Diarybot greets the user and asks if he or she wishes to continue on writing. What the user writes is sent to Diarybot's skills for Basic or Responsive chats. Basic chat returns a simple message to thank the user for writing. Responsive chat asks 5 follow-up questions before thanking the user at the end. <Figure 4.3> is the initializing screen of the Diarybot conversation.

### 4.2.1. Emotional Intelligence

In Responsive chat, Diarybot finds two types of keywords: key negative sentiments and a key relationship. For this functionality, it uses aforementioned sentiment analysis skills to retrieve negative sentiments in the users' text via a trained deep learning algorithm for linguistic analysis serviced by ADAM Open AI API.[③] The API returns all negative sentiments in Korean morphemes, with weights automatically calculated by the algorithm. In this process, it returns modifiers in independent morphemes as well, thereby losing the direct dependencies in language. Still, the weights imply the intensity of the sentiments. Diarybot then ranks these weights in order to return the top most negative sentiments in writing, excluding morphemes that are not in complete adjective or verb form. These words are incorporated in the follow-up question by the bot, as indicated in <Table 4.1>. If no sentiment is expressed in the users' writing, the bot refers to a template question that simply asks the user to review their feelings at the moment upon writing. <Figure 4.4> illustrates Diarybot's sentiment analysis process in search for emotion keywords. If no emotion keywords are found, the bot asks the user to

[③] ADAM Open AI is RESTful-based public AI API service. It offers a total of 60 API services for linguistic as well as audio and image data analyses, and combine these APIs for virtual assistants, intelligent robots, etc. Its linguistic analysis scores an about 99% accuracy rate. Source: *ADAMS.ai Open API*. <<https://www.adams.ai/overview>>

review any feelings from writing about the traumatic event.



I hate go back home. Would rather die if I need to face my husband or his mother. But I live in the middle of nowhere. Terrified of going back and thinking of what to say.

Sentiment Words Retrieval
Text Preprocessing
Detecting Negative Sentiments
Retrieving Top Weighted Words
hate (-11)
die (-11)

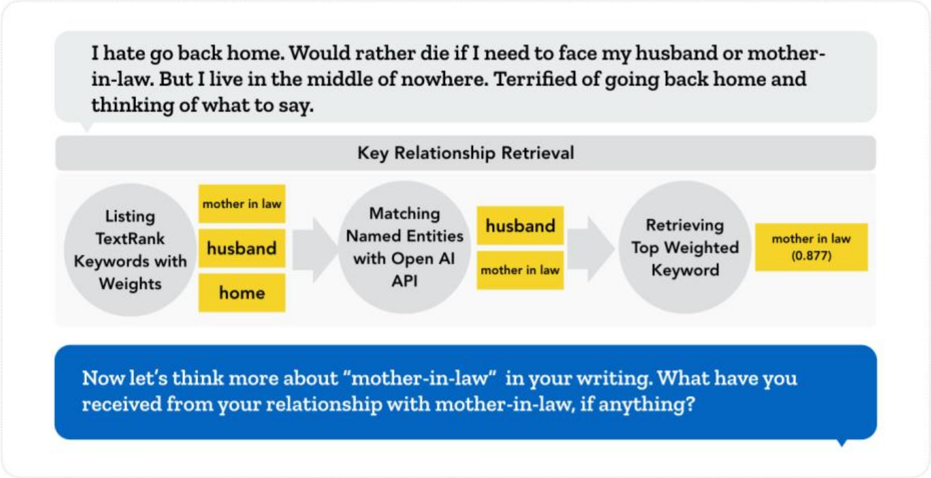
In your writing today, feelings of "hate" and "dying" were found. What impact have they had on your body and mind?

**Figure 4.4. A graphic illustration of Diarybot's sentiment retrieval. The user input at the top is a mock example, where a hypothetical user discusses most troubled incident. The bot analyzes the input and retrieves the most prevalent negative sentiments in yellow boxes on the right. The prompt incorporates these in the prompt at the bottom.**

The three responsive prompts invite a trainee to reflect upon a key relationship, that is "X." To retrieve X, Diarybot uses a two-track keyword extraction method, as illustrated in <Figure 4.5>. First, a TextRank algorithm [119,159] searches for keywords in the text given by the user. These are cross-examined against the list of familial, social and occupational relations that the named entity recognition API[4] finds in the writing. If a match is found, the X is retrieved. This is to precisely aim at the key relationship amongst many possible candidates in the text. If the TextRank keyword does not match, the relation word is retrieved

[4] The named entity recognition is serviced via public Open AI API in Korea, supported by the Ministry of Science, Technology, Information and Communication. Its AI-based linguistic analysis services include an automatic recognition of named entities that pertain to human relationship categories included in the exhaustive TTA Standard Named Entity Tagset (TTAK.KO-10.0852) that lists 15 categories and 146 sub-categories. Source: *Linguistic Analysis Skill at the Public Open AI·Data Service Portal*. <<http://aiopen.etri.re.kr/guide_wiseNLU.php>>

for X to fit into the context of the responsive prompt. Finally, in case of no match and no relation word, the TextRank keyword is ignored and Diarybot returns a template prompt.



I hate go back home. Would rather die if I need to face my husband or mother-in-law. But I live in the middle of nowhere. Terrified of going back home and thinking of what to say.

Now let's think more about "mother-in-law" in your writing. What have you received from your relationship with mother-in-law, if anything?

Key Relationship Retrieval
Listing TextRank Keywords with Weights
mother in law
husband
home
Matching Named Entities with Open AI API
husband
mother in law
Retrieving Top Weighted Keyword
mother in law (0.877)

**Figure 4.5. A graphic illustration of Diarybot's key relationship retrieval. Again, the example comes from a mock user. The keywords from TextRank algorithm are cross-examined against all named entities. The final keyword is included in the prompt at the bottom.**

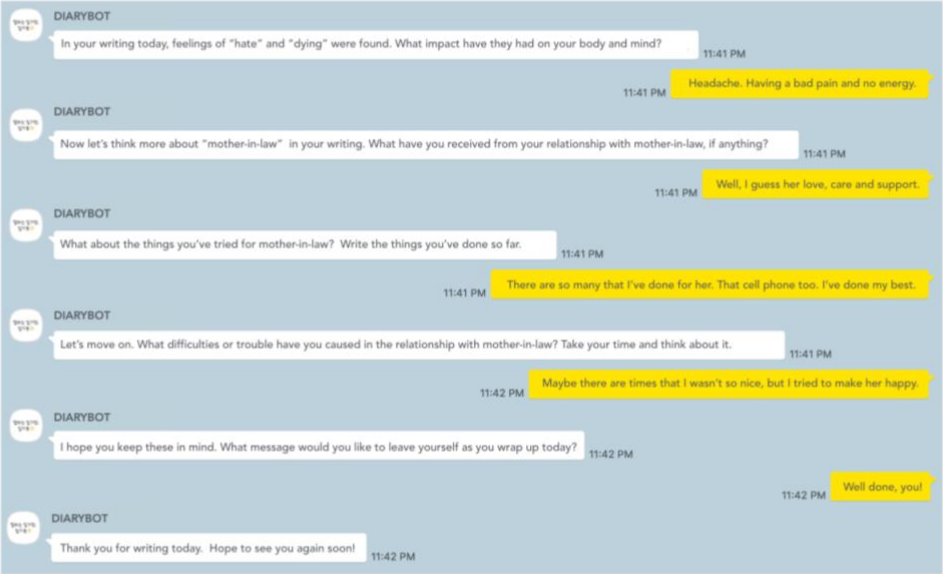
### 4.2.2. Procedural Intelligence

For Basic chat, the sequence of welcome-writing-exit can be managed by KakaoTalk's pre-defined functionality on the developer's platform. For Responsive chat, however, Diarybot is linked to skills from external modules, Flow Manager and Response Generator.

***Flow Manager:*** In order to make sure that the bot returns each prompt in order. To achieve this, the Diarybot system assigns a unique session for each user by their ID. It processes sequential information to make sure the user moves from one stage to another in the welcome-writing-follow up-exit sequence.

***Response Generator:*** Response Generator identifies keywords and assemble responses as illustrated above in <Figure 4.5>. An example of the follow-up conversation in Responsive chat is illustrated in <Figure

4.6>.



DIARYBOT

In your writing today, feelings of "hate" and "dying" were found. What impact have they had on your body and mind? 11:41 PM

11:41 PM Headache. Having a bad pain and no energy.

DIARYBOT

Now let's think more about "mother-in-law" in your writing. What have you received from your relationship with mother-in-law, if anything? 11:41 PM

11:41 PM Well, I guess her love, care and support.

DIARYBOT

What about the things you've tried for mother-in-law? Write the things you've done so far. 11:41 PM

11:41 PM There are so many that I've done for her. That cell phone too. I've done my best.

DIARYBOT

Let's move on. What difficulties or trouble have you caused in the relationship with mother-in-law? Take your time and think about it. 11:41 PM

11:42 PM Maybe there are times that I wasn't so nice, but I tried to make her happy.

DIARYBOT

I hope you keep these in mind. What message would you like to leave yourself as you wrap up today? 11:42 PM

11:42 PM Well done, you!

DIARYBOT

Thank you for writing today. Hope to see you again soon! 11:42 PM

**Figure 4.6. A mock follow-up conversation with Diarybot in Responsive chat. All conversational exchanges are between a hypothetical user and reproduced in English. The white text balloons are Diarybot's responsive prompts. For emotional awareness, it finds "hate" and "dying." For social awareness, it finds "mother-in-law." Finally, for self-awareness, it asks the user to leave a message for the self to wrap up.**

# 4.3. Experimental User Study

To find out how users experience Diarybot conversation, a user study was designed for a controlled experiment. To contrast the two chats with a baseline, a Google document was set up, which included the same expressive writing instructions as those by the bot.

## 4.3.1. Participants

A total of 30 participants (14 male, Median=28 years, min=23, max=41) were recruited from an online post within a university campus. Most participants were undergraduate and graduate students from a variety of disciplines. Prior to the experiment, one researcher explained the purpose of the study and sought voluntary consent to participate. The selection criteria were the willingness to participate in a four-day reflection on the most difficult life experience(s), and at least a week-long experience of using KakaoTalk. At completion, participants received a gift voucher equating to $20 in value as compensation for their time.

### 4.3.2. Task

The study ran from October 21, 2019 to November 15, 2019. The study lasted for four days to maintain the original expressive writing setup [200]. To capture feedback on their experiences, participants were asked to complete surveys every day before and after the writing, and before and after the study. They were also asked to take part in a post-hoc interview for a deeper probe into their reflection. The entirety of research activities took approximately 120 minutes in total.

### 4.3.3. Procedure

Participants were randomly assigned to each of the three writing conditions. They were invited to a quiet study room with a sizable desk and a comfortable chair. There was a laptop for writing, with snacks in a basket. When participants came on the first day they filled out a pre-survey. They were then left in private to write. Time taken for writing varied, but it took about 20 minutes a day. Upon their notice the researcher came back to save the conversation, and they filled out a post-survey. The same procedure was repeated for the next three days. Last day's procedure included a semi-structured interview.

### 4.3.4. Safeguarding of Study Participants and Ethics

## Approval

Three measures were taken to ensure safety of the participants. First, in the introductory session, participants were ensured they could leave the study if they found any portion of the procedure difficult, or did not want to continue in the writing or research. Furthermore, the researcher took precautions to read participants' writings and survey responses immediately after each session to check for any indicators of distress or risky behavior that could be alarming, such as disclosures of self-harm or intend to harm others [188]. If this was the case, the researcher would discuss with the participant what would be an appropriate action to take, such as to consult the university's health support, which never happened. Finally, on each day, the researcher reiterated the day's procedure at the beginning of the activity, and took questions to confirm whether the participant was experiencing any confusion or uneasy health symptoms from writing. The researcher's contact information was provided for any case of emergency, should they choose to leave the study at any point. The purpose, procedure and instruments in this study were carefully reviewed and approved by Seoul National University's institutional review board (IRB 1910/002-020).

### 4.3.5. Surveys and Interviews

This study took a mixed-methods approach to gain a detailed understanding of the users' reflection experience. This included a set of survey instruments to capture: (i) participants' health and psychological wellbeing; (ii) their evaluations on the conversational experience upon reflection; and (iii) their perceptions of the chatbot design. To gather deeper insight, a semi-structured interview was conducted at completion of the study. All survey data was analyzed using Python's statistical analysis packages, and interviews were anonymized and transcribed for a thematic analysis.

### (1) Health and Psychological Wellbeing

One of the key interests of the expressive writing procedure is the relationship between writing and health [204]. Since participants come from a nonclinical population, this study took two types of survey instruments to identify any signs of physical and mental discomfort, using Pennebaker's questionnaire [218]. This 16-item survey was taken on a 7-point Likert scale (1: not at all; 7: a great deal), right before and after the writing.

To measure participants' psychological wellbeing, two widely used instruments were used on a 5-point and 7-point Likert scale, respectively. The Schwartz Outcome Scale [94] (SOS-10; e.g., "I feel hopeful about my future.") and 7-point Clinical Outcome in Routine Evaluation [65] (CORE-10; e.g. "Over the last week, I have felt unhappy.") are both widely used to measure wellbeing in a relatively short span of time, and representative of all levels of patients [94] and common mental health problems [9]. The wellbeing surveys were collected at the beginning and the end of the study.

### (2) Reflection Experience

To understand how participants approached the writing activity and reflection, Pennebaker's writing questionnaire [218] was used on a 7-point Likert scale. The items included: "How much did you want to talk about what you wrote today?", and "How much did you hold yourself back from talking about what you wrote today?". To capture chatbot-guided reflection experience, 4 additional items were added: "While writing, I felt like I was heard.", "While writing, I felt like I was talking to someone.", "I could gain a new perspective on what I wrote about." and "I could have a better understanding of what I wrote about." This survey was collected every day after writing.

On the last day, post-survey included Pennebaker's last day of writing questionnaire [197], also on a 7-point Likert scale, to capture users' overall experience of reflecting on their trauma with Diarybot.

### (3) Perceptions on Chatbot Design

To characterize how participants felt about using Diarybot, survey items were taken from the social robot acceptance toolkit [96]. The survey items spanned 10 qualities of a social robot: Anxiety; Attention; Intention to Use; Perceived Adaptability, Enjoyment, Sociability and Usefulness; Social Influence; Social Presence; and Trust. These items were included in the post-study survey, measured on a 5-point Likert scale.

### (4) Interviews

The study concluded with a semi-structured interview that asked participants about their overall experience of the writing activity; what they thought of the chatbot design and interactions, as well as their understanding of the personal life events after reflecting on them in the study. Each interview lasted for about 20 minutes.

All interviews were audio recorded, fully transcribed, anonymized, and then subjected to thematic analysis by following the 6-phase process [29]. To this end, the interview transcripts were reviewed and segmented for a process of open coding, which generated free-phrased labels. To allow an enough time to familiarize with the data, the open coding process was repeated. The labels were reviewed and renamed for initial codes. The codes were examined several rounds before initial themes were formed. As advised by Braun and Clarke [29], themes were reviewed again to see if they fit into the overall study scheme and research questions: how participants evaluated their self-reflection experience with the chatbot; how they responded to the disclosure and guidance strategies; and what they found supportive and disruptive from the overall experience. The global themes generated included: telling narratives; chatbot interactions for support; and adapting user behavior.

## 4.4. Results

The findings show that participants experienced different types of self-reflection in Basic and Responsive chats, in terms of user expression, interaction and engagement. Both quantitative and qualitative findings are as follows.

### 4.4.1. Quantitative Findings

Despite no telltale differences in health and wellbeing, participants had an emotional experience reflecting on a past trauma. Those who used Diarybot for Basic and Responsive chats showed different perceptions of the bot, indicating that the interactions incurred different reflections.

Amongst the total of 120 writings, participant responses to the expressive writing instructions were, on average, 178.5 words long and included descriptions of: conflicts in social relationships (n=31, 25.8%), family crises (n=29, 24.2%), low self-esteem (n=17, 14.2%), failed love (n=12, 10%), failures (n=12, 10%), work stress (n=11, 9.2%) and other (n=8, 6.7%). Describing the reasons why those narratives were chosen, participants most often indicated that it was a difficult or traumatic experience (n=47, 39.2%) or a recent trouble (n=23, 19.2%). Less often they described motivations to wrap up the study and give closure (n=15, 12.5%), that it was the event that came to mind (n=11, 9.2%), to continue the narrative from the last day (n=10, 8.3%), wanting to write about something that was never told (n=8, 6.7%), and other non-specified reasons (n=6, 5%).

**(1) Users experience mood swings but no changes in wellbeing**

The overall experience had little impact on participants' wellbeing in terms of both SOS-10 and CORE-10 scores. Given such a short span of writing, it is natural and understandable that there is little change. Most expressive writing studies, the participants are called in for follow-up in about months' time [198,218]. Additionally, before and after the writing, there was not any noticeable change in participants' physical symptoms, but emotional states. In fact, a three-way mixed ANOVA analysis indicates there were significant symptoms and mood changes, with a mixed up-and-down trend each day: *cold hands* ($F_{1,27}$=7.976, p<0.01), *sad* ($F_{1,27}$=6.975, *p*<0.05), *guilty* ($F_{1,27}$=10.357, *p*<0.005), *happy* ($F_{1,27}$=8.795, *p*<0.01), and *fatigued* ($F_{1,27}$= 14.925, *p*<0.001). Also, since users could choose to write different topics each day, changes were observed in progress of time: *sweaty hands* ($F_{2.18,56.66}$=6.374, p<0.005), *nervous* ($F_{2.46,66.43}$=5.864, *p*<0.005), *sad* ($F_{2.40,64.83}$=4.233, *p*<0.05), and *contented* ($F_{3,81}$=4.608, *p*<0.01). The analysis included Mauchly's test for sphericity, and if needed, Greenhouse-Geiser corrections were applied. Finally, also observed were interaction effects on writing and time: *headache* ($F_{1.88,49}$=3.512, p<0.05), *nervous* ($F_{2.37,63.99}$=5.118, *p*<0.01), as well as writing and writing interface: *fatigued* ($F_{2,27}$=3.616, *p*<0.05). Still, the generalized eta-squared values for both interaction effects were less than

0.05, suggesting that writing about a trauma itself was the main factor in the symptom and mood changes. Later in the interviews, participants said the feelings lingered “a bit, about an hour” (P6), but “faded” in time (P3). This finding is also in line with literature that expressive writing can feel taxing at times given the subject nature [200,205].

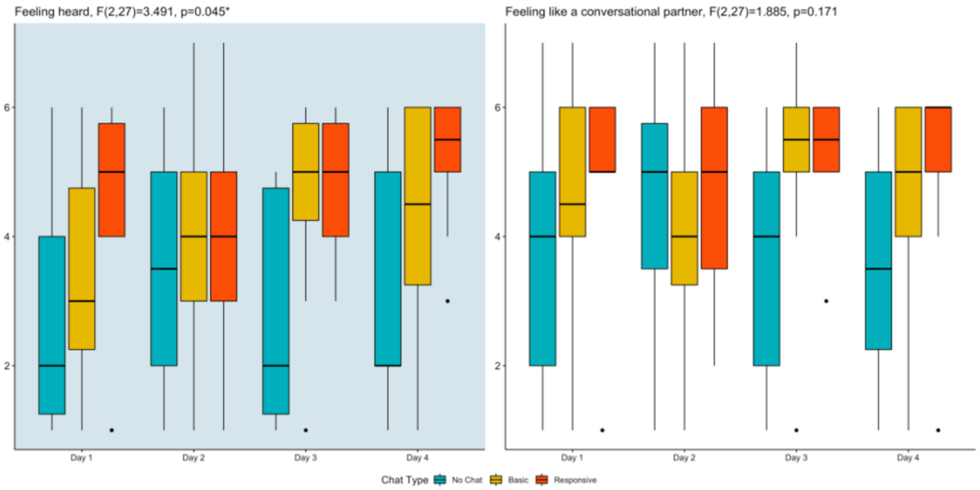


**Figure 4.7. Daily participant responses to how much it felt like being heard, as opposed to having a reciprocated conversation. As indicated on the left, participants in the Basic and Responsive chat conditions felt significantly more like being heard compared to the baseline condition.**

### (2) Diarybot enables social sharing of emotions

In both Diarybot chat conditions, participants felt like they were being heard by the bot. Responses to daily post-writing surveys in <Figure 4.7> show a significant difference between both chats and the baseline, as indicated by a two-way mixed ANOVA analysis ($F_{2,27}$=3.491, *p*=0.045). However, they did not feel like they were engaging in an active conversation with Diarybot ($F_{2,27}$=1.885, *p*>0.1). This shows that Diarybot was a type of reflection partner that would play a role of soundboard to the participants, while not so much to be like a friend who would actively engage in the conversation and chit-chat. What's interesting is that it was possible to incur such a feeling in the Basic chat interaction. This means that even the bot medium itself, with minimized interaction, can provide a different writing experience from writing on a Google document.

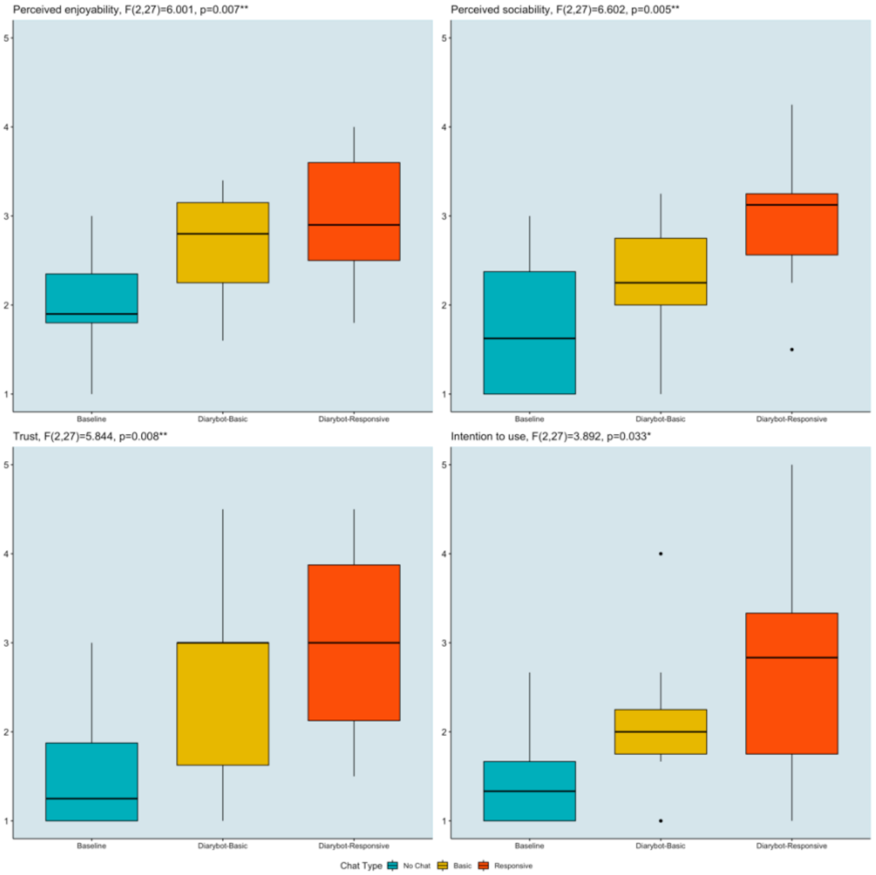


**Figure 4.8. Participant responses to social robot acceptance measures. Significant differences were observed for Responsive chat in four social acceptance items: perceived enjoyability, perceived sociability, trust and intention to use.**

### (3) Increased interaction leads to engagement

As seen in <Figure 4.8>, engaging in Responsive chat with Diarybot was not like having more active conversational exchanges compared to the Basic chat, but it led to a more sociable perception of Diarybot, as shown in participants' post-survey responses to chatbot perceptions. The two-way ANOVA results show that participants assessed the interaction with Diarybot in Responsive chat to be significantly more enjoyable ($F_{2,27}$=6.001, $p$=0.007), more sociable ($F_{2,27}$=6.602, $p$=0.005), trustworthy ($F_{2,27}$=5.844, $p$=0.008) and willing to use again ($F_{2,27}$=3.892, $p$=0.033). The post-hoc Tukey HSD results also suggest that compared to the baseline, the Responsive chat provided a much more enjoyable and sociable interaction that would lead to increased engagement <Table 4.2>. This shows that the participants did find the increased interaction in the Responsive chat amusing enough for continued engagement, though the interaction was not quite reciprocated.

Moreover, the increased interaction did not lead to a consistent finding of new perspectives or renewed understanding of the past trauma. The daily post-writing responses to these questionnaires show more variance within each group, suggesting that the follow-up guidance may or may not impact how they think about the trauma after writing or chatting with the bot. How participants think about the experience is further detailed in the qualitative findings.

**Table 4.2. Tukey's HSD test results.**

| | **Comparison 1** | | | | **Comparison 2** | | | | **Comparison 3** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Baseline – Basic Chat** | | | | **Baseline – Responsive Chat** | | | | **Basic Chat – Responsive Chat** | | | |
| | *estimate* | *conf.low* | *conf.high* | *p.adj* | *estimate* | *conf.low* | *conf.high* | *p.adj* | *estimate* | *conf.low* | *conf.high* | *p.adj* |
| ***Post-writing*** | | | | | | | | | | | | |
| Feeling heard | 1.05 | 0.1624 | 1.9375 | 0.0160* | 1.55 | 0.6624 | 2.4375 | 0.0002*** | 0.50 | -0.3875 | 1.3875 | 0.3770 |
| ***Post-experiment*** | | | | | | | | | | | | |
| Emotional expressiveness | 0.9 | 0.0653 | 1.7346 | 0.0326* | 0.1 | -0.7346 | 0.9346 | 0.9530 | -0.8 | -1.6346 | 0.0346 | 0.0622 |
| Difficulty in writing | -1.7 | -3.2461 | -0.1538 | 0.0290* | -0.30 | -1.8461 | 1.2461 | 0.8810 | 1.4 | -0.1461 | 2.9461 | 0.0816 |
| ***Acceptance*** | | | | | | | | | | | | |
| Perceived enjoyability | 0.7 | -0.0345 | 1.4345 | 0.0640 | 1.0 | 0.2654 | 1.7345 | 0.0062** | 0.3 | -0.4345 | 1.0345 | 0.5750 |
| Perceived sociability | 0.575 | -0.2440 | 1.3940 | 0.2090 | 1.200 | 0.3809 | 2.0190 | 0.0032** | 0.625 | -0.1940 | 1.4440 | 0.1600 |
| Trust | 1.00 | -0.0765 | 2.0765 | 0.0725 | 1.45 | 0.3734 | 2.5265 | 0.0067** | 0.45 | -0.6265 | 1.5265 | 0.5610 |
| Intention to use | 0.5666 | -0.5003 | 1.6336 | 0.3980 | 1.200 | 0.1330 | 2.2669 | 0.0251* | 0.633 | -0.4336 | 1.7003 | 0.3200 |

**(4) Guidance influences user behavior**

Responses to post-experiment survey reveal that reflecting on traumatic experiences within chatbot conversations was significantly less difficult, and easier to express feelings <Figure 4.9>. A two-way ANOVA on post-study survey shows that participants who took part in Basic chat felt significantly less difficult to write ($F_{2,27}$=4.234, *p*=0.025), and easier in expressing their feelings in writing about their trauma ($F_{2,27}$=4.294, *p*=0.024). Both baseline and Responsive chat participants scored along the middle. The post-hoc Tukey HSD analysis, as in <Table 3.2>, shows that the difference stems from the baseline and Basic chat, both for emotional expressiveness (*adjusted p*=0.0236) and difficulty in writing (*adjusted p*=0.0290). This result suggests that participants find it easier to engage with the chatbot to form a narrative about their trauma when the chatbot had a minimal interaction with them, only welcoming them, asking them to write, and thanking them for writing. On the other hand, when only the expressive writing instructions were given on a Google document sheet, participants felt it much more difficult to go about the writing, and experienced a certain level of inhibition in their emotions. Likewise, with increased interactions in Responsive chat the ease and expressiveness were moderated, with Diarybot asking further follow-up questions about their writing.

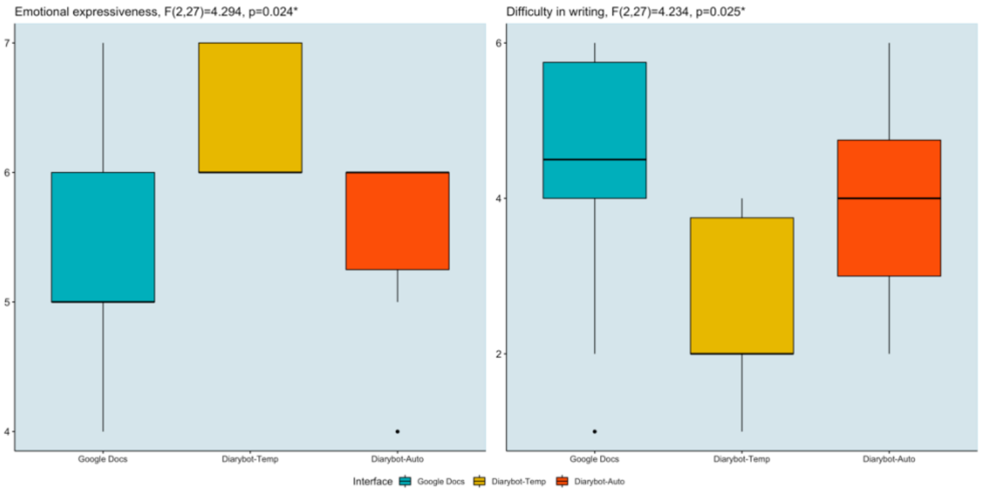


**Figure 4.9. Last day participant responses to the overall ease of emotional expressiveness and difficulty in writing. For both, Basic chat participants responded with a significant difference for more emotional expressiveness and less difficulty in writing.**

## 4.4.2. Qualitative Findings

Upon completion of the four-day writing, all participants were asked to participate in a 20-minute semi-structured interview, which was to gain a deeper understanding of the reflection participants had with and without Diarybot. The interviews reveal interesting findings for designing chatbots and guidance strategies for chatbot-mediated reflection.

### (1) Resolvedness leads to different reflection outcomes

Most importantly, what participants said in the interviews suggest that above any interaction design, writing about traumatic events yields different individual experiences and responses depending on how resolved the events feel to the participants. This was most visible in the baseline participants, who had to be responsible for the four-day writing the whole time without additional technological interaction or guidance.

For P7, the writing experience was essentially a reflection process to "reorganize some forgotten thoughts, which wouldn't necessarily be new discoveries." P1 said the only different the writing would have made was "mood swings." However, for P7 and P18 the writing helped, in that it prompted "a typical reflection session" (P18).

In other words, the reception of expressive writing depended much on how the individual was trying to make out of it. For example, inhibition of feelings was a strong motivation for P5, who sought much not to be swayed by recurrence of any negative feelings from the trauma:

> *"I did not want to be emotional. In addition, it was about the most negative experience in life. You know, people do not linger on such things every day. We put it inside. So in a way the writing was an opportunity to think about it again, but not more. Like, it was already on my mind and I only wrote it out? It didn't help with news ideas or perspectives."*

Expressive writing itself most likely incurred an emotional burden for the participants. P7 and P15 said that it was "depressing" since they had to revisit the past traumas, which P18 worried about:

> *"It helped me. I mean, I feel like this can help if the person is in a psychologically healthy state. People with higher depressive symptoms, like, anti-social people, I mean, more hysterical people may feel even more down from this experience. It's just a thought, though."*

However, in the end the writing could have gains for participants. P1 said, "You know, at last, you look at things from a new perspective, and learn new things." P9 said she could turn this experience a positive one in the end:

> *"It was the last day today, right? So I had made up my mind to*

> *wrap everything up. You know, that is what happened but from that experience I learned this. I was actually thinking about it that way but writing it out helped that process. You know, I was more positive? To justify those things. I wanted to finish this up with lessons learned."*

What baseline participants said in the interviews testify to the fact that in approaching negative life experiences, individuals have different motivations as well as different levels of perceived unresolvedness of those events. Having to revisit them in any way has an emotional cost, and for the benefits of expressive writing to outweigh it, one needs to have a strong resolve or motivation to lead it towards their advantage. In this study, this was provided in the form of a chatbot and its follow-up questions.

**(2) Chatbots can be designed for different types of reflection**

Participants who took part in Basic and Responsive chats with Diarybot went through similar experiences with respect to the expressive writing procedure, but their responses show that they had different experiences overall. While Basic chat participants liked that they could say things that they wouldn't otherwise say to someone else, Responsive chat participants liked that Diarybot would give "nudges" that would distance themselves from the trauma and reexamine what had happened.

For P2, a Basic chat participant, Diarybot would feel "strikingly different" from journaling or writing a diary. The very presence of Diarybot on a messenger app creates "the feeling that someone's there." This feeling of virtual interlocutor recurs in most Basic chat participants: P17 says "yeah, like, you are sitting down there to pour out something to somebody else," and P4 said "there was definitely the feeling that

Diarybot was listening to me." In P8's words:

> *"Right, it's a chatbot. It feels different. There's the, like, expectation that it would respond back to you. I don't have those expectations for memos and word documents. They aren't supposed to talk to you. But this chatbot is a conversational partner. That makes it different."*

In this way, having Diarybot made reflecting on and writing about the trauma a storytelling process. P8 said he wasn't lonely because as he wrote about what happened, he "wasn't alone." P10, also a Basic chat participant, felt it was different from writing in her diary because "it felt like someone's watching, like, telling stories." This made P10 "talk casually as well." According to P15, writing feels like a composition, where "you have to put everything in an organized way," while chatbot feels like a conversation, where "you just talk, talk naturally, and there's more emotional side to it rather than having it all nicely typed up for an essay." P2, P8, and P20 also added that it was "a good experience" to have their stories told to a nonemotional other, "who wouldn't judge [them]." P26 especially preferred to write this way, for she would like someone to listen to her stories, and in a more comfortable manner than when she'd keep diaries.

For Responsive chat participants, the experience had a different nuance added to it. they mostly felt like the follow-up process was an "efficient means" (P12) of detaching from the emotional aspect of the trauma but "processing it through" (P29). More specifically, they liked the procedure and directiveness within the flow of the prompts provided. P16 said "Er… I'd say the conversation proceeded as if it were a professional one… the questions were rather repetitive, but it asked me how I felt, and what I think… I'd say it wasn't bad." P12 felt like the process was like a psychotherapy:

> *"I've had some (therapy sessions) myself, and I've heard others like it, and there's not much to it; you know, you'd say what's troubling you, and you say 'I don't know what to do' sort of things… and the therapist would say 'oh, is that right?' That's all. I mean, they'd never give you answers. I thought Diarybot would just be the same."*

P11 agreed, "I had come to a psychotherapist once. The process in which she'd approach my problems was like what Diarybot did. And when I told her what my problems were, she'd almost repeat it, 'I hear you saying these things, is that right?' or 'Is this how you feel?' The way Diarybot asked me questions was like how she'd branching into my problems."

Most of all, the follow-up questions that Diarybot asked offered a chance to rethink the situation in a new light, "maybe breaking out of the box" (P30). For 12, the questions provoked her to think about something she hadn't thought about: "especially the last question, it was something that I'd never say or ask myself to do. It had a big impact on me, in a positive way." P11 had a similar experience:

> *"It's like Diarybot helps me distance myself from [the event]. It would ask me someone I wrote about, and it suddenly puts him, or her, in a new light, like from a different perspective… that would help me take a step back and think about what to do, like to console myself, think again, reflect further, you know."*

In P25's words: "Diarybot would put things in a different perspective, distancing a 'me'-point of view of things." At first it would feel "unexpected" to P28, which would eventually lead him to think:

> *"What have I received from my friend? It had me pause for a moment. Then I thought, well, a glass of beer, those things, but I*

> *wouldn't really answer that, and I kept thinking, what are his influences, on me, I remembered those. Something I haven't thought of."*

Put together, P29 and P30 would say "taking the third-party view" of things "was really helpful," which "put a brake on the emotional outpouring but start thinking about 'what now?', and a more objective picture."

In sum, the Basic and Responsive chats offered different reflections on participants' writing about traumatic experiences. Having a virtual company released the burden of writing up a composition of a trauma narrative in Basic chat and turned the experience into almost a personal but social gathering for telling stories, explaining what happened and how it felt. On the other hand, having Diarybot following up with the stories offered a new perspective in looking at things, a pleasant detachment to invite new self-insight.

**(3) Interaction helps engagement, but needs to be varied**

Engaging with Diarybot, however, was not altogether a satisfying experience. All Basic chat participants would leave with some regret, wishing Diarybot "had had more interaction" in the conversation (P17). P14 suggested "it'd be much better if Diarybot would give something like, you know, fillers or encouragements? Just to signal that it's following me." P23 said, "you know, it's a bot. When you say it's a bot, there's usually more." P10 wanted the bot to feel "more personal, you know, since there was no response in return." In P23's words:

> *"If there were something more, anything to let the conversation going, I'd feel more interested, more fun I guess, and I'd think 'I want to do it'."*

Increased interaction in Responsive chat was "definitely a plus," according to P16. What participants especially liked about Responsive chat was that it referred back to what they wrote about, and that it gave "adaptive responses" over time (P30). P29 said:

> *"What I liked about Diarybot was that it'd say, 'there were these feelings' from what I wrote, and it would ask me questions based on that. I thought it was based on what I wrote."*

P28 had a similar view:

> *"What I liked best was that the bot responses changed. What was it? I think I wrote different stories each day for some reason. All four were different, but on the first two days the bot would give me similar responses. But on the second and third days there were some people involved in my writing and the bot caught those and it would ask me like, 'let's talk more about that person,' which was really interesting. Then I thought, 'oh? the bot can say more different things than I expected,' I liked that. That, that it could change. I think today it was also different…"*

However, Responsive chat participants pointed out that sometimes the bot only responded with template questions and they did not fit exactly into the context of their story. P13 rather complained:

> *"Coming to write every day and telling Diarybot what troubles me can be a bit demanding… you know, I came to write about what's stressing me out the most right now. But the bot only asked me what can be done, what needed to be done, something like that throughout. To be honest, I couldn't see the point of the questions because if I had known the answers myself I wouldn't need to talk to Diarybot in the first place."*

P12, who only experienced template questions for the entire session, also added a similar view:

> *"I think the best question for me was the last one, when it asked me to leave me a message. I took it as a self-encouragement, and it felt really nice. The questions that came before were a bit confusing to me. I was not exactly sure what the bot was referring to when it said what should be done."*

### (4) Users readily adapt to chatbot guidance

Responsive chat participants, depending on how and what they wrote each day, could have different prompts. <Table 4.3> shows the number of days that the participants had responsive prompts.

**Table 4.3. Number of days that Responsive prompts were retrieved for Responsive chat participants.**

| Days | Number of Participants |
|---|---|
| 0 | 2 |
| 1 | 4 |
| 2 | 2 |
| 3 | 1 |
| 4 | 1 |

Most Responsive chat participants had one or two days of responsive prompts in the experiment, which means that they had a chance to engage in both template and responsive prompts in the follow up process. P29 said that seeing Diarybot specifically responding to what she wrote changed her perception of the chatbot to be "more reliable." This is well detailed in P31's words:

> *"On the first two days, the questions felt rather like templates. I'd*

> *write about the experience, and the bot asked me how it affected my body and mind, and what I could do and couldn't. But on the third day, I was surprised, there was the keyword right off from what I wrote, and the bot would exactly point out on the relationship. The questions also changed. It was a nice surprise and I think I could write in more detail thanks to that. On the fourth day it went back to the former questions but had a different nuance. But different phrases can also affect the writing, right? I thought the system could change. I guess, I guess what I wrote was more and less the same, but what I thought about the system changed, which I think was an important difference."*

The discovery of Diarybot's behavior led to adjusting user behavior as well, as in choosing topics for writing and how participants wrote things. For example, as experiment proceeded, P25 would choose topics that would suit the responsive prompts:

> *"It was on the second day. It just happened that there was some person in my writing, but it wasn't about the person. But the bot would hold onto that person and kept asking questions about him. So I learned that the bot focuses on people, so the next two days I talked about people issues. I thought 'let's not talk about abstract things."*

As for P30, she tried to keep the person in the writing consistent for Diarybot to understand, perhaps hoping it to *catch* the word:

> *"Especially today, I happened to write about people issues throughout the days, but I didn't deliberately do that. But when you talk about certain people, there can be different ways to call them, it could be 'him,' or 'that friend,' you know. I tried to keep it consistent, since I didn't want the bot to be confused. For*

> *example I could talk about my boyfriend. I could say 'bf' or 'he,' but I repeated 'boyfriend' throughout… it's longer but I tried to fit into the bot's pattern."*

Yet frustrations also stemmed from the fact that Diarybot could only respond with a fixed pattern and only pick up on relationship issues. Once participants learned that Diarybot would catch only person-related keywords, their interest waned a bit, or they tried to adjust their writing behavior. For 22, the conversation lost its realness: "The feeling of having a real conversation lessened, over time, because the questions would repeat themselves in an order. So I thought it was just a program in the end." P11 would also agree: "I wrote about different friends on the last two days, but Diarybot would ask me questions about just 'a friend.' I'd wish it would ask me different questions about it."

P25, who most actively engaged with Diarybot, described how her perceptions of Diarybot changed throughout the experiment:

> *"On the first day, I thought, 'oh, it feels like talking with a human being.' However, over time I could see the pattern, and since then I just felt that this was only a bot. You know, for four days, I learned what the bot would ask, and it did just that. So I thought, it did not ask me these questions because I said these things, but only it was supposed to ask these questions. Like, even if I said 'banana banana banana,' the bot would still ask the same questions? Of course I didn't do that, but I guess the surprise waned for me. I guess I got accustomed to it."*

## 4.5. Implications

The findings of this study indicate that Diarybot can be a potential reflection partner for an individual, to engage in an explaining or an exploring reflection, inspiring self-insight and awareness. Little change was observed in wellbeing, which was in fact expected. The study was rather brief, the participants did not have health issues. A meta-analysis of expressive writing studies has indicated that expressive writing has modest benefits within nonclinical population [8]. Overall, chatbots were received as a nonjudgmental listener to the participants, which makes the whole experience as *telling* stories of a trauma instead of writing. Guided prompts in Responsive chat offered an opportunity to reconsider the event in a new light, asking to review their actions as well as others'. Finally, once Responsive chat participants realized the bot led a different set of prompts depending on what they wrote, they made alterations in their narrative and chose to adapt to their assumed workings of Diarybot. Based on the findings, this section delves into the depth and patterns of chatbot interaction, which, in turn, calls for a careful reconsideration on design transparency in HCI.

### 4.5.1. Telling Stories to a Chatbot

The case with Diarybot calls attention for the depth and complexity of interactions with conversational agents. In this study, a mere existence of a chatbot on a messenger app, asking participants to write about a traumatic experience was enough to create a very different user experience from writing on a Google document. In a way, this finding supports the earlier work in that emotional disclosure in Pennebaker's expressive writing is in fact not conducted alone but with an implicit audience that is the researchers themselves [214]. Social disclosure matters, as it provides the motivation to write the narrative for a reader.

Prior research has pointed out that given the lack of specific instructions in Pennebaker's expressive writing, individuals are left free to choose a self-selected writing style [90]. Despite the potential myriad of individual writing styles, writing itself has been an effective means of revisiting traumatic experiences and making reinterpretations [201,205]. Nonetheless, the inconsistent findings in nonclinical populations [90] and improved health outcomes in writing about an imaginary trauma [83] still pose a question in understanding how exactly the writing leads an individual to a healing process.

This work sides with the social disclosure view, but with an explicit, or virtual, audience rather than implicit audience. In both baseline and Basic chat, participants were aware that the researchers collected their data. That each led significantly different reflection experiences indicates that it was more about how the medium led participants to approach the writing. Participants in the baseline condition said it was difficult to "fill up the blank space" (P1) and went "back and forth" to edit and revise the writing (P16). On the other hand, participants in the Basic chat condition could express themselves emotionally and felt less difficult talking about their experiences. This supports the idea that the telling of stories, instead of writing a narrative [124,195], might play a role in the potential wellbeing of the expressive writer. That the linguistic markers such as the type of works and pronoun use in the narrative correlate with health outcomes, rather than the quality of narrative [216], may support the idea that other factors in disclosure should be considered other than narrative composition.

The chatbot interactions immediately turns a composition into an interlocution. Engaging in the chat with Diarybot made the participants feel they "needed to explain [this] to the bot" (P29). Without much further interaction, having the audience free of social stigma and responsibility "relieved the burden" of disclosure (P8). This makes an important implication to HCI researchers that what participates in an interaction, as much as how it participates, also matters in designing technology.

### 4.5.2. Designing for Disclosure

An increased interaction in Responsive chat received mixed user interpretations. While the survey finding shows that it was not as easy to write or express feelings in the Responsive chat as in the Basic chat, further interactions with the bot were associated with higher perceived sociability and enjoyability, as well as trustworthiness and intention to use the bot in the future. In the interviews, participants especially liked that Diarybot generated responses from their writing, and pointed out on a key relationship. Over the course of the four-day experiment, however, participants gradually learned the routines of the interaction. Knowing what Diarybot will say next and the keyword that it will pick up from the writing received mixed user reactions: Participants either altered their narrative to see if they could surprise themselves, or their expectations for the bot waned as the experiment progressed.

This finding carries much importance to HCI researchers. First, it suggests that there is a strong preference for an increased interaction with a chatbot. Interaction incentivizes engagement, and guided prompts do not hamper with users' wanting to lead the interaction [186] as they still exert their will to lead the interaction by trying to vary the prompt keyword. Nonetheless, there is an even stronger preference for a *varied* interaction. Participants' excitement in talking with Diarybot waned as they felt like the responses were templated. In fact, this phenomenon regarding users wanting to have a control over the agent and yet having overblown expectations had already been foretold by Norman's notion of human-agent interaction [184].

What does this mean for designing chatbot guidance for disclosure? Decades ago, Horvitz [108] suggested the idea of mixed-initiative user interface with both reasoning machinery and direct manipulation. The principles of mixed-initiative user interfaces enable efficient human-AI collaboration, yet they also pose a systematic problem of systems having to guess user needs. More recently, Amershi et al [3] have proposed a set of design guidelines for human-AI interaction that are applicable to the stage of interaction: initial, during and over-time. The exhaustive set of principles listed in [3] is also centered towards having the agent remain in the role of a supporter, not a leader. A similar line of research on human-AI collaboration also revealed that users would like to take the lead in the interaction [186].

Users may not always know his or her wants or needs, especially when they reflect on a negative life experience and do not know what to make out of it. This work embraces the previously mentioned principles of human-agent interaction and makes a further attempt to suggest that agent guidance needs to be designed with an element of *planned* surprise. In fact, conversational agents are already designed this way; the inner workings of the chat algorithms are hidden from the user, and the only direct manipulation the user is allowed to make is to order a task. When chatbots offer guided questions to prompt the user for further thinking, the user in fact does not lose control. The conversation is centered on the user's narrative and the chatbot only delivers prompts for further disclosure. In other words, the bot only *nudges*. The notion of planned surprise means unpredictable guidance that responds to changing user context, so that the user does not lose control but can stay engaged in the interaction.

### 4.5.3. Rethinking Predictability and Transparency

The discussion so far challenges the two principles of usability in intelligent interfaces: predictability and transparency. Predictability refers to the extent to which a user can predict the effects of her actions [117]. Transparency is the extent to which she can understand system actions and/or has a clear picture of how the system works. In fact, it has been pointed out that systems that adapt to their users and change their behavior to better fit user needs may violate the principle of predictability and possibly also not be transparent and may hinder users' control over the system [107]. However, the findings of this work show that, to a certain extent, unpredictability in the system may work toward the benefit of an increased user engagement. Once participants found that Diarybot changed questions, their assumptions for the adaptability of the system changed as well, as in P28's words. Furthermore, noticing that Diarybot took a key relationship from their writing, P30 kept the words consistent within the writing hoping for the bot to pick it up. This shows that unpredictability triggers user needs to take back control of the system, yet once achieved, the interest may fade. The same applies for transparent systems. Not knowing exactly how the bot worked resulted in participant explorations around their narrative as to see how the bot would respond to them.

While managing user expectations is important in order not to mislead or frustrate users during their interaction with the bot [3], it is also important to maintain some enigma in its workings to support continued user engagement, especially when users seek to be inspired. In fact, in social sharing of emotions, people are willing to share with others who can offer new perspectives or interpretations of an emotional event [266]. An emotional event necessitates a cognitive articulation, for which people actively engage with others to find socially acceptable ways to define the experience [220]. Thus the design tensions in predictability and transparency in fact can open up new opportunities to design human-agent interactions that are changeable and even fluid; language connects unfathomable ideas and makes interpretations. That participants filled the void of the algorithmic enigma behind Diarybot's response shows that nudging is an important business. In other words, having participants try to guess, understand and interpret the bot's algorithmic intentions can create planned surprise. In turn, users may surprise themselves by participating in the interaction for their newfound articulations of an emotional event.

## 4.6. Summary

In this chapter, the design and implementation of *Diarybot*, a chatbot that encourages explaining and exploring reflections were discussed. Based on Pennebaker's expressive writing, Diarybot was offered in two versions: Basic and Responsive. The Basic chat only offered the user to write about a traumatic experience on a chatbot interface. On the other hand, Responsive chat delivered a set of follow-up questions derived from what a user has written in response to the initial expressive writing instructions. The findings show that the two chats could successfully mediate explaining and exploring reflections; while the stories could be shared in the Basic chat, different point-of-views could be considered in the Responsive chat for further thinking about the shared problem.

More importantly, the findings reveal that the different levels of interactions between Basic chat and Responsive chat can yield different reflection experiences on participants' past trauma. It was easier to share life's trauma and confess associated thoughts and feelings with it, when the writing was merely transformed into a conversation in Basic chat. The increased interaction in Responsive chat, however, led to higher user engagement. In addition, it was observed Responsive chat participants tested assumptions on the bot algorithm, trying to make sense of their experience. This leads to a potential tension in design of chat interactions for reflection assistants. Since users are veiled from the workings of a chatbot, expectations surge at first; yet engagement may wane as users learn the routines of the interaction. Varied and layered interactions may help; however, it may risk user controllability. The tensions highlight the heightened need to manage user expectations, and to design the "chats," as opposed to the "bot," to maintain continued user engagement.

The next chapter will discuss the overall findings of Bonobot and Diarybot studies, and further engage in challenges and opportunities in designing for chatbot-guided reflection. Finally, it will explore meaning-making as a novel interaction design metaphor for intelligent agents, taking self-reflection as a joint venture between agents and users.

# Chapter 5. Designing Chatbots for Self-Reflection: Supporting Guided Disclosure

In this study, two chatbots, Bonobot and Diarybot, were designed and implemented to encourage user narratives in support of the following: transformative, explaining and exploring reflection on life's most difficult experiences and unresolved stress. The results of the studies show the potential of designing for chatbot-guided reflection, and design strategies that can help users share and scaffold their articulation of negative life experiences. As for Bonobot, the evoking stage induced by motivational interviewing (MI) skills was well-received by participants as its questions served as an effective means of inspiring motivational boost for behavior change. Additionally, other MI-adherent statements that reflected on the particular subject matter, i.e. graduate school life, were also positively received. This finding indicates that carefully designed conversational sequences can serve as appropriate guidance to support a transformative reflection.

Yet designing Bonobot as if a human MI counsellor resulted in participants' demanding support that exceeds both human and machine capabilities. Bonobot actively led the conversation, asking open questions and providing words of empathy and encouragement. Participants were only encouraged to think about and answer the questions, and respond back to the bot. Such a design strategy could encourage a clearly purpose-driven reflection where users are directed to think about a certain idea without further ado or digression. However, as a result, user expectations may surge and there is a limited space for users to explore their own ideas. In addition, it was observed from a few participants that even though such a change-oriented, transformative reflection may be necessary, it is not always wanted by the users. This finding supports the rationale for constructing the design space in this work; a reflection process may not always be linear but even be regressive, and users need not take a single path in reflecting on a life's significant event but sometimes circumvent or detour. In sum, reflecting on stressors at graduate school with Bonobot could mostly encourage a change-oriented narrative for a motivational boost, yet challenges remained for self-determination and engagement issues.

To encourage users' own thought processes in reflection, the study proceeded on to designing more user-oriented reflections so that they are less bounded by guidance but allow users to make sense of the experience on their own. Employing expressive writing in psychology, Diarybot was offered in two chat versions, Basic and Responsive. Each supported a chat to support expressive writing narratives and a follow-up was added in Responsive chat. An experimental user study was set up to assign 10 randomly selected participants to each chat, with an expressive writing baseline in Google document. Though the writing prompt was essentially the same, chatbot participants felt significantly more heard than the baseline. Having a virtual yet explicit audience made a *reader* for the writing, which may support the earlier work on social disclosure [214]. This shows a mere presence of a bot instance or medium may as well create user expectations for interaction; without further understanding of the specific bot functionalities, users were ready to tell their story instead of composing it.

Yet more chat interactions in Responsive chat received significantly higher ratings in user perceptions of enjoyability, sociability, trust and intention to use. Inevitably, interaction incentivizes engagement; users found Responsive chat far more fun and enjoyable. It is still interesting that chatting with Diarybot in Responsive chat made reflecting on life's most negative experiences somewhat *enjoyable*. Some participants even approached the researcher and asked if they could continue using the bot after the experiment, while no such request was received from baseline. This finding speaks for the potential of employing natural language interfaces as an instrument for mediating companionship. Not only do we love to tell stories to others but we are also wired to do so [81,255]. Participants projected a role of a listener to Diarybot, and the interaction became that of a storytelling rather than a written composition. Potential lies in conversational agents that has a powerful comparative advantage to mimic the most natural human communication that is to talk.

As for the expressive writing activity, however, it was Basic chat that was much easier and less difficult to write than the other two. It is interesting that while the Responsive chat interactions were enjoyable, they were not the easiest or most comfortable in expression. Some Responsive chat participants said that the follow-up questions were sometimes out of context and felt like they did not fit into what they had expected. Still, surprisingly, participants gradually learned the routines of interaction and adapted their narratives to them. It was as if they were helping the bot understand – or they wanted to have the interaction unfold as they thought they'd like it to. Had the participants figured out how the algorithm worked behind the scenes, they might not have been able to construct their narratives as freely as they could have done. This points to an intriguing question whether AI needs to be *explained*. In other words, transparency might hamper with the user trying to make up his or her own narrative, or he or she will foresee what the agent will say next. Virtually no communication exists in such manner. Therefore, tensions exist in designing intelligent agents that engage in human-like communication, especially for reflection activities.

Despite a number of conversational agents both in research and in industry, there has been limited discussion on their design in the broader context of human-AI interaction. Most agents are task-based, running errands and achieving repetitive tasks. Yet more and more agents are starting to operate in a social context. Microsoft's XioaIce is starting to build a friendship with a user [244], and we have so-called chatbot therapists [68,73,115]. Still, these chatbots are mostly discussed in terms of their performance: how naturally they talk like humans, how effectual their treatment programs are, etc. This work has proposed the design and implementation of reflection assistant chatbots, Bonobot and Diarybot, which pioneer a *symbolic* interaction between user narrative and the bot algorithm. Users communicate their understanding of reality to the bot, and the bot's algorithmically retrieved responses are not perceived as mere characters but symbols with meaning and purpose. This indicates that in such a cognitive activity, the agent or AI is no longer an ensemble of numbers and computational algorithms; its agency is created in the hyperspace with its functions. In the advent of AI technologies, grave responsibilities are upon the engineers and HCI researchers alike to design "responsible AI" beyond explainable AI (XAI) [145] for lay users.

This chapter now discusses what chatbots as reflection companions mean in a broader context and how it may extend the existing research in HCI (e.g. [1,107,108,117,146,291]). More specifically, it discusses that chatbots as virtual and social partner has a unique place in encouraging a goal-oriented inner conversation, namely the cognitive processing of stressors while also assisting their emotional processing. Yet to support such, a careful consideration on the design of the chat interactions should be warranted. Tensions arise in the interactions between a human user and only an anthropomorphic agent, especially in terms of autonomy and adaptivity, as well as affordances of AI. To ensure user engagement yet embrace tensions in design, meaning-making is proposed as a novel design metaphor for AI in mental wellbeing. This metaphor supports the reflective thinking process that is process-oriented and often nonlinear, highlights AI in companionship of users instead of servanthood, and advances the current understanding of human-AI interaction from a mechanical one to a symbolic one, establishing AI as an relational agent [18,19] that walks with us on our life path.

## 5.1. Designing for Guided Disclosure

To enable reflective thinking on life's most troubling experiences in the best constructive manner requires user self-disclosure scaffolded by carefully designed guidance. This section delves further into how this could be made possible with mere text-based conversations with a chatbot. Based on the previous findings, this work presents chatbots as a private conversational partner that can successfully nudge cognitive processing of once troubling experiences, while encouraging user engagement via guided disclosure.

### 5.1.1. Chatbots as Virtual Confidante

One of the most important design decisions for chatbots in this work was to help users share as much as possible. The key design strategy was to translate a reflection activity – looking at inner thoughts and feelings

– into a disclosure activity – telling inner thoughts and feelings. This design decision was grounded on earlier research that putting stress into words does not only encourage users to reveal their thoughts and feelings about them, but also lead to a cathartic, or even therapeutic, effect [203].

This work takes it further and argues that chatbots can mediate an even more honest level of disclosure for their human-like and nonhuman qualities. Ironically speaking, that chatbots can talk like humans makes it possible for many users to start telling stories. However, the fact that they cannot inherently understand what they mean helps the disclosure process. The very existence of a virtual audience made participants feel easier to write about their stress and how to go about it in writing. In Bonobot and Diarybot studies alike, participants said that talking with the chatbot was easier and less burdensome than with their friends, families and other acquaintances. Some participants that talked with Bonobot, while they felt it was much like a counsellor or a therapist they met before, felt more comfortable with Bonobot because they would not have to worry about the thoughts and feelings of an interlocutor. For Diarybot, some participants never had any experience of telling stories of their trauma to others, and when they did, they were glad that they did so. Some only decided to do so because it was a chatbot. Participants were well aware of the fact that they were talking to a chatbot, not a human being, which was better than the other way around. Here, chatbots are at a unique place in human-computer interaction in that while their human-like features support human interactions, the interactions are essentially free of human bias.

Earlier work has shown that people readily engage in social sharing of emotions for various purposes including venting, help seeking, bonding, empathy and so on [88]. Emotion induces social sharing [43,219], and it benefits the discloser [285]. The benefits are however not voluntarily sought out by many people due to a variety of reasons, including saving face and shame [50]. Expressive writing is one of the best alternatives, writing in private but enabling the communication of disclosing emotions. However, it has been argued that writing in an experimental condition is also an inherently social process, accompanied by the researcher himself [214]. Thus chatbots can very well suit the role of listener in the social disclosure process, and design should address it.

In this vein, ensuring user safety and privacy will have to be the first and most necessary condition for any chatbot designed to encourage reflection on a sensitive and private topic. Not only does the system need to be equipped with all technical requirements to protect privacy and personal information, but it is also critical that the bot makes it communicates it to the user explicitly. Both Bonobot and Diarybot had a privacy notice for the user to feel safe. Bonobot delivered a short message not to worry about personal information, and Diarybot also mentioned that it will keep every conversation private. Participants later said in the interviews that these messages held them rest assured about what to say and how to describe it in the conversation.

### 5.1.2. Routine and Variety in Interaction

The chatbot interactions in this study mostly took the form of open-ended questions, because questions invite answers [229] and therefore can keep the conversation going, as well as serve as an effective strategy to draw out thoughts and feelings. In Bonobot, open-ended questions in motivational interviewing helped users to explore their problems. In Diarybot, expressive writing led self-disclosure of traumatic events. Yet the questions were carefully designed and arranged in iteration to find the best possible conversational sequence to make a chatbot conversation. Informed by conversation analysis (CA) research (e.g. [228,229,234]), the order and sequence of chatbot utterances and turns were decided to take after a naturally occurring real-life conversation.

Next, contextual understanding is pivotal in natural conversations [172]; yet is hardly achieved with the technological state of the art. In this work, a minimal contextual understanding was aimed for and achieved by extracting keywords from user input. For Bonobot, graduate school-related keywords were pre-defined and weighted for selection and reassemble in responses. In Diarybot, sentiments and key relationships were algorithmically selected and returned to the user in the form of questions. Both strategies are limited in that they cannot capture the flow and context of the conversation, especially the user intent; however, it intrigued participants to stay engaged in the conversation and continue within the conversation.

In addition, it is trickier but critical for the bot to deliver responses that convey an accurate empathy. Empathy is like a glue that builds the bonding between the user and the agent [277]. The bonding can lead to user satisfaction and reciprocated empathy from users, which can help with disclosure [99]. In this work, Bonobot attempted reflective listening and other MI-related skills for the relational component of motivational interviewing. Though most were appreciative, they demanded more personalized and accurate expressions that suit the context of the conversation. Still, they liked them when the bot responses correctly matched with the graduate school context, which were more phrased in a more targeted fashion. Thus, to ensure user engagement, guidance must be designed and organized in a way that follows the implicit rules of human communication.

Though chatbots communicate in a natural and familiar fashion, what they communicate need be varied and surprising in an inspiring manner. Users expect more than just chit-chat when they are about to share stories from the deep inside. More specifically, their likings toward Bonobot's evoking questions and Diarybot's follow-up in Responsive chat show that their pursuit for understanding and meaning. What these questions had in common was that they were directive; both Bonobot's and Diarybot's questions had a goal for the participants to focus on a certain path of thinking in the reflection process: for Bonobot, it was the idea of change [165], and for Diarybot, perspective switching [23].

Moreover, the questions were challenging in that they made participants explain themselves. Throughout the conversation, they had to explain their motives, reasons and frame of reference to respond to the bot's questions. This can help the users in two ways: First, it prevents brooding and rumination. Research on self-reflection has cautioned the risk of rumination or brooding (e.g. [104,261]). When reflecting on life's most difficult experiences, it is natural that one may fall into negative thoughts and feelings. The problem is these may backfire on reflective thinking. Moreover, Rime suggests that in addition to social sharing of emotions, the emotional event can finally come to a closure when one can cognitively articulate its meaning [219]. Instead of venting and seeking moral support, the event needs to be seen in a new light and defined with an interpretation. If this is not satisfied, Rime argues that the emotional route will repeat. Hence if the nudge provided by the bot can gently stir thinking on unexplored aspects of an emotional experience, it can lead to discovering self-insight and gaining self-knowledge, as evidenced by the participants in the study.

### 5.1.3. Reflection as Continued Experience

Reflection is not a static activity, and thoughts and feelings change over time. Hence it is also important for a reflection assistant to adapt to the changing user context for a long-term use. Here, context awareness, apart from contextual understanding in natural language processing, is important for two reasons. First, users would like to make sure that a mutual understanding is reached in communicating emotions [181]. For example, in the Diarybot study where participants had to repeat the procedure for four days, some participants inquired whether Diarybot caught on the different meanings of the use of the same word, "friend," for the bot returned questions about the friend. Since there was no sign to indicate whether Diarybot told the friend from yesterday apart from the friend today, the participant wanted to make sure. Moreover, the user may stay in the same frame of reference but his or her receptions and interpretations may change over time. Previous research suggests that negative life events are seen from different lenses over time [168]. The design of reflection assistants should take this into consideration that perspectives and interpretations may change, and be ready to support different paths in thinking.

Finally, adapting to user context also means that the bot walks the user through the thought process. As the user tries to reflect on a problem with the bot, it wouldn't repeat the same questions, which would wane engagement and eventually fail user expectations. Diversifying chatbot guidance according to the user's life and usage context also needs to be considered in the design process. To pick up where the user left off, continue on from a previous thought, and revisit feelings from a web of episodic memories will also be a technological challenge as much as a design challenge. To design a reflection assistant is truly an art in and of itself, since it essentially models after a friend who listens and advises -- one to whom we can confide in and who asks questions that matter and support us when we need it. Intelligent agents now face a number of challenges as they try to tap into the subjective realm of meaning, e.g. reflection and wellbeing [189]. How meanings are explained, explored and transformed via AI-assisted reflection processes will largely depend on how users perceive it, by effectively managing user expectations.

## 5.2. Tensions in Design

Intelligent agents are here to stay [184], and they already accompany every aspect of our lives. Though the technical complexities of the AI-induced agents and systems are increasing, how we define and shape the interaction with AI is relatively less explored. Understanding how users perceive and respond to the unseen technology will be key in designing AI interactions and designing their experiences around agents that are embedded in our lives.

### 5.2.1. Adaptivity

In this work, chatbots are user-adaptive, meaning that they are user interfaces that adapt to user based on processes of user model acquisition and application that involve some form of learning, inference, or decision making [117]. In other words, the bots deliver responses from user input, which results in compromising predictability and design transparency, the two important usability principles. One of the key features of chatbots in this study was that they try to follow up with user input. Participants responded to this in two broad ways. First, they tried to figure out the workings of the chatbot by making alterations in their narratives. Second, they showed an increased engagement once they found out that the bot was able to ask different questions depending on their narrative. Earlier work on interactive systems points out that users like to have control [107], and in direct manipulation systems that results in predictable outcomes, what users get being mapped onto their input. However, in a naturally occurring conversation an interlocutor does not always respond in the exactly same manner all the time. Using natural language as input modality, the chatbot is also expected to mimic human behavior in communication. This means that while the conversational routine may be predictable, the bot responses may not. Also, because users are not exposed to the inner workings of the response generator, an inherent lack exists in its internal transparency [107]. In fact, Jameson [117] has noted that for user-adaptive systems predictability and transparency can work at a global level, e.g. the layout and overall behavior of the system. Yet he has also cautioned that anthropomorphic representations of adaptive systems may invoke unrealistically high user expectations on system competence, e.g. natural language capabilities and task understanding [117]. Thus when it comes to chatbots, there is a tradeoff in system predictability and transparency: the system needs to be predictable and transparent to allow users control; however, doing so may result in failing their expectations for a human-like behavior. To tackle this problem, as Liberman and Selker [146] suggested, giving users the ability to adjust the degree of initiative may be an option; nonetheless, it may also cause confusion in the global predictability of the system. In the end, one-size cannot fit all; no one talks in a single style, and the responses are bound to change since we are human. This is an interesting yet challenging design problem for HCI researchers, practitioners and engineers alike, to design for predictability and transparency of chatbots as user-adaptive systems.

### 5.2.2. Autonomy

Another interesting observation from reflecting with chatbots was that participants would willingly trade their freedom and will to make decisions, i.e. self-determination, with answers that they may expect to get from the bot. In fact, seeking answers and solution behavior is one of the natural responses to stressors. Oftentimes, people engage in social sharing of emotions, but it does not resolve the stress or difficulties before the emotions are finally articulated, defined and labeled [219]. In the process people may even internally experience confusion, frustration, and sometimes ambivalence between opposing views. Out of helplessness, participants showed the tendency to turn to the bot and delegate the decision-making to the bot. For intelligent agents to show human-like capabilities it risks a dependency relationship that users may give up on control [146]. Though it had earlier been pointed out that users may confuse anthropomorphic agents with fellow human beings [135,243], Lieberman and Selker [146] note that people are good at differentiating the two, though there is always the danger that people will treat the agent as a real person, overextending their humanlike behavior. The participants' behavior toward the bot, however, takes a slightly different view: they were not only treating the agent as if it were a human being, but also as if it had superhuman capabilities. For instance, they wanted the bot not only to search information for them but also to make instant analyses for their problem situation. This indicates that they were very well aware of the fact that they were not talking to a human being; knowing that they were talking to a machine agent led them to put off all the tasks and intelligence needed in coping. Here, user expectations are high and they are even willing to delegate their decision-making to the bot. Whether they will actually act upon the machine-made decisions is a different matter. Instead, the more important question is: how should designers respond to user requests like this? Instead of full and perfect automation of an AI agent, we can instead let humans participate in the design loop of technology. Ge Wang makes an argument for "humans-in-the-loop" for designing interactive AI systems [291]. Instead of designing interactive AI as a system that is perfectly designed once-and-for-all, he insists that human users participate in the design process so that the machine learns to help [291]. This view works for reflection assistants that try to help users articulate and interpret their emotional experiences. Because individuals themselves are the sole survivors of the emotional experiences, such a system should value their agency and autonomy. Moreover, it fits the process of reflection as well [247]. Incorporating granularity in the design and scaffolding it fits stepwise processes for designing different types of reflection. Better yet, such an approach will safely reserve user control and autonomy in interacting with the agents, and serve as an effective means of managing expectations toward the agents as a tool, not an "oracle" [291].

### 5.2.3. Algorithmic Affordance

The discussion on chatbot predictability and transparency, as well as user control and autonomy leads to rethinking the concept of affordance in intelligent agents. In HCI, affordances are defined as "the perceived or actual properties of the thing, primarily those fundamental properties that determine just how the thing could possibly be used" [183]. Though the term was originally coined by ecological psychologist James Gibson [75,76], Don Norman's affordances has been established as an important pillar of HCI and interaction design. Especially with graphical user interfaces of the time, affordances were well positioned for providing the very appealing visual cues to the operation of things around us [183]. Over time, we see that agents increasingly become interface-less; they lose the physical-visual cues that hint their functions and operations. For conversational agents, it's more common nowadays that most of them are merely floating chat screens or borrow a friend instance on chat apps. The disappearance of distinctive visual metaphor makes it even more elusive to differentiate them from one another. Though conversational agents use natural language as the primary modality for interaction, the scope and variety of functions that they perform are strikingly different. Even in this work, the linguistic capabilities as well as the conversational styles of the two bots are different. The inconsistency of chat interactions is precisely what makes it much harder to manage user expectations. What is needed, then, is a concerted term for the affordances to refer to workings of disembodied conversational agents for design. The term, "algorithmic affordance" is proposed here to describe the affordance of intelligent agents whose operations are invisible to the user before use. Much of the chat functionality of many AI-induced systems is hidden from the user. It is very well expected that they can perform certain tasks very well, but exactly how they do it and how users should engage with them to make them work is not explained in *a priori* but experienced and learned *a posteriori*. Yet this proposal risks lowered user engagement, as participants in this study pointed out after they assumed and finally learned conversational routines of Responsive chat in Diarybot. This can be remedied by diversifying interactions and natural language output. The bigger challenge rather lies in how to hold agents explainable as well as responsible for telling users how they do what they do, and find ways to make it work. Should agents be held accountable for explaining its algorithmic affordances? Would users actually want to know them all, or would they rather want to wait to be surprised? How "natural" should chatbots be in communicating with users anyway? For agents to engage in deeply human cognitive processes such as self-reflection, their chat interactions convey more than just a receipt or token of exchange. Every utterance needs to have a point or a purpose, carrying a note of human empathy at the same time hinting its shrewd mechanical intelligence. It is a truly daunting task for chatbot designers alone. It requires a joint and orchestrated effort from HCI researchers, engineers, interaction designers, and communications and linguistics experts. In the advent of everything that is AI, algorithmic affordance is only beginning to take shape yet at quite a speed. In the end, it pulls the fundamental question in designing human-AI interaction: how should we human users interact with the machine black box? The next section will ponder the idea of meaning-making with AI, to suggest that algorithmic affordance is in the eye of the beholder.

## 5.3. Meaning-Making as Design Metaphor

This work is centered on designing and implementing chatbots to aid an individual's voluntary reflection on life's struggles. The chatbots in this study, Bonobot and Diarybot, intervene the process of scaffolding the reflection via questions and feedback, aiming to trigger new ideas and lead users on a non-ruminating path. So far, it's been discussed what potential chatbots have as reflection assistants and how their invisible algorithmic workings may create tensions in design, user expectations, and designing for conversational user experiences. At the heart of this problem is the gap that can never be closed: humans can think, while machines cannot, though they may appear *as if* they do. Users engage in a social interaction with chatbots, and in providing guidance for their self-reflection chatbots carry a purpose and meaning, both implicitly and explicitly. Thus they engage in an interpretive process, trying to define the bot's actions. The bot responses are not taken for granted or at face value but attached and ascribed meanings that users try to negotiate and construct, not by themselves but with the bot. It is a *symbolic* interaction [21,87], and this work suggests *meaning-making* as a novel metaphor in human-AI interaction. The interaction essentially mediates users' inner conversation [87] to sort out the meaning of an adversity in life. AI plays a role of nudging the process.

### 5.3.1. Meaning in Reflection

In psychology, reflecting on negative experiences in life necessitates a search for meaning, a coherent understanding of the event to find solace and value in life. Victor Frankl was one of the first to emphasize "man's will to meaning" following the Holocaust [71]. Existential psychologists further pondered on meaning in humans' coping with adversities (e.g. [11,283]. The role of meaning is increasingly being valued in promoting wellbeing as well [206,226]. In fact, there are evidence-based studies that show the experience of meaning for wellbeing [93,113,151,226,254].

What exactly is meaning? In social psychologist Roy Baumeister's terms, meaning is a "mental representation of possible relationships among things, events, and relationships" [12]. Put more eloquently, also social psychologist Shelley Taylor [262] wrote:

> *"Meaning is an effort to understand the event: why it happened and what impact it has had. The search for meaning attempts to*

> *answer the question, 'What is the significance of the event?' Meaning is exemplified by the results of an attributional search that answers the question, 'What caused the event to happen?'. . . Meaning is also reflected in the answer to the question, 'What does my life mean now?'"*

The achievement of meaning shares much with the goal of designing guided disclosure for different reflection processes: by scaffolding the thoughts upon life's most difficult experiences, users are encouraged to ponder on what the events had meant for their life. Hence despite its fluidity, meaning or the achievement of meaning is critical in confronting highly stressful life experiences [189].

Recovering from a stressful event involves reducing the discrepancy between its appraised meaning and global beliefs and goals [120]. The process or activity of meaning-making refers to the processes in which people engage to reduce this discrepancy [189]. Park [189] has delineated four categorical schemes in meaning-making [189]: automatic/deliberate, assimilation/accommodation processes, searching for comprehensibility/ significance, and cognitive/emotional processing. These schemes are not mutually exclusive, but particularly relevant to this work is the cognitive and emotional processing of meaning making.

Cognitive processing emphasizes the reworking of one's beliefs, while emotional processing highlights the experiencing and exploring of one's emotions. It is more invested in exposure and habituation along with the regulation of negative affect [61] and attempts to understand one's feelings [252]. In contrast, cognitive processing emphasizes the cognitive aspects of integrating experiential information with preexisting schemas [118,278]. It involves reappraisals and repeated comparisons between one's experience and existing beliefs to modify one or the other [52,84], which is achieved through thoughtful reflection, including awareness of the emotions an event evokes and the effect it might have on one's future [28]. All reflection processes in this work were designed to help users be aware of emotions and think through their impact on themselves, and the findings support so. Expressive writing studies have also suggested that both emotional and cognitive processes contribute to meaning-making (e.g., [246,270]). Though routes may be different, all are searching for meaning.

### 5.3.2. Meaning-Making as Interaction

Then how can meaning be achieved? Meanings are made from efforts to reduce discrepancies between appraised and global meanings [189]. According to Park [189], meanings can be in different forms. To name a few: sense of having "made sense," acceptance, reattributions, causal understanding, perceptions of growth, positive changes in life, changed self-identity, reappraised meaning of stressors, changed global beliefs and goals, and restored or changed sense of meaning in life. What these have in common is that meaning involves a *reinterpretation* of impact.

It is an important question for an HCI research to ask how the discrepancies can be resolved via human-computer interaction. Meaning research is relatively new in HCI, despite many findings on meaningful interactions (e.g. [25,41,109,127,134]). Recently, Mekler and Hornbaek offered a framework of meaning in interaction [156], yet meaning here refers to the quality of interaction, not the process of interaction itself. In psychology, meaning-making is a process in which one understands, construes, or makes sense of life events, relationships, and the self [114]. In this sense, meaning is an outcome of an intrapersonal experience. Thus the goal of designing human-computer interaction of a thoughtful reflection for meaning-making is to translate such inner workings of the self. Since little is investigated on meaning-making process in HCI, the metaphor of "meaning-making" is borrowed from educational critics Neil Postman and Charles Weingartner [211], on teaching and learning:

> *"[Meaning-making] is, to begin with, much less static than the others. It stresses a process view of minding, including the fact that "minding" is undergoing constant change. [It] also forces us to focus on the individuality and the uniqueness of the meaning maker (the minder). In most of the other metaphors there is an assumption of "sameness" in all learners. The "garden" to be cultivated, the darkness to be lighted, the foundation to be built upon, the clay to be molded—there is always the implication that all learning will occur in the same way. The flowers will be the same color, the light will reveal the same room, the clay will take the same shape, and so on. Moreover, such metaphors imply boundaries, a limit to learning. How many flowers can a garden hold? How much water can a bucket take? What happens to the learner after his mind has been molded? How large can a building be, even if constructed on a solid foundation? The "meaning maker" has no such limitation. There is no end to his educative process. He continues to create new meanings..."*

Not surprisingly, this view on meaning-making aligns with adjustment to stressful events in psychology. Attempting to make meaning is not always linear but ongoing, for which an individual strives to make sense of discrepancies continuously. Efforts to make meaning gradually move toward reducing the discrepancies [189], yet in such process it may spiral downhill. If this is to be done by an individual, he or she has to make a long way to make sense of the stressful event, the self, and the others around the self, because meanings are not easily achieved. This work has sought to find a solution to minimize meaning-making attempts that may result in rumination but encourage constructive cognitive processing via chatbot interactions. Findings suggest that while it could not always be guaranteed that meanings were made, all attempts were successful in leading the thought process on a positive path. Thus this work proposes meaning-making as a novel design metaphor for chat interactions, or further, human-AI interaction where the agent partners with users in trying to make sense of life's agony.

### 5.3.3. Making Meanings with AI

Can AI make meanings for us? Trying to answer this question would be in vain as because machines simply cannot. Meanings are highly subjective in nature and can change over time [189]. Reflecting on life's unresolved stress from struggles, suffering and sorrow is difficult; one often shuns away from doing it. Nonetheless, when done, it can teach beautiful lessons to learn and grow, enriching the next chapter of life.

This work contributes the design of a technology that can help us *think*. It talks to us in ways that scaffold our thinking process in a stepwise manner. It asks us what we wish to be different and asks about key relationships from our stories. While this can be done with a family member or a friend, but the outcome would not be the same. Agents, whether they are chatbots, embodied agents or voice user interfaces, are and will never be human. Interactions with machine agents can pride on computational efficiency, especially in terms of decision-making and logic. This means that they can be designed to support a mechanical, highly structured exploration of an often-complicated emotional event. Both Bonobot and Diarybot in this work delivered a structured conversation;

unlike human conversations, the interaction was designed in a way to follow an algorithm to navigate the human mind. This is also not like the "therapist" chatbots that emulate a human counsellor or deliver a treatment. These chatbot therapists, they risk user engagement in the long run, since their conversations are not continuous but only reach a dead-end that is a repeated therapy exercise. Also, users with a critical condition are encouraged to eventually connect to a human counsellor in the end. The same applies to peer chat interactions, though the opposite. While peers can learn to deliver programmed chats, humans are excellent at wit and caprice that their improvisations and empathy will best work towards therapeutic transference. Agents are not human, and that is their best policy. For long we have endeavored to make machines work like us, and it has been successful. However, the undesirable consequence is that expectations fail and engagement wanes. This work argues that their algorithms be best manipulated, designed and perfected to serve logical, systemic and organized thinking on life's complications.

In addition, the interactions with agents are, in fact, not social. They are social in nature that the interactions mimic those between humans; however, they are not designed as human beings and their interactions will not end up in contributing to community knowledge but only self-knowledge. In other words, human-AI interaction can only benefit the user himself or herself, unlike human-human interaction. However, interacting with agents can impact the society. Technologies are scalable, and their impact reaches millions. Though they may not be connected, but technologies link them. In the famous words of Sherry Turkle, we are in this alone "but" together in two ways. First, in AI-abound society, technologies may isolate us but surround us. Second, we do not feel alone since we are in interactions with AI. In the end, the virtual togetherness would create an inherently subjective experience especially in terms of reflection.

Finally, perhaps the more important question to ask would be: Can we make meanings with AI? This work shows promise. Conversational agents with general artificial intelligence will continue to advance in natural language understanding, creating an as-if-human experience for users. Yet again, this work makes the point that the benefit of engaging an AI agent rather than a family, friend or counsellor is for its exceptional computational power. Algorithmic affordances of AI may or may not allow us to comprehend why is it that AI does what it does, but we will certainly be able to make sense of what it is, in our own subjective world. In such process the discrepancies will be resolved, both between us and machines, as well as us and life. Meanings will be made. This work highlights the partnership with AI to make sense of the most personal experiences in life. Instead of having AI the smart know-it-all, this work invites AI to be an intelligent *nudge* – what it nudges into will be actively sought and interpreted by users, with assumptions made and meaning created in the process. That will precisely be how the reflecting individual continues to engage in the making of meanings with AI.

# Chapter 6. Conclusion

## 6.1. Research Summary

This work aimed to design and implement conversational agents that encourage user narrative and self-reflection in mental wellbeing. More specifically, it was motivated to design a chatbot that can engage in self-reflection processes in which users can explain their difficult life events, explore untapped meanings, and promote change in behavior if needed. Though self-reflection does not usually involve a third-party intervention, reviewing prior work has informed that it tends to involve brooding, to which appropriate guidance can help. Thus this work has constructed design space with user disclosure and chatbot guidance, where depending on the levels of disclosure and guidance four reflection subspaces were delineated and investigated.

Most technologies designed for reflection have focused on making technologies a medium to revisit the past experiences and review self-tracking data. Increasing the level of guidance and support for disclosure would help technologies engage in other types of reflections such as explaining, exploring and transformative. In this work, conversational agents were suggested as the best means of technological intervention to scaffold the reflection process. Based on levels of disclosure and guidance, three chatbots, one of which takes after motivational interviewing and the other two based on expressive writing, were designed and implemented. First, Bonobot was designed for a transformative reflection via a structured conversation to promote behavior change. Diarybot was designed to encourage explaining and exploring reflections based on expressive writing and follow-up prompts upon rethinking a past trauma. Bonobot was implemented as a web application, and Diarybot on a messenger app, in two versions: Basic and Responsive. The chatbots were each set up for an experimental user study with 30 participants.

Findings of the study are as follows. A qualitative method was used in the user study with Bonobot, to dig deeper into the conversational user experiences of using a chatbot for self-reflection. Participants mostly appreciated the evoking questions from motivational interviewing, for a chance of refreshing their goals and perspectives on a stressful situation. In addition, Bonobot was mostly in charge of the conversation by leading questions and providing MI-adherent responses such as reflections and affirmations. This active guidance was appreciated, but only when it worked in the correct context of conversation. Finally, perhaps due to Bonobot's proactive guidance throughout the conversation, participants had a number of requests for additional chat functionality that was essentially beyond human and machine intelligence. The Bonobot study, taken together, shows that while a chatbot can lead a goal-oriented conversation, it may risk user autonomy and independence in decision by imposing too much guidance.

Diarybot was designed and implemented based on lessons learned from the Bonobot study. This time, the aim was to let the user explore their own narrative, rather than the bot leading the user to explore. Moreover, two different types of chat, Basic chat and Responsive chat, were created to compare the chatbot effect and the chat interaction effect, both against baseline created in a Google document. The findings show that chatbots can play a role of a virtual audience and/or a reader for participants to tell their stories. Participants in Basic and Responsive chat conditions rated significantly higher for their feeling heard, as opposed to Google document participants. Yet increased interaction in Responsive chat received significantly higher ratings on perceived enjoyability, perceived sociability, trust and intention to use. This shows that more chat interactions with Diarybot instantaneously transformed the expressive writing activity into a fun and enjoyable conversation with a nonhuman agent. Perhaps the most interesting finding was observed from the survey results on ease of emotional expression and difficulty in writing the highest and lowest, respectively, for Basic chat; while in contrast higher user engagement and proactive adaptation to the bot algorithm were observed in Responsive chat. Taken together, making chat exchanges with Diarybot was fun and enjoyable, and participants would willingly make alterations in their own trauma narratives to suit the workings of the bot. During the four-day experiment, they gradually learned the patterns of the bot conversation and tried to fit their writing into it, by keeping words consistent and choosing the topic of writing to make the algorithm work properly. This observation poses an interesting question to HCI community regarding the fundamental principles of UX on interface predictability and transparency. Unpredictable outcomes of interaction incentivize continued user engagement; moreover, opaque, unexplained design intrigued proactive narrative making.

The findings point to interesting tensions and boundaries between AI and human-AI interaction. As AI increasingly permeates into our lives, they are shaping different realities. As new interactions emerge, users adapt to the workings of agents without completely decoding the black box, creating tensions in system predictability and design transparency as well as user engagement. Study participants showed a great interest in figuring out how the bot worked, but once they recognized the pattern, their engagement waned bit by bit. Furthermore, unrealistically high expectations for an intelligent agent may also risk user autonomy or self-determination. Participants wanted to be assisted by the bot with solutions to resolve their stress once and for all. Managing expectations would be a key challenge in designing AI to engage in reflection.

Moreover, these tensions lead to rethink perceived affordances, and suggest algorithmic affordances as an alternative. It captures the chasm between user expectations and AI functionality, as well as the challenge to capture vast possibilities of user receptions to the algorithms. In other words, algorithmic affordances signify the hidden and invisible workings of AI as the black box, and it depends more on user's own conceptions and expectations of the bot that may determine user experiences. Moreover, algorithmic affordances are especially important in reflection design, in that users communicate with AI in language. Language carries symbols, negotiated meanings, and social construction of the reality we live in. The bot guidance is not mere words of meaningless response but perceived as having an intention and a purpose. The symbolic interaction between users and AI in reflecting on life's most difficult experiences leads to the proposal of making meanings with AI. In the end, AI nudges us humans into tap into the unexplored meanings of life's miseries and sorrows. Meanings are created when the cognitive gaps are closed in the interpretive process. In the era where AI agents prevail, what we may need is something beyond explainable AI; it's responsible AI which will keep us intact when traumas get rewired in our narrative.

## 6.2. Limitations and Future Work

This work is bound with limitations. In terms of conversation design for chatbots, it used a limited number of rule-based sequences to generate the bot responses. Had there been more chat sequences available, user experiences would have prevailed in many different possible ways that would help to dig deeper into their perceptions of the bot and experiences in reflection. Moreover, both Bonobot and Diarybot conversations are not without cultural bias in participant recruitment. Bonobot participants were Korean nationals; however, they talked in English as Bonobot's was implemented in English for MI skills. Given the requirement of the study to talk about stressful experiences, nuances in language might have translated differently in their second language use. Diarybot participants were also Korean nationals who were regular users of the messenger app Kakao, on which Diarybot was built. Their familiarity with the app might have had preconceptions about the bot and its functionalities.

In the experimental setup, the study could only recruit a limited number of participants for a limited duration, due to resource constraints. Though Diarybot replicated the original setup of Pennebaker's expressive writing [196,200], the recruited 30 participants had to be assigned to three conditions, resulting in a handful number of participants in each. Future work may use a larger sample for statistical confidence in data analysis. On the other hand, all 30 participants had a chance to talk with Bonobot, but only for once. Further interactions with Bonobot might have revealed user experiences in the continued exercise of transformative reflection, which can also be explored in future work. Additionally, participants were recruited in a higher-ed institution, which, in a greater context, may not represent the general population in literacy and familiarity with technology. In this work, this was mostly taken care of by recruiting participants from as diverse academic backgrounds and age groups as possible.

Another concern may be that of participants' psychological wellbeing. Both Bonobot and Diarybot experiments recruited participants from non-clinical population, meaning that they would voluntarily share their stress and traumas, but they were not explicitly with any mental health conditions. Given the institutional setup, it was quite difficult to reach participants with clinical issues. Plus, given the resource constraints, it was also difficult to conduct a longitudinal observation of their wellbeing. Despite the difficulties, however, future work should address the trends and differences in psychological wellbeing as a result of short, moderate and long-term interaction with AI reflection assistant. It is, however, also worth noting that mental wellbeing is a highly variable and complicated subject that is hardly improved via chatbot interactions. The focus should rather be on how users interact with these technologies and how such an interaction may or may not play a role in their wellbeing, which is the central interest of HCI researchers and interaction designers.

Last but not least, this work has explored an operationalized design space with disclosure support and guidance, in which many different types of chatbot-guided reflections can be designed. Chatbots in this work were only based on person-centered methods, i.e. expressive writing and motivational interviewing. Other approaches to reflection in the future will expand the scope and comprehensiveness of self-reflection with agent intermediaries.

## 6.3. Final Remarks

This work has pioneered the less explored intersection between HCI and self-reflection, designing chatbot technology as an active mediator for one’s reflective pondering on emotionally troubled experiences. This problem is in fact a huge design challenge in that it spans various disciplines including HCI, psychology, linguistics and communication theories. However, the true challenge has been deciphering what the chatbot interactions have meant for the users in narrating their pain and sorrows. Through the design of Bonobot and Diarybot, this work has achieved a series of chat interactions that could successfully scaffold a reflection behavior. Moreover, findings gathered from experimental user studies point to tensions that challenge existing notions in HCI that would open up new directions in design. Finally, this work offers telling evidence for the need for such an interdisciplinary research in HCI community in the advent of new reality in which intelligent agents serve us, listen to us and help us make meanings in life. The findings are also relevant to the industrial and commercial applications of conversational agents, especially to inform the emerging conversational UX design and natural conversation framework (NCF). At last, this study contributes to the necessary yet insufficient discussion on designing AI as our invisible cohabitant, and as a new companion to our journey in self-discovery.

# Bibliography

[1] Hussein A. Abbass. 2019. Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust. *Cogn Comput* 11, 2 (April 2019), 159–171. DOI:https://doi.org/10.1007/s12559-018-9619-0

[2] M. Alvarez-Jimenez, S. Bendall, R. Lederman, G. Wadley, G. Chinnery, S. Vargas, M. Larkin, E. Killackey, P. D. McGorry, and J. F. Gleeson. 2013. On the HORYZON: Moderated online social therapy for long-term recovery in first episode psychosis. *Schizophrenia Research* 143, 1 (January 2013), 143–149. DOI:https://doi.org/10.1016/j.schres.2012.10.009

[3] Saleema Amershi, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Eric Horvitz, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, and Paul N. Bennett. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, ACM Press, Glasgow, Scotland Uk, 1–13. DOI:https://doi.org/10.1145/3290605.3300233

[4] Ian Anderson, Julie Maitland, Scott Sherwood, Louise Barkhuus, Matthew Chalmers, Malcolm Hall, Barry Brown, and Henk Muller. 2007. Shakra: Tracking and Sharing Daily Activity Levels with Unaugmented Mobile Phones. *Mobile Netw Appl* 12, 2 (June 2007), 185–199. DOI:https://doi.org/10.1007/s11036-007-0011-7

[5] Hal Arkowitz, Henny A. Westra, William R. Miller, and Stephen Rollnick (Eds.). 2007. *Motivational interviewing in the treatment of psychological problems* (1st Edition ed.). The Guilford Press, New York.

[6] Sue Atkins and Kathy Murphy. 1993. Reflection: a review of the literature. *Journal of Advanced Nursing* 18, 8 (1993), 1188–1192. DOI:https://doi.org/10.1046/j.1365-2648.1993.18081188.x

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]* (May 2016). Retrieved May 3, 2020 from http://arxiv.org/abs/1409.0473

[8] Karen A. Baikie and Kay Wilhelm. 2005. Emotional and physical health benefits of expressive writing. *Advances in Psychiatric Treatment* 11, 5 (September 2005), 338–346. DOI:https://doi.org/10.1192/apt.11.5.338

[9] Michael Barkham, Bridgette Bewick, Tracy Mullin, Simon Gilbody, Janice Connell, Jane Cahill, John Mellor-Clark, David Richards, Gisela Unsworth, and Chris Evans. 2013. The CORE-10: A short measure of psychological distress for routine use in the psychological therapies.

*Counselling and Psychotherapy Research* 13, (January 2013), 1–13.

[10] Lisa J. Barney, Kathleen M. Griffiths, and Michelle A. Banfield. 2011. Explicit and implicit information needs of people with depression: a qualitative investigation of problems reported on an online depression support forum. *BMC Psychiatry* 11, 1 (May 2011), 88. DOI:https://doi.org/10.1186/1471-244X-11-88

[11] Alexander Batthyany and Pninit Russo-Netzer. 2014. Psychologies of Meaning. In *Meaning in Positive and Existential Psychology*, Alexander Batthyany and Pninit Russo-Netzer (eds.). Springer, New York, NY, 3–22. DOI:https://doi.org/10.1007/978-1-4939-0308-5_1

[12] Roy F. Baumeister. 1991. *Meanings of Life*. Guilford Press.

[13] Eric P.S. Baumer. 2015. Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), Association for Computing Machinery, Seoul, Republic of Korea, 585–594. DOI:https://doi.org/10.1145/2702123.2702234

[14] Eric P.S. Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 conference on Designing interactive systems* (DIS '14), Association for Computing Machinery, Vancouver, BC, Canada, 93–102. DOI:https://doi.org/10.1145/2598510.2598598

[15] Nancy Beckham. 2007. Motivational interviewing with hazardous drinkers. *Journal of the American Academy of Nurse Practitioners* 19, 2 (2007), 103–110. DOI:https://doi.org/10.1111/j.1745-7599.2006.00200.x

[16] Alison Bell and Stephen Rollnick. 1996. Motivational interviewing in practice: A structured approach. In *Treating substance abuse: Theory and technique*. The Guilford Press, New York, NY, US, 266–285.

[17] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3, Feb (2003), 1137–1155.

[18] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2010. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence* 24, 6 (July 2010), 648–666. DOI:https://doi.org/10.1080/08839514.2010.492259

[19] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction* 12, 2 (June 2005), 293–327.

[20] Nataly Birbeck, Shaun Lawson, Kellie Morrissey, Tim Rapley, and Patrick Olivier. 2017. Self Harmony: Rethinking Hackathons to Design and Critique Digital Technologies for Those Affected by Self-Harm. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), Association for Computing Machinery, Denver, Colorado, USA, 146–157. DOI:https://doi.org/10.1145/3025453.3025931

[21] Herbert Blumer. 1986. *Symbolic Interactionism: Perspective and Method*.

University of California Press.

[22] Gillie Bolton. 2010. *Reflective Practice: Writing and Professional Development*. SAGE.

[23] Mark Bond and James W. Pennebaker. 2012. Automated computer-based feedback in expressive writing. *Computers in Human Behavior* 28, 3 (May 2012), 1014–1018. DOI:https://doi.org/10.1016/j.chb.2012.01.003

[24] Roger J. Booth, Keith J. Petrie, and James W. Pennebaker. 1997. Changes in Circulating Lymphocyte Numbers Following Emotional Disclosure: Evidence of Buffering? *Stress Medicine* 13, 1 (1997), 23–29. DOI:https://doi.org/10.1002/(SICI)1099-1700(199701)13:1<23::AID-SMI714>3.0.CO;2-E

[25] Julia Ayumi Bopp, Elisa D. Mekler, and Klaus Opwis. 2016. Negative Emotion, Positive Experience? Emotionally Moving Moments in Digital Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), Association for Computing Machinery, San Jose, California, USA, 2996–3006. DOI:https://doi.org/10.1145/2858036.2858227

[26] David Boud. 2001. Using journal writing to enhance reflective practice. *New Directions for Adult and Continuing Education* 2001, 90 (2001), 9–18. DOI:https://doi.org/10.1002/ace.16

[27] David Boud, Rosemary Keogh, and David Walker. 1985. *Reflection: Turning Experience Into Learning*. Kogan Page.

[28] J. E. Bower, M. E. Kemeny, S. E. Taylor, and J. L. Fahey. 1998. Cognitive processing, discovery of meaning, CD4 decline, and AIDS-related mortality among bereaved HIV-seropositive men. *J Consult Clin Psychol* 66, 6 (December 1998), 979–986. DOI:https://doi.org/10.1037//0022-006x.66.6.979

[29] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (January 2006), 77–101. DOI:https://doi.org/10.1191/1478088706qp063oa

[30] Joan E. Broderick, Doerte U. Junghaenel, and Joseph E. Schwartz. 2005. Written Emotional Expression Produces Health Benefits in Fibromyalgia Patients. *Psychosomatic Medicine* 67, 2 (April 2005), 326–334. DOI:https://doi.org/10.1097/01.psy.0000156933.04566.bd

[31] Leslie R. Brody and Suzanne H. Park. 2004. Narratives, mindfulness, and the implicit audience. *Clinical Psychology: Science and Practice* 11, 2 (June 2004), 147–154. DOI:https://doi.org/10.1093/clipsy/bph065

[32] Brené Brown. 2012. Daring greatly. Retrieved February 16, 2020 from https://search.lib.byu.edu/byu/record/lee.6995454?holding=1uga7u6hr0c7io1f

[33] Elissa J. Brown and Richard G. Heimberg. 2001. Effects of Writing About Rape: Evaluating Pennebaker's Paradigm with a Severe Trauma. *J Trauma Stress* 14, 4 (October 2001), 781–790. DOI:https://doi.org/10.1023/A:1013098307063

[34] A. Bruce, I. Nourbakhsh, and R. Simmons. 2002. The role of

expressiveness and attention in human-robot interaction. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, 4138–4142 vol.4. DOI:https://doi.org/10.1109/ROBOT.2002.1014396

[35] Rebecca A. Burwell and Stephen R. Shirk. 2007. Subtypes of Rumination in Adolescence: Associations Between Brooding, Reflection, Depressive Symptoms, and Coping. *Journal of Clinical Child & Adolescent Psychology* 36, 1 (March 2007), 56–65. DOI:https://doi.org/10.1080/15374410709336568

[36] James Calderhead. 1989. Reflective teaching and teacher education. *Teaching and Teacher Education* 5, 1 (January 1989), 43–51. DOI:https://doi.org/10.1016/0742-051X(89)90018-8

[37] Gillian Cameron, David Cameron, Gavin Megaw, R. R. Bond, Maurice Mulvenna, Siobhan O'Neill, C. Armour, and Michael McTear. 2018. Best practices for designing chatbots in mental healthcare – A case study on iHelpr. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI-2018)*. DOI:http://dx.doi.org/10.14236/ewic/HCI2018.129

[38] Charles S. Carver, Michael F. Scheier, and Jagdish K. Weintraub. 1989. Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology* 56, 2 (February 1989), 267–283. DOI:https://doi.org/10.1037/0022-3514.56.2.267

[39] Christine Cheepen. 1988. *The predictability of informal conversation*. Pinter.

[40] Mei-whei Chen and Nan J. Giblin. 2017. *Individual counseling and therapy: Skills and techniques* (1st Edition ed.). Routledge, New York.

[41] EunJeong Cheon and Norman Makoto Su. 2018. The Value of Empty Space for Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, Montreal QC, Canada, 1–13. DOI:https://doi.org/10.1145/3173574.3173623

[42] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*. Retrieved May 2, 2020 from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8075

[43] Véronique Christophe and Bernard Rimé. 1997. Exposure to the social sharing of emotion: Emotional impact, listener responses and secondary social sharing. *European Journal of Social Psychology* 27, 1 (1997), 37–54. DOI:https://doi.org/10.1002/(SICI)1099-0992(199701)27:1<37::AID-EJSP806>3.0.CO;2-1

[44] Karen Church, Eve Hoggan, and Nuria Oliver. 2010. A study of mobile mood awareness and communication through MobiMood. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (NordiCHI '10), Association for Computing Machinery,

Reykjavik, Iceland, 128–137. DOI:https://doi.org/10.1145/1868914.1868933

[45] 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. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), Association for Computing Machinery, Glasgow, Scotland Uk, 1–12. DOI:https://doi.org/10.1145/3290605.3300705

[46] Sheldon Cohen. 1994. Perceived stress scale. Retrieved August 1, 2018 from http://mindgarden.com/documents/PerceivedStressScale.pdf

[47] Kenneth Mark Colby. 1975. *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. Elsevier.

[48] Sunny Consolvo, David W. McDonald, and James A. Landay. 2009. Theory-driven design strategies for technologies that support behavior change in everyday life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), Association for Computing Machinery, Boston, MA, USA, 405–414. DOI:https://doi.org/10.1145/1518701.1518766

[49] Jacqueline Corcoran. 2016. *Motivational interviewing: A workbook for social workers*. Oxford University Press.

[50] Patrick W. Corrigan. 2005. Dealing with stigma through personal disclosure. In *On the stigma of mental illness: Practical strategies for research and social change*. American Psychological Association, Washington, DC, US, 257–280. DOI:https://doi.org/10.1037/10887-012

[51] Catherine Crane, Thorsten Barnhofer, and J. Mark G. Williams. 2007. Reflection, brooding, and suicidality: A preliminary study of different types of rumination in individuals with a history of major depression. *British Journal of Clinical Psychology* 46, 4 (2007), 497–504. DOI:https://doi.org/10.1348/014466507X230895

[52] J. Creswell, Sue Lam, Annette Stanton, Shelley Taylor, Julienne Bower, and David Sherman. 2007. Does Self-Affirmation, Cognitive Processing, or Discovery of Meaning Explain Cancer-Related Health Benefits of Expressive Writing? *Personality & social psychology bulletin* 33, (March 2007), 238–50. DOI:https://doi.org/10.1177/0146167206294412

[53] Antonia S. Csillik. 2013. Understanding motivational interviewing effectiveness: Contributions from Rogers' client-centered approach. *The Humanistic Psychologist* 41, 4 (2013), 350–363. DOI:https://doi.org/10.1080/08873267.2013.779906

[54] Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the ATIS task: the ATIS-3 corpus. In *Proceedings of the workshop on Human Language Technology* (HLT '94), Association for Computational Linguistics,

Plainsboro, NJ, 43–48. DOI:https://doi.org/10.3115/1075812.1075823

[55] Gavin Daker-White and Anne Rogers. 2013. What is the potential for social networks and support to enhance future telehealth interventions for people with a diagnosis of schizophrenia: a critical interpretive synthesis. *BMC Psychiatry* 13, 1 (November 2013), 279. DOI:https://doi.org/10.1186/1471-244X-13-279

[56] Sharon Danoff-Burg, Catherine E. Mosher, Asani H. Seawell, and John D. Agee. 2010. Does narrative writing instruction enhance the benefits of expressive writing? *Anxiety, Stress, & Coping* 23, 3 (May 2010), 341–352. DOI:https://doi.org/10.1080/10615800903191137

[57] Ineke Demeyer, Evi De Lissnyder, Ernst H. W. Koster, and Rudi De Raedt. 2012. Rumination mediates the relationship between impaired cognitive control for emotional information and depressive symptoms: A prospective study in remitted depressed adults. *Behaviour Research and Therapy* 50, 5 (May 2012), 292–297. DOI:https://doi.org/10.1016/j.brat.2012.02.012

[58] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero. 2013. Recent advances in deep learning for speech research at Microsoft. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8604–8608. DOI:https://doi.org/10.1109/ICASSP.2013.6639345

[59] Antoine Douaihy, Thomas M. Kelly, and Melanie A. Gold (Eds.). 2015. *Motivational interviewing: A guide for medical trainees* (1st Edition ed.). Oxford University Press, Oxford.

[60] Hugh Dubberly, Rajiv Mehta, Shelley Evenson, and Paul Pangaro. 2010. Reframing health to embrace design of our own well-being. *interactions* 17, 3 (May 2010), 56–63. DOI:https://doi.org/10.1145/1744161.1744175

[61] Anke Ehlers and David M. Clark. 2006. Predictors of Chronic Posttraumatic Stress Disorder: Trauma Memories and Appraisals. In *Pathological anxiety: Emotional processing in etiology and treatment*. The Guilford Press, New York, NY, US, 39–55.

[62] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems. In *Proceedings of the 24th International Conference on World Wide Web* (WWW '15), International World Wide Web Conferences Steering Committee, Florence, Italy, 278–288. DOI:https://doi.org/10.1145/2736277.2741667

[63] Eva M. Epstein, Denise M. Sloan, and Brian P. Marx. 2005. Getting to the Heart of the Matter: Written Disclosure, Gender, and Heart Rate. *Psychosom Med* 67, 3 (2005), 413–419. DOI:https://doi.org/10.1097/01.psy.0000160474.82170.7b

[64] Brian A. Esterling, Michael H. Antoni, Mary Ann Fletcher, Scott Margulies, and Neil Schneiderman. 1994. Emotional disclosure through writing or speaking modulates latent Epstein-Barr virus antibody titers.

*Journal of Consulting and Clinical Psychology* 62, 1 (1994), 130–140. DOI:https://doi.org/10.1037/0022-006X.62.1.130

[65] Chris Evans, John Mellor-Clark, Frank Mar. 2000. CORE: Clinical Outcomes in Routine Evaluation. *Journal of Mental Health* 9, 3 (January 2000), 247–255. DOI:https://doi.org/10.1080/jmh.9.3.247.255

[66] Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ* 328, 7449 (May 2004), 1166. DOI:https://doi.org/10.1136/bmj.328.7449.1166

[67] Allan Fenigstein. 1987. On the Nature of Public and Private Self-Consciousness. *Journal of Personality* 55, 3 (1987), 543–554. DOI:https://doi.org/10.1111/j.1467-6494.1987.tb00450.x

[68] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. 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 Ment Health* 4, 2 (June 2017). DOI:https://doi.org/10.2196/mental.7785

[69] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (OZCHI '10), Association for Computing Machinery, Brisbane, Australia, 216–223. DOI:https://doi.org/10.1145/1952222.1952269

[70] B. J. Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December (December 2002), 5:2. DOI:https://doi.org/10.1145/764008.763957

[71] Viktor E. (Viktor Emil) Frankl. 2006. *Man's search for meaning*. Boston : Beacon Press. Retrieved June 23, 2020 from http://archive.org/details/manssearchforme00fran

[72] Pasquale G. Frisina, Joan C. Borod, and Stephen J. Lepore. 2004. A Meta-Analysis of the Effects of Written Emotional Disclosure on the Health Outcomes of Clinical Populations. *The Journal of Nervous and Mental Disease* 192, 9 (September 2004), 629–634. DOI:https://doi.org/10.1097/01.nmd.0000138317.30764.63

[73] Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, and Michiel Rauws. 2018. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment Health* 5, 4 (December 2018). DOI:https://doi.org/10.2196/mental.9782

[74] Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. 2018. Towards Empathetic Human-Robot Interactions. In *Computational Linguistics and Intelligent Text Processing* (Lecture Notes in Computer Science), Springer International Publishing, Cham, 173–193.

DOI:https://doi.org/10.1007/978-3-319-75487-1_14

[75] James J. Gibson. 1977. The theory of affordances. In *Perceiving, acting, and knowing: toward an ecological psychology*, John Bransford Robert E Shaw (ed.). Hillsdale, N.J. : Lawrence Erlbaum Associates, pp.67-82. Retrieved May 19, 2020 from https://hal.archives-ouvertes.fr/hal-00692033

[76] James J. Gibson. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

[77] Yori Gidron, Tuvia Peri, John F. Connolly, and Arieh Y. Shalev. 1996. Written disclosure in posttraumatic stress disorder: Is it beneficial for the patient? *Journal of Nervous and Mental Disease* 184, 8 (1996), 505–507. DOI:https://doi.org/10.1097/00005053-199608000-00009

[78] Oscar F. Gonçalves. 1994. Cognitive Narrative Psychotherapy: The Hermeneutic Construction of Alternative Meanings. *Journal of Cognitive Psychotherapy; New York* 8, 2 (1994), 105-108,110-125.

[79] Charles Goodwin. 1996. Paul Drew & John Heritage (eds.), Talk at Work: Interaction in institutional settings. (Studies in interactional sociolinguistics, 8.) Cambridge & New York: Cambridge University Press, 1992. Pp. x, 580. Hb $79.95, pb $29.95. *Language in Society* 25, 4 (December 1996), 616–620. DOI:https://doi.org/10.1017/S0047404500020844

[80] Benjamin M. Gorman and David R. Flatla. 2018. MirrorMirror: A Mobile Application to Improve Speechreading Acquisition. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, Montreal QC, Canada, 1–12. DOI:https://doi.org/10.1145/3173574.3173600

[81] Jonathan Gottschall. 2012. *The Storytelling Animal: How Stories Make Us Human*. Houghton Mifflin Harcourt.

[82] Leslie S. Greenberg, Jeanne C. Watson, Robert Elliot, and Arthur C. Bohart. 2001. Empathy. *Psychotherapy: Theory, Research, Practice, Training* 38, 4 (2001), 380–384. DOI:https://doi.org/10.1037/0033-3204.38.4.380

[83] M. A. Greenberg, C. B. Wortman, and A. A. Stone. 1996. Emotional expression and physical health: revising traumatic memories or fostering self-regulation? *J Pers Soc Psychol* 71, 3 (September 1996), 588–602. DOI:https://doi.org/10.1037//0022-3514.71.3.588

[84] Melanie A. Greenberg. 1995. Cognitive Processing of Traumas: The Role of Intrusive Thoughts and Reappraisals1. *Journal of Applied Social Psychology* 25, 14 (1995), 1262–1296. DOI:https://doi.org/10.1111/j.1559-1816.1995.tb02618.x

[85] Melanie A. Greenberg and Arthur A. Stone. 1992. Emotional disclosure about traumas and its relation to health: Effects of previous disclosure and trauma severity. *Journal of Personality and Social Psychology* 63, 1 (1992), 75–84. DOI:https://doi.org/10.1037/0022-3514.63.1.75

[86] Stephanie Greer, Danielle Ramo, Yin-Juei Chang, Michael Fu, Judith

Moskowitz, and Jana Haritatos. 2019. Use of the Chatbot "Vivibot" to Deliver Positive Psychology Skills and Promote Well-Being Among Young People After Cancer Treatment: Randomized Controlled Feasibility Trial. *JMIR mHealth and uHealth* 7, 10 (2019), e15018. DOI:https://doi.org/10.2196/15018

[87] Emory A. Griffin, Andrew Ledbetter, and Glenn Grayson Sparks. 2015. *A First Look at Communication Theory*. McGraw-Hill Education.

[88] James J. Gross (Ed.). 2007. *Handbook of emotion regulation*. Guilford Press, New York.

[89] James J. Gross and Robert W. Levenson. 1997. Hiding feelings: The acute effects of inhibiting negative and positive emotion. *Journal of Abnormal Psychology* 106, 1 (1997), 95–103. DOI:https://doi.org/10.1037/0021-843X.106.1.95

[90] Adam J. Guastella and Mark R. Dadds. 2006. Cognitive-Behavioral Models of Emotional Writing: A Validation Study. *Cogn Ther Res* 30, 3 (June 2006), 397–414. DOI:https://doi.org/10.1007/s10608-006-9045-6

[91] Adam J. Guastella and Mark R. Dadds. 2009. Sequential Growth in Cognitive-behavioral Emotion-processing: A Laboratory Study. *Cognitive Therapy and Research* 33, 4 (August 2009), 368–374. DOI:https://doi.org/10.1007/s10608-008-9199-5

[92] Jürgen Habermas. 1978. *Knowledge and Human Interests*. Heinemann Educational.

[93] Benjamin W. Hadden and C. Veronica Smith. 2019. I Gotta Say, Today Was a Good (and Meaningful) Day: Daily Meaning in Life as a Potential Basic Psychological Need. *J Happiness Stud* 20, 1 (January 2019), 185–202. DOI:https://doi.org/10.1007/s10902-017-9946-y

[94] Greg Haggerty, Margaret Blake, Melissa Naraine, Caleb Siefert, and Mark A. Blais. 2010. Construct validity of the Schwartz outcome scale-10: comparisons to interpersonal distress, adult attachment, alexithymia, the five-factor model, romantic relationship length and ratings of childhood memories. *Clin Psychol Psychother* 17, 1 (February 2010), 44–50. DOI:https://doi.org/10.1002/cpp.643

[95] Xiaodong He and Li Deng. 2017. Deep Learning for Image-to-Text Generation: A Technical Overview. *IEEE Signal Processing Magazine* 34, 6 (November 2017), 109–116. DOI:https://doi.org/10.1109/MSP.2017.2741510

[96] M. Heerink, B. Krose, V. Evers, and B. Wielinga. 2009. Measuring acceptance of an assistive social robot: a suggested toolkit. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 528–533. DOI:https://doi.org/10.1109/ROMAN.2009.5326320

[97] Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS Spoken Language Systems Pilot Corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*. Retrieved May 3, 2020 from

https://www.aclweb.org/anthology/H90-1021

[98] J. Hendricks, D. Mooney, and C. Berry. 1996. A practical strategy approach to use of reflective practice in critical care nursing. *Intensive Crit Care Nurs* 12, 2 (April 1996), 97–101. DOI:https://doi.org/10.1016/s0964-3397(96)81042-1

[99] Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. 2008. Effects of self-disclosure and empathy in human-computer dialogue. In *2008 IEEE Spoken Language Technology Workshop*, 109–112. DOI:https://doi.org/10.1109/SLT.2008.4777852

[100] Clara E. Hill. 2005. Therapist techniques, client involvement, and the therapeutic relationship: Inextricably intertwined in the therapy process. *Psychotherapy: Theory, Research, Practice, Training* (January 2005). DOI:https://doi.org/10.1037/0033-3204.42.4.431

[101] Clara E. Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action, 3rd ed.* American Psychological Association, Washington, DC, US.

[102] Clara E. Hill, Janet E. Helms, Victoria Tichenor, Sharon B. Spiegel, Kevin E. O'Grady, and Elgin S. Perry. 1988. Effects of therapist response modes in brief psychotherapy. *Journal of Counseling Psychology* 35, 3 (July 1988), 222–233. DOI:https://doi.org/10.1037/0022-0167.35.3.222

[103] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* 29, 6 (November 2012), 82–97. DOI:https://doi.org/10.1109/MSP.2012.2205597

[104] J Gregory Hixon and William B Swann. When Does Introspection Bear Fruit? Self-Reflection, Self-Insight, and Interpersonal Choices. 9.

[105] Annabell Ho, Jeff Hancock, and Adam S Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *J Commun* 68, 4 (August 2018), 712–733. DOI:https://doi.org/10.1093/joc/jqy026

[106] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. 2006. SenseCam: a retrospective memory aid. In *Proceedings of the 8th international conference on Ubiquitous Computing* (UbiComp'06), Springer-Verlag, Orange County, CA, 177–193. DOI:https://doi.org/10.1007/11853565_11

[107] K. Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers* 12, 4 (February 2000), 409–426. DOI:https://doi.org/10.1016/S0953-5438(99)00006-5

[108] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*, ACM Press, Pittsburgh, Pennsylvania, United States, 159–166. DOI:https://doi.org/10.1145/302979.303030

[109] Steven Houben, Connie Golsteijn, Sarah Gallacher, Rose Johnson, Saskia Bakker, Nicolai Marquardt, Licia Capra, and Yvonne Rogers. 2016. Physikit: Data Engagement Through Physical Ambient Visualizations in the Home. 1608–1619. DOI:https://doi.org/10.1145/2858036.2858059

[110] Jon Houck. 2008. Motivational Interviewing Skill Code (MISC) 2.1. Retrieved January 28, 2019 from https://casaa.unm.edu/download/misc.pdf

[111] Thomas K. Houston, Lisa A. Cooper, and Daniel E. Ford. 2002. Internet Support Groups for Depression: A 1-Year Prospective Cohort Study. *AJP* 159, 12 (December 2002), 2062–2068. DOI:https://doi.org/10.1176/appi.ajp.159.12.2062

[112] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (CIKM '13), Association for Computing Machinery, San Francisco, California, USA, 2333–2338. DOI:https://doi.org/10.1145/2505515.2505665

[113] Veronika Huta and Richard Ryan. 2010. Pursuing pleasure or virtue: The differential and overlapping well-being benefits of hedonic and eudaimonic motives. *Journal of Happiness Studies* 11, (October 2010), 735–762.

[114] Michael Ignelzi. 2000. Meaning-Making in the Learning and Teaching Process. *New Directions for Teaching and Learning* 2000, 82 (2000), 5–14. DOI:https://doi.org/10.1002/tl.8201

[115] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. 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 (2018), e12106. DOI:https://doi.org/10.2196/12106

[116] Ellen Isaacs, Artie Konrad, Alan Walendowski, Thomas Lennig, Victoria Hollis, and Steve Whittaker. 2013. Echoes from the Past: How Technology Mediated Reflection Improves Well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), ACM, New York, NY, USA, 1071–1080. DOI:https://doi.org/10.1145/2470654.2466137

[117] Anthony Jameson. 2002. Adaptive interfaces and agents. In *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. L. Erlbaum Associates Inc., USA, 305–330.

[118] Ronnie Janoff-Bulman. 2010. *Shattered Assumptions*. Simon and Schuster.

[119] Jin-pyo Hong and Jung-won Cha. 2009. Retrieving Important Sentences in Korean using TextRank Algorithm. *Communications of the Korean Institute of Information Scientists and Engineers* 36, 1C (June 2009), 311–314.

[120] Stephen Joseph and P. Linley. 2005. Positive Adjustment to Threatening Events: An Organismic Valuing Theory of Growth Through Adversity. *Review of General Psychology* 9, (September 2005). DOI:https://doi.org/10.1037/1089-2680.9.3.262

[121] Marije Kanis and Willem Paul Brinkman. 2007. What do people like? the design of a mobile tool to harness and share positive thoughts. In *Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!* (ECCE '07), Association for Computing Machinery, London, United Kingdom, 191–198. DOI:https://doi.org/10.1145/1362550.1362589

[122] Katy Kaplan, Mark S. Salzer, Phyllis Solomon, Eugene Brusilovskiy, and Pamela Cousounis. 2011. Internet peer support for individuals with psychiatric disabilities: A randomized controlled trial. *Social Science & Medicine* 72, 1 (January 2011), 54–62. DOI:https://doi.org/10.1016/j.socscimed.2010.09.037

[123] Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. 3128–3137. Retrieved May 3, 2020 from https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.html

[124] Jody Koenig Kellas, Haley Kranstuber Horstman, Erin K. Willer, and Kristen Carr. 2015. The Benefits and Risks of Telling and Listening to Stories of Difficulty Over Time: Experimentally Testing the Expressive Writing Paradigm in the Context of Interpersonal Communication Between Friends. *Health Communication* 30, 9 (September 2015), 843–858. DOI:https://doi.org/10.1080/10410236.2013.850017

[125] Dennis J. King and Janice Miner Holden. 1998. Disclosure of Trauma and Psychosomatic Health: An Interview With James W. Pennebaker. *Journal of Counseling & Development* 76, 3 (1998), 358–363. DOI:https://doi.org/10.1002/j.1556-6676.1998.tb02552.x

[126] Howard Kirschenbaum. 2004. Carl Rogers's Life and Work: An Assessment on the 100Th Anniversary of His Birth. *Journal of Counseling and Development* 82, 1 (Winter 2004), 116.

[127] Alexandra Kitson, Thecla Schiphorst, and Bernhard E. Riecke. 2018. Are You Dreaming? A Phenomenological Study on Understanding Lucid Dreams as a Tool for Introspection in Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, Montreal QC, Canada, 1–12. DOI:https://doi.org/10.1145/3173574.3173917

[128] Jacqueline D. Kloss and Stephen A. Lisman. 2002. An exposure-based examination of the effects of written emotional disclosure. *British Journal of Health Psychology* 7, 1 (2002), 31–46. DOI:https://doi.org/10.1348/135910702169349

[129] Stacey H. Kovac and Lillian M. Range. 2000. Writing Projects: Lessening Undergraduates' Unique Suicidal Bereavement. *Suicide and Life-*

*Threatening Behavior* 30, 1 (2000), 50–60. DOI:https://doi.org/10.1111/j.1943-278X.2000.tb01064.x

[130] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger (eds.). Curran Associates, Inc., 1097–1105. Retrieved May 3, 2020 from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[131] Paul Kwon and Megan L. Olson. 2007. Rumination and depressive symptoms: Moderating role of defense style immaturity. *Personality and Individual Differences* 43, 4 (September 2007), 715–724. DOI:https://doi.org/10.1016/j.paid.2007.01.012

[132] Brian M. Landry. 2008. Storytelling with digital photographs: supporting the practice, understanding the benefit. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '08), Association for Computing Machinery, Florence, Italy, 2657–2660. DOI:https://doi.org/10.1145/1358628.1358738

[133] N Landsteiner. 2005. Elizabot. Retrieved August 1, 2018 from https://www.masswerk.at/elizabot/

[134] Sophie Landwehr Sydow, Jakob Tholander, and Martin Jonsson. 2017. "It's a Bomb!" -- Material Literacy and Narratives of Making. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), Association for Computing Machinery, Denver, Colorado, USA, 121–132. DOI:https://doi.org/10.1145/3025453.3025529

[135] Jaron Lanier. 1996. My Problem with Agents. *Wired*. Retrieved May 18, 2020 from https://www.wired.com/1996/11/myprob/

[136] Richard S. Lazarus and Susan Folkman. 1984. *Stress, appraisal, and coping* (1st Edition ed.). Springer Publishing Company, New York.

[137] Andrew LeCompte. 2000. *Creating harmonious relationships: A practical guide to the power of true empathy* (1st Edition ed.). Atlantic Books, Portsmouth, N.H.

[138] Reeva Lederman, Greg Wadley, John Gleeson, Sarah Bendall, and Mario Álvarez-Jiménez. 2014. Moderated online social therapy: Designing and evaluating technology for mental health. *ACM Trans. Comput.-Hum. Interact.* 21, 1 (February 2014), 5:1–5:26. DOI:https://doi.org/10.1145/2513179

[139] Reineke Lengelle, Tom Luken, and Frans Meijers. 2016. Is self-reflection dangerous? Preventing rumination in career learning. *Australian Journal of Career Development* 25, 3 (October 2016), 99–109. DOI:https://doi.org/10.1177/1038416216670675

[140] Stephen J. Lepore, Pablo Fernandez-Berrocal, Jennifer Ragan, and Natalia Ramos. 2004. It's not that bad: Social challenges to emotional disclosure enhance adjustment to stress. *Anxiety, Stress, & Coping* 17, 4 (December 2004), 341–361. DOI:https://doi.org/10.1080/10615800412331318625

[141] Stephen C. Levinson. 2007. Optimizing person reference – perspectives from usage on Rossel Island. *Person Reference in Interaction: Linguistic, Cultural and Social Perspectives*, 29–72. DOI:https://doi.org/10.1017/CBO9780511486746.004

[142] Guo Li, Xiaomu Zhou, Tun Lu, Jiang Yang, and Ning Gu. 2016. SunForum: Understanding Depression in a Chinese Online Community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16), Association for Computing Machinery, San Francisco, California, USA, 515–526. DOI:https://doi.org/10.1145/2818048.2819994

[143] Ian Li, Anind K. Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11*, ACM Press, Beijing, China, 405. DOI:https://doi.org/10.1145/2030112.2030166

[144] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. *arXiv:1603.06155 [cs]* (June 2016). Retrieved May 3, 2020 from http://arxiv.org/abs/1603.06155

[145] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (April 2020), 1–15. DOI:https://doi.org/10.1145/3313831.3376590

[146] Henry Lieberman. Agents for the User Interface. 21.

[147] Jill Littrell. 1998. IS THE REEXPERIENCE OF PAINFUL EMOTION THERAPEUTIC? *Clinical Psychology Review* 18, 1 (January 1998), 71–102. DOI:https://doi.org/10.1016/S0272-7358(97)00046-9

[148] Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. 2017. A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions* 10, (December 2017), 39–46. DOI:https://doi.org/10.1016/j.invent.2017.10.002

[149] Sonja Lyubomirsky, Lorie Sousa, and Rene Dickerhoof. 2006. The costs and benefits of writing, talking, and thinking about life's triumphs and defeats. *Journal of Personality and Social Psychology* 90, 4 (2006), 692–708. DOI:https://doi.org/10.1037/0022-3514.90.4.692

[150] Alasdair Macdonald. 2011. *Solution-focused therapy: Theory, research and practice* (2nd Edition ed.). SAGE Publications Ltd, Los Angeles, Calif.

[151] Kyla A. Machell, Todd B. Kashdan, Jerome L. Short, and John B. Nezlek. 2015. Relationships Between Meaning in Life, Social and Achievement Events, and Positive and Negative Affect in Daily Life. *Journal of Personality* 83, 3 (2015), 287–298. DOI:https://doi.org/10.1111/jopy.12103

[152] Molly Magill, Jacques Gaume, Timothy R. Apodaca, Justin Walthers, Nadine R. Mastroleo, Brian Borsari, and Richard Longabaugh. 2014. The technical hypothesis of motivational interviewing: A meta-analysis of MI's key causal model. *Journal of Consulting and Clinical Psychology* 82, 6

(December 2014), 973–983. DOI:https://doi.org/10.1037/a0036833

[153] Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. 2008. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08), Association for Computing Machinery, Florence, Italy, 477–486. DOI:https://doi.org/10.1145/1357054.1357131

[154] Mark Matthews and Gavin Doherty. 2011. In the mood: engaging teenagers in psychotherapy using mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), Association for Computing Machinery, Vancouver, BC, Canada, 2947–2956. DOI:https://doi.org/10.1145/1978942.1979379

[155] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), Association for Computing Machinery, Austin, Texas, USA, 849–858. DOI:https://doi.org/10.1145/2207676.2208525

[156] Elisa D. Mekler and Kasper Hornbæk. 2019. A Framework for the Experience of Meaning in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, ACM Press, Glasgow, Scotland Uk, 1–15. DOI:https://doi.org/10.1145/3290605.3300455

[157] Belinda Melling and Terry Houguet-Pincham. 2011. Online peer support for individuals with depression: A summary of current research and future considerations. *Psychiatric Rehabilitation Journal* 34, 3 (2011), 252–254. DOI:https://doi.org/10.2975/34.3.2011.252.254

[158] Jack Mezirow. 1991. *Transformative Dimensions of Adult Learning*. Jossey-Bass, 350 Sansome Street, San Francisco, CA 94104-1310 ($27.

[159] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Barcelona, Spain, 404–411. Retrieved January 20, 2020 from https://www.aclweb.org/anthology/W04-3252

[160] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger (eds.). Curran Associates, Inc., 3111–3119. Retrieved May 3, 2020 from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[161] Murray G. Millar and Abraham Tesser. 1986. Effects of affective and cognitive focus on the attitude–behavior relation. *Journal of Personality and Social Psychology* 51, 2 (1986), 270–276. DOI:https://doi.org/10.1037/0022-3514.51.2.270

[162] William R. Miller and Theresa B. Moyers. 2017. Motivational interviewing and the clinical science of Carl Rogers. *J Consult Clin Psychol* 85, 8 (August 2017), 757–766. DOI:https://doi.org/10.1037/ccp0000179

[163] William R. Miller and Stephen Rollnick. 1991. *Motivational interviewing: Preparing people to change addictive behavior*. The Guilford Press, New York.

[164] William R. Miller and Stephen Rollnick. 2002. *Motivational interviewing: preparing people for change* (2nd Edition ed.). Guilford Press, New York.

[165] William R Miller and Stephen Rollnick. 2013. *Motivational interviewing: Helping people change* (3rd Edition ed.). Guilford Press, New York, NY, USA.

[166] William R. Miller and Gary S. Rose. 2009. Toward a theory of motivational interviewing. *American Psychologist* 64, 6 (September 2009), 527–537. DOI:https://doi.org/10.1037/a0016830

[167] Regina Miranda and Susan Nolen-Hoeksema. 2007. Brooding and reflection: Rumination predicts suicidal ideation at 1-year follow-up in a community sample. *Behaviour Research and Therapy* 45, 12 (December 2007), 3088–3095. DOI:https://doi.org/10.1016/j.brat.2007.07.015

[168] Terence R. Mitchell, Leigh Thompson, Erika Peterson, and Randy Cronk. 1997. Temporal Adjustments in the Evaluation of Events: The "Rosy View." *Journal of Experimental Social Psychology* 33, 4 (July 1997), 421–448. DOI:https://doi.org/10.1006/jesp.1997.1333

[169] Ine Mols, Elise van den Hoven, and Berry Eggen. 2016. Technologies for Everyday Life Reflection: Illustrating a Design Space. In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction* (TEI '16), Association for Computing Machinery, Eindhoven, Netherlands, 53–61. DOI:https://doi.org/10.1145/2839462.2839466

[170] Jennifer A. Moon. 1999. *Reflection in Learning & Professional Development: Theory & Practice*. Stylus Publishing, P.

[171] Robert J. Moore. 2018. A Natural Conversation Framework for Conversational UX Design. In *Studies in Conversational UX Design*, Robert J. Moore, Margaret H. Szymanski, Raphael Arar and Guang-Jie Ren (eds.). Springer International Publishing, Cham, 181–204. DOI:https://doi.org/10.1007/978-3-319-95579-7_9

[172] Robert J. Moore and Raphael Arar. 2019. *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*. ACM Books, New York.

[173] Robert J. Moore, Raphael Arar, Guang-Jie Ren, and Margaret H. Szymanski. 2017. Conversational UX design. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*, ACM Press, Denver, Colorado, USA, 492–497. DOI:https://doi.org/10.1145/3027063.3027077

[174] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M

Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *J Med Internet Res* 20, 6 (June 2018). DOI:https://doi.org/10.2196/10148

[175] Robert R. Morris, Stephen M. Schueller, and Rosalind W. Picard. 2015. Efficacy of a Web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *J. Med. Internet Res.* 17, 3 (March 2015), e72. DOI:https://doi.org/10.2196/jmir.4167

[176] Theresa B. Moyers, Tim Martin, Jon M. Houck, Paulette J. Christopher, and J. Scott Tonigan. 2009. From in-session behaviors to drinking outcomes: A causal chain for motivational interviewing. *Journal of Consulting and Clinical Psychology* 77, 6 (2009), 1113–1124. DOI:https://doi.org/10.1037/a0017189

[177] Theresa B. Moyers and William R. Miller. 2013. Is Low Therapist Empathy Toxic? *Psychol Addict Behav* 27, 3 (September 2013), 878–884. DOI:https://doi.org/10.1037/a0030274

[178] Theresa B. Moyers, L.N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. 2016. The Motivational Interviewing Treatment Integrity code (MITI 4): Rationale, preliminary reliability and validity. *J Subst Abuse Treat* 65, (June 2016), 36–42. DOI:https://doi.org/10.1016/j.jsat.2016.01.001

[179] Maurice Mulvenna, Jennifer Boger, and Raymond Bond. 2017. Ethical by Design: A Manifesto. In *Proceedings of the European Conference on Cognitive Ergonomics 2017 - ECCE 2017*, ACM Press, Ume&#229;, Sweden, 51–54. DOI:https://doi.org/10.1145/3121283.3121300

[180] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '94), ACM, New York, NY, USA, 72–78. DOI:https://doi.org/10.1145/191666.191703

[181] Frédéric Nils and Bernard Rimé. 2012. Beyond the myth of venting: Social sharing modes determine the benefits of emotional disclosure. *European Journal of Social Psychology* 42, 6 (2012), 672–681. DOI:https://doi.org/10.1002/ejsp.1880

[182] Richard E. Nisbett and Timothy D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84, 3 (1977), 231–259. DOI:https://doi.org/10.1037/0033-295X.84.3.231

[183] Donald A. Norman. 1988. *The psychology of everyday things*. Basic Books, New York, NY, US.

[184] Donald A. Norman. 1994. How might people interact with agents. *Commun. ACM* 37, 7 (July 1994), 68–71. DOI:https://doi.org/10.1145/176789.176796

[185] Sally A. Norman, Mark A. Lumley, John A. Dooley, and Michael P. Diamond. 2004. For Whom Does It Work? Moderators of the Effects of Written Emotional Disclosure in a Randomized Trial Among Women With Chronic Pelvic Pain. *Psychosomatic Medicine* 66, 2 (April 2004), 174–183. DOI:https://doi.org/10.1097/01.psy.0000116979.77753.74

[186] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, Montreal QC, Canada, 1–13. DOI:https://doi.org/10.1145/3173574.3174223

[187] Kathleen O'Leary, Arpita Bhattacharya, Sean A. Munson, Jacob O. Wobbrock, and Wanda Pratt. 2017. Design Opportunities for Mental Health Peer Support Technologies. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), Association for Computing Machinery, Portland, Oregon, USA, 1470–1484. DOI:https://doi.org/10.1145/2998181.2998349

[188] Kathleen O'Leary, Stephen M. Schueller, Jacob O. Wobbrock, and Wanda Pratt. 2018. "Suddenly, we got to become therapists for each other": Designing Peer Support Chats for Mental Health. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, Montreal QC, Canada, 1–14. DOI:https://doi.org/10.1145/3173574.3173905

[189] Crystal Park. 2010. Making Sense of the Meaning Literature: An Integrative Review of Meaning Making and Its Effects on Adjustment to Stressful Life Events. *Psychological bulletin* 136, (March 2010), 257–301. DOI:https://doi.org/10.1037/a0018301

[190] Crystal L. Park and Carol Joyce Blumberg. 2002. Disclosing Trauma Through Writing: Testing the Meaning-Making Hypothesis. *Cognitive Therapy and Research* 26, 5 (October 2002), 597–616. DOI:https://doi.org/10.1023/A:1020353109229

[191] Andrea Grimes Parker. 2014. Reflection-through-performance: personal implications of documenting health behaviors for the collective. *Personal Ubiquitous Comput.* 18, 7 (October 2014), 1737–1752. DOI:https://doi.org/10.1007/s00779-014-0780-5

[192] Jessica A. Pater, Oliver L. Haimson, Nazanin Andalibi, and Elizabeth D. Mynatt. 2016. "Hunger Hurts but Starving Works": Characterizing the Presentation of Eating Disorders Online. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16), Association for Computing Machinery, San Francisco, California, USA, 1185–1200. DOI:https://doi.org/10.1145/2818048.2820030

[193] Rakesh Patibanda, Florian "Floyd" Mueller, Matevz Leskovsek, and Jonathan Duckworth. 2017. Life Tree: Understanding the Design of Breathing Exercise Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (CHI PLAY '17), Association for Computing Machinery, Amsterdam, The Netherlands, 19–31. DOI:https://doi.org/10.1145/3116595.3116621

[194] S. Tejaswi Peesapati, Victoria Schwanda, Johnathon Schultz, Matt

Lepage, So-yae Jeong, and Dan Cosley. 2010. Pensieve: supporting everyday reminiscence. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, ACM Press, Atlanta, Georgia, USA, 2027. DOI:https://doi.org/10.1145/1753326.1753635

[195] J. W. Pennebaker. 2000. Telling stories: the health benefits of narrative. *Lit Med* 19, 1 (2000), 3–18. DOI:https://doi.org/10.1353/lm.2000.0011

[196] J. W. Pennebaker and S. K. Beall. 1986. Confronting a traumatic event: toward an understanding of inhibition and disease. *J Abnorm Psychol* 95, 3 (August 1986), 274–281. DOI:https://doi.org/10.1037//0021-843x.95.3.274

[197] J. W. Pennebaker, M. Colder, and L. K. Sharp. 1990. Accelerating the coping process. *J Pers Soc Psychol* 58, 3 (March 1990), 528–537. DOI:https://doi.org/10.1037//0022-3514.58.3.528

[198] J. W. Pennebaker, J. K. Kiecolt-Glaser, and R. Glaser. 1988. Disclosure of traumas and immune function: health implications for psychotherapy. *J Consult Clin Psychol* 56, 2 (April 1988), 239–245. DOI:https://doi.org/10.1037//0022-006x.56.2.239

[199] J. W. Pennebaker, T. J. Mayne, and M. E. Francis. 1997. Linguistic predictors of adaptive bereavement. *J Pers Soc Psychol* 72, 4 (April 1997), 863–871. DOI:https://doi.org/10.1037//0022-3514.72.4.863

[200] J. W. Pennebaker and J. D. Seagal. 1999. Forming a story: the health benefits of narrative. *J Clin Psychol* 55, 10 (October 1999), 1243–1254. DOI:https://doi.org/10.1002/(SICI)1097-4679(199910)55:10<1243::AID-JCLP6>3.0.CO;2-N

[201] James Pennebaker. 2002. What our words can say about us: Toward a broader language psychology. 15, (January 2002), 8–9.

[202] James W. Pennebaker. 1985. Traumatic experience and psychosomatic disease: Exploring the roles of behavioural inhibition, obsession, and confiding. DOI:https://doi.org/10.1037/h0080025

[203] James W. Pennebaker. 1993. Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour Research and Therapy* 31, 6 (July 1993), 539–548. DOI:https://doi.org/10.1016/0005-7967(93)90105-4

[204] James W. Pennebaker. 1997. Writing About Emotional Experiences as a Therapeutic Process. *Psychol Sci* 8, 3 (May 1997), 162–166. DOI:https://doi.org/10.1111/j.1467-9280.1997.tb00403.x

[205] James W. Pennebaker and Cindy K. Chung. 2011. *Expressive Writing: Connections to Physical and Mental Health*. Oxford University Press. DOI:https://doi.org/10.1093/oxfordhb/9780195342819.013.0018

[206] Christopher Peterson, Nansook Park, and Martin E. P. Seligman. 2005. Orientations to happiness and life satisfaction: the full life versus the empty life. *J Happiness Stud* 6, 1 (March 2005), 25–41. DOI:https://doi.org/10.1007/s10902-004-1278-z

[207] Keith J. Petrie, Roger J. Booth, James W. Pennebaker, Kathryn P. Davison, and Mark G. Thomas. 1995. Disclosure of trauma and immune

response to a hepatitis B vaccination program. *Journal of Consulting and Clinical Psychology* 63, 5 (1995), 787–792. DOI:https://doi.org/10.1037/0022-006X.63.5.787

[208] Ria Poole, Daniel Smith, and Sharon Simpson. 2015. How Patients Contribute to an Online Psychoeducation Forum for Bipolar Disorder: A Virtual Participant Observation Study. *JMIR Mental Health* 2, 3 (2015), e21. DOI:https://doi.org/10.2196/mental.4123

[209] Benjamin Poppinga, Stefan Oehmcke, Wilko Heuten, and Susanne Boll. 2013. Storyteller: in-situ reflection on study experiences. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services* (MobileHCI '13), Association for Computing Machinery, Munich, Germany, 472–475. DOI:https://doi.org/10.1145/2493190.2494655

[210] Jennifer Porter. 2017. Why You Should Make Time for Self-Reflection (Even If You Hate Doing It). *Harvard Business Review*. Retrieved February 13, 2020 from https://hbr.org/2017/03/why-you-should-make-time-for-self-reflection-even-if-you-hate-doing-it

[211] Neil Postman and Charles Weingartner. 1969. *Teaching as a subversive activity*. New York, Delacorte Press. Retrieved June 23, 2020 from http://archive.org/details/teachingassubver00post

[212] John Powell and Aileen Clarke. 2007. Investigating Internet Use by Mental Health Service Users: Interview Study. *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems* (2007), 1112.

[213] P. J. Price. 1990. Evaluation of Spoken Language Systems: the ATIS Domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*. Retrieved May 3, 2020 from https://www.aclweb.org/anthology/H90-1020

[214] Alison M. Radcliffe, Mark A. Lumley, Jessica Kendall, Jennifer K. Stevenson, and Joyce Beltran. 2010. Written Emotional Disclosure: Testing Whether Social Disclosure Matters. *J Soc Clin Psychol* 26, 3 (May 2010), 362–384. DOI:https://doi.org/10.1521/jscp.2007.26.3.362

[215] Tristine Rainer. 1978. *The new diary : how to use a journal for self-guidance and expanded creativity*. Los Angeles : J.P. Tarcher ; New York : Distributed by St. Martin's Press. Retrieved April 19, 2020 from http://archive.org/details/newdiary00tris

[216] Nairán Ramírez-Esparza and James W. Pennebaker. 2006. Do good stories produce good health?: Exploring words, language, and culture. *Narrative Inquiry* 16, 1 (January 2006), 211–219. DOI:https://doi.org/10.1075/ni.16.1.26ram

[217] Juan Ramos. 2003. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the 1st Instructional Conference on Machine Learning*, 133–142.

[218] J. M. Richards, W. E. Beal, J. D. Seagal, and J. W. Pennebaker. 2000. Effects of disclosure of traumatic events on illness behavior among

psychiatric prison inmates. *J Abnorm Psychol* 109, 1 (February 2000), 156–160. DOI:https://doi.org/10.1037//0021-843x.109.1.156

[219] Bernard Rimé. 2009. Emotion Elicits the Social Sharing of Emotion: Theory and Empirical Review. *Emotion Review* 1, 1 (January 2009), 60–85. DOI:https://doi.org/10.1177/1754073908097189

[220] Bernard Rimé, Batja Mesquita, Stefano Boca, and Pierre Philippot. 1991. Beyond the emotional event: Six studies on the social sharing of emotion. *Cognition and Emotion* 5, (January 1991), 435–465. DOI:https://doi.org/10.1080/02699939108411052

[221] C. R. Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *J Consult Psychol* 21, 2 (April 1957), 95–103.

[222] John Rolfe. 1998. *Expanding Nursing Knowledge: Understanding and Researching your own Practice* (2 edition ed.). Butterworth-Heinemann, Oxford ; Boston.

[223] Stephen Rollnick and Jeff Allison. 2003. Motivational Interviewing. In *The Essential Handbook of Treatment and Prevention of Alcohol Problems* (1 edition), Nick Heather and Tim Stockwell (eds.). Wiley, Chichester, West Sussex, England ; Hoboken, NJ.

[224] Stephanie S. Rude, Kacey Little Maestas, and Kristin Neff. 2007. Paying attention to distress: What's wrong with rumination? *Cognition and Emotion* 21, 4 (June 2007), 843–864. DOI:https://doi.org/10.1080/02699930601056732

[225] Richard M. Ryan and Edward L. Deci. 2001. On Happiness and Human Potentials: A Review of Research on Hedonic and Eudaimonic Well-Being. *Annual Review of Psychology* 52, 1 (2001), 141–166. DOI:https://doi.org/10.1146/annurev.psych.52.1.141

[226] Carol Ryff and Burton Singer. 2008. Know Thyself and Become What You Are: A Eudaimonic Approach to Psychological Well-Being. *Journal of Happiness Studies* 9, (February 2008), 13–39. DOI:https://doi.org/10.1007/s10902-006-9019-0

[227] Harvey Sacks. 1985. Notes on methodology. *Structures of Social Action*, 21–27. DOI:https://doi.org/10.1017/CBO9780511665868.005

[228] Harvey Sacks and Emanuel A. Schegloff. 1979. Two preferences in the organization of reference to persons in conversation and their interaction. DOI:https://doi.org/10.1017/CBO9780511486746.003

[229] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 4 (1974), 696–735. DOI:https://doi.org/10.2307/412243

[230] Jessica M. Sales, Natalie A. Merrill, and Robyn Fivush. 2013. Does making meaning make it better? Narrative meaning-making and well-being in at-risk African-American adolescent females. *Memory* 21, 1 (January 2013), 97–110. DOI:https://doi.org/10.1080/09658211.2012.706614

[231] Ruhi Sarikaya. 2017. The Technology Behind Personal Digital Assistants: An overview of the system architecture and key components. *IEEE Signal*

*Processing Magazine* 34, 1 (January 2017), 67–81. DOI:https://doi.org/10.1109/MSP.2016.2617341

[232] Corina Sas and Alan Dix. 2009. Designing for reflection on experience. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '09), Association for Computing Machinery, Boston, MA, USA, 4741–4744. DOI:https://doi.org/10.1145/1520340.1520730

[233] Emanuel Schegloff. 2007. Sequence Organization in Interaction: A Primer in Conversation Analysis. *Sequence Organization in Interaction: A Primer in Conversation Analysis I* 1, (January 2007), 1–300. DOI:https://doi.org/10.1017/CBO9780511791208

[234] Emanuel A. Schegloff. 1968. Sequencing in conversational openings1. *American Anthropologist* 70, 6 (December 1968), 1075–1095. DOI:https://doi.org/10.1525/aa.1968.70.6.02a00030

[235] Emanuel Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53, (June 1977), 361–382. DOI:https://doi.org/10.2307/413107

[236] D. Schoen. 1983. The Reflective Practitioner. In *Basic Books. New York*, CUMINCAD. Retrieved May 3, 2020 from http://papers.cumincad.org/cgi-bin/works/Show?9e1a

[237] D. Schoen. 1987. Educating the Reflective Practitioner. In *San Francisco: Josey-Bass Publishers*, CUMINCAD. Retrieved May 3, 2020 from http://papers.cumincad.org/cgi-bin/works/Show?54c7

[238] Beate Schrank, Ingrid Sibitz, Annemarie Unger, and Michaela Amering. 2010. How Patients With Schizophrenia Use the Internet: Qualitative Study. *Journal of Medical Internet Research* 12, 5 (2010), e70. DOI:https://doi.org/10.2196/jmir.1550

[239] Mari Sengoku, Hiroaki Murata, Takanobu Kawahara, Kaori Imamura, and Kazuyuki Nakagome. 2010. Does daily Naikan therapy maintain the efficacy of intensive Naikan therapy against depression? *Psychiatry and Clinical Neurosciences* 64, 1 (2010), 44–51. DOI:https://doi.org/10.1111/j.1440-1819.2009.02049.x

[240] Stuart M. Shieber. 1994. Lessons from a Restricted Turing Test. *arXiv:cmp-lg/9404002* (April 1994). Retrieved May 3, 2020 from http://arxiv.org/abs/cmp-lg/9404002

[241] Rebecca M. Shingleton and Tibor P. Palfai. 2016. Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research. *Patient Education and Counseling* 99, 1 (January 2016), 17–35. DOI:https://doi.org/10.1016/j.pec.2015.08.005

[242] Ben Shneiderman. 1982. The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology* 1, 3 (1982), 237–256. DOI:https://doi.org/10.1080/01449298208914450

[243] Ben Shneiderman. 1997. Direct manipulation for comprehensible, predictable and controllable user interfaces. In *Proceedings of the 2nd*

*international conference on Intelligent user interfaces - IUI '97*, ACM Press, Orlando, Florida, United States, 33–39. DOI:https://doi.org/10.1145/238218.238281

[244] Heung-yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers Inf Technol Electronic Eng* 19, 1 (January 2018), 10–26. DOI:https://doi.org/10.1631/FITEE.1700826

[245] Denise M. Sloan and Brian P. Marx. 2004. A Closer Examination of the Structured Written Disclosure Procedure. *Journal of Consulting and Clinical Psychology* 72, 2 (April 2004), 165–175. DOI:https://doi.org/10.1037/0022-006X.72.2.165

[246] Denise M. Sloan, Brian P. Marx, Eva M. Epstein, and Jennifer L. Dobbs. 2008. Expressive writing buffers against maladaptive rumination. *Emotion* 8, 2 (April 2008), 302–306. DOI:https://doi.org/10.1037/1528-3542.8.2.302

[247] Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), Association for Computing Machinery, Denver, Colorado, USA, 2696–2707. DOI:https://doi.org/10.1145/3025453.3025516

[248] Joshua M. Smyth and James W. Pennebaker. 1999. *Sharing One's Story: Translating Emotional Experiences into Words as a Coping Tool*. Oxford University Press. Retrieved May 3, 2020 from https://www.oxfordclinicalpsych.com/view/10.1093/med:psych/9780195119343.001.0001/med-9780195119343-chapter-4

[249] Joshua M. Smyth, Arthur A. Stone, Adam Hurewitz, and Alan Kaell. 1999. Effects of Writing About Stressful Experiences on Symptom Reduction in Patients With Asthma or Rheumatoid Arthritis: A Randomized Trial. *JAMA* 281, 14 (April 1999), 1304–1309. DOI:https://doi.org/10.1001/jama.281.14.1304

[250] Joshua Smyth, Nicole True, and Joy Souto. 2001. Effects of Writing About Traumatic Experiences: The Necessity for Narrative Structuring. *Journal of Social and Clinical Psychology* 20, 2 (June 2001), 161–172. DOI:https://doi.org/10.1521/jscp.20.2.161.22266

[251] Anna Ståhl, Kristina Höök, Martin Svensson, Alex S. Taylor, and Marco Combetto. 2009. Experiencing the Affective Diary. *Personal Ubiquitous Comput.* 13, 5 (June 2009), 365–378. DOI:https://doi.org/10.1007/s00779-008-0202-7

[252] Annette Stanton, Sarah Kirk, Christine Cameron, and Sharon Danoff-Burg. 2000. Coping through emotional approach: Scale construction and validation. *Journal of Personality and Social Psychology* 78, (June 2000), 1150–1169. DOI:https://doi.org/10.1037/0022-3514.78.6.1150

[253] Annette L. Stanton, Sharon Danoff-Burg, Lisa A. Sworowski, Charlotte A. Collins, Ann D. Branstetter, Alicia Rodriguez-Hanley, Sarah B. Kirk, and

Jennifer L. Austenfeld. 2002. Randomized, Controlled Trial of Written Emotional Expression and Benefit Finding in Breast Cancer Patients. *JCO* 20, 20 (October 2002), 4160–4168. DOI:https://doi.org/10.1200/JCO.2002.08.521

[254] Michael F. Steger. 2012. Experiencing meaning in life: Optimal functioning at the nexus of well-being, psychopathology, and spirituality. In *The human quest for meaning: Theories, research, and applications, 2nd ed*. Routledge/Taylor & Francis Group, New York, NY, US, 165–184.

[255] Greg J. Stephens, Lauren J. Silbert, and Uri Hasson. 2010. Speaker–listener neural coupling underlies successful communication. *Proc Natl Acad Sci U S A* 107, 32 (August 2010), 14425–14430. DOI:https://doi.org/10.1073/pnas.1008662107

[256] Molly M. Stevens, Gregory D. Abowd, Khai N. Truong, and Florian Vollmer. 2003. Getting into the Living Memory Box: Family archives & holistic design. *Pers Ubiquit Comput* 7, 3 (July 2003), 210–216. DOI:https://doi.org/10.1007/s00779-003-0220-4

[257] Margaret Stroebe, Wolfgang Stroebe, Henk Schut, Emmanuelle Zech, and Jan van den Bout. 2002. Does disclosure of emotions facilitate recovery from bereavement? Evidence from two prospective studies. *Journal of Consulting and Clinical Psychology* 70, 1 (2002), 169–178. DOI:https://doi.org/10.1037/0022-006X.70.1.169

[258] Peter Suedfeld and James W. Pennebaker. 1997. Health Outcomes and Cognitive Aspects of Recalled Negative Life Events. *Psychosomatic Medicine* 59, 2 (April 1997), 172–177.

[259] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger (eds.). Curran Associates, Inc., 3104–3112. Retrieved May 3, 2020 from http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[260] Yoshimitsu Takahashi, Chiyoko Uchida, Koichi Miyaki, Michi Sakai, Takuro Shimbo, and Takeo Nakayama. 2009. Potential Benefits and Harms of a Peer Support Social Network Service on the Internet for People With Depressive Tendencies: Qualitative Content Analysis and Social Network Analysis. *Journal of Medical Internet Research* 11, 3 (2009), e29. DOI:https://doi.org/10.2196/jmir.1142

[261] Keisuke Takano and Yoshihiko Tanno. 2009. Self-rumination, self-reflection, and depression: Self-rumination counteracts the adaptive effect of self-reflection. *Behaviour Research and Therapy* 47, 3 (March 2009), 260–264. DOI:https://doi.org/10.1016/j.brat.2008.12.008

[262] Shelley E. Taylor. 1983. Adjustment to threatening events: A theory of cognitive adaptation. *American Psychologist* 38, 11 (1983), 1161–1173. DOI:https://doi.org/10.1037/0003-066X.38.11.1161

[263] Josephine Tchetagni, Roger Nkambou, and Jacqueline Bourdeau. 2007. Explicit Reflection in Prolog-Tutor. *International Journal of Artificial*

*Intelligence in Education* 17, 2 (January 2007), 169–215.

[264] Anja Thieme, Madeline Balaam, Jayne Wallace, David Coyle, and Siân Lindley. 2012. Designing wellbeing. In *Proceedings of the Designing Interactive Systems Conference* (DIS '12), Association for Computing Machinery, Newcastle Upon Tyne, United Kingdom, 789–790. DOI:https://doi.org/10.1145/2317956.2318075

[265] Anja Thieme, Jayne Wallace, Paula Johnson, John McCarthy, Siân Lindley, Peter Wright, Patrick Olivier, and Thomas D. Meyer. 2013. Design to promote mindfulness practice and sense of self for vulnerable women in secure hospital services. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), Association for Computing Machinery, Paris, France, 2647–2656. DOI:https://doi.org/10.1145/2470654.2481366

[266] Peggy A. Thoits. 1984. Coping, social support, and psychological outcomes: The central role of emotion. *Review of Personality & Social Psychology* 5, (1984), 219–238.

[267] P. D. Trapnell and J. D. Campbell. 1999. Private self-consciousness and the five-factor model of personality: distinguishing rumination from reflection. *J Pers Soc Psychol* 76, 2 (February 1999), 284–304. DOI:https://doi.org/10.1037//0022-3514.76.2.284

[268] Wendy Treynor, Richard Gonzalez, and Susan Nolen-Hoeksema. 2003. Rumination Reconsidered: A Psychometric Analysis. *Cognitive Therapy and Research* 27, 3 (June 2003), 247–259. DOI:https://doi.org/10.1023/A:1023910315561

[269] Alan M. Turing. 1950. Computing Machinery and Intelligence. *Parsing the Turing Test, ISBN 978-1-4020-9624-2. Springer Science+Business Media B.V., 2009, p. 23* (1950), 23. DOI:https://doi.org/10.1007/978-1-4020-6710-5_3

[270] Philip Ullrich and Susan Lutgendorf. 2002. Journaling about stressful events: Effects of cognitive processing and emotional expression. *Annals of behavioral medicine: a publication of the Society of Behavioral Medicine* 24, (February 2002), 244–50. DOI:https://doi.org/10.1207/S15324796ABM2403_10

[271] Paul Verhaeghen, Jutta Joorman, and Rodney Khan. 2005. Why we sing the blues: the relation between self-reflective rumination, mood, and creativity. *Emotion* 5, 2 (June 2005), 226–232. DOI:https://doi.org/10.1037/1528-3542.5.2.226

[272] Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv:1506.05869 [cs]* (July 2015). Retrieved May 3, 2020 from http://arxiv.org/abs/1506.05869

[273] Liisa Voutilainen, Anssi Peräkylä, and Johanna Ruusuvuori. 2010. Recognition and Interpretation: Responding to Emotional Experience in Psychotherapy. *Research on Language and Social Interaction* 43, 1 (February 2010), 85–107. DOI:https://doi.org/10.1080/08351810903474799

[274] Richard S. Wallace. 2009. The Anatomy of A.L.I.C.E. In *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, Robert Epstein, Gary Roberts and Grace Beber (eds.). Springer Netherlands, Dordrecht, 181–210. DOI:https://doi.org/10.1007/978-1-4020-6710-5_13

[275] Joseph Weizenbaum. 1966. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (January 1966), 36–45. DOI:https://doi.org/10.1145/365153.365168

[276] Richard M. Wenzlaff and Daniel M. Wegner. 2000. Thought Suppression. *Annual Review of Psychology* 51, 1 (2000), 59–91. DOI:https://doi.org/10.1146/annurev.psych.51.1.59

[277] Irene G. Wilkinson. 2019. In Praise of Empathy: The Glue that holds Caring Communities Together in a Fractured World. *Canadian Journal of Family and Youth / Le Journal Canadien de Famille et de la Jeunesse* 11, 1 (January 2019), 234–291. DOI:https://doi.org/10.29173/cjfy29415

[278] Rhonda M. Williams, Mary C. Davis, and Roger E. Millsap. 2002. Development of the cognitive processing of trauma scale. *Clinical Psychology & Psychotherapy* 9, 5 (2002), 349–360. DOI:https://doi.org/10.1002/cpp.343

[279] Timothy D Wilson and Dana S Dunn. 1986. Effects of introspection on attitude-behavior consistency: Analyzing reasons versus focusing on feelings. *Journal of Experimental Social Psychology* 22, 3 (May 1986), 249–263. DOI:https://doi.org/10.1016/0022-1031(86)90028-4

[280] Timothy D. Wilson, Dana S. Dunn, Dolores Kraft, and Douglas J. Lisle. 1989. Introspection, Attitude Change, and Attitude-Behavior Consistency: the Disruptive Effects of Explaining Why we Feel the Way we Do. In *Advances in Experimental Social Psychology*, Leonard Berkowitz (ed.). Academic Press, 287–343. DOI:https://doi.org/10.1016/S0065-2601(08)60311-1

[281] Morgan Worthy, Albert L. Gary, and Gay M. Kahn. 1969. Self-disclosure as an exchange process. *Journal of Personality and Social Psychology* 13, 1 (1969), 59–63. DOI:https://doi.org/10.1037/h0027990

[282] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. 2017. Achieving Human Parity in Conversational Speech Recognition. *arXiv:1610.05256 [cs, eess]* (February 2017). Retrieved May 3, 2020 from http://arxiv.org/abs/1610.05256

[283] Irvin D. Yalom. 1980. *Existential psychotherapy*. Basic Books, New York, NY, US.

[284] Kam-shing Yip. 2006. Self-reflection in Reflective Practice: A Note of Caution. *Br J Soc Work* 36, 5 (July 2006), 777–788. DOI:https://doi.org/10.1093/bjsw/bch323

[285] Emmanuelle Zech and Bernard Rimé. 2005. Is talking about an emotional experience helpful? effects on emotional recovery and perceived benefits. *Clinical Psychology & Psychotherapy* 12, 4 (2005), 270–287.

DOI:https://doi.org/10.1002/cpp.460

[286] Vibeke Zoffmann, Åsa Hörnsten, Solveig Storbækken, Marit Graue, Bodil Rasmussen, Astrid Wahl, and Marit Kirkevold. 2016. Translating person-centered care into practice: A comparative analysis of motivational interviewing, illness-integration support, and guided self-determination. *Patient Education and Counseling* 99, 3 (March 2016), 400–407. DOI:https://doi.org/10.1016/j.pec.2015.10.015

[287] 2018. Optimal Workshop. Retrieved September 6, 2018 from https://www.optimalworkshop.com/reframer

[288] Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. *Google AI Blog*. Retrieved May 27, 2020 from http://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

[289] Much more than a chatbot: China's Xiaoice mixes AI with emotions and wins over millions of fans. *Asia News Center*. Retrieved May 27, 2020 from https://news.microsoft.com/apac/features/much-more-than-a-chatbot-chinas-xiaoice-mixes-ai-with-emotions-and-wins-over-millions-of-fans/

[290] IEEE SA - The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Retrieved May 3, 2020 from https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[291] Humans in the Loop: The Design of Interactive AI Systems. *Stanford HAI*. Retrieved May 17, 2020 from https://hai.stanford.edu/blog/humans-loop-design-interactive-ai-systems

# 국문초록

최근 인공지능(Artificial Intelligence; AI) 기술은 우리 삶의 면면을 매우 빠르게 바꿔놓고 있다. 특히 애플의 시리(Siri)와 구글 어시스턴트(Google Assistant) 등 자연어 인터페이스(natural language interfaces)의 확장은 곧 인공지능 에이전트와의 '대화'가 인터랙션의 주요 수단이 될 것임을 능히 짐작케 한다. 실상 인공지능 에이전트는 실생활에서 콘텐츠 추천과 온라인 쇼핑 등 다양한 서비스를 제공하고 있지만, 이들의 대부분은 과업-지향적이다. 즉 인공지능은 우리의 삶을 편리하게 하지만, 과연 편안하게 할 수 있는가? 본 연구는 편하지만 편하지 않은 현대인을 위한 기술의 역할을 고민하는 데에서 출발한다.

자아성찰(self-reflection), 즉 자신에 대해 깊이 생각해 보는 활동은 자기인식과 자기이해를 도모하고 배움과 목표의식을 고취하는 등 분야를 막론하고 널리 연구 및 적용되어 왔다. 하지만 자아성찰의 가장 큰 어려움은 스스로 건설적인 성찰을 도모하기 힘들다는 것이다. 특히, 부정적인 감정적 경험에 대한 자아성찰은 종종 우울감과 불안을 동반한다. 극복이 힘든 경우 상담 또는 치료를 찾을 수 있지만, 사회적 낙인과 잣대의 부담감으로 꺼려지는 경우가 다수이다.

'성찰 디자인'(Reflection Design)은 인간-컴퓨터상호작용(HCI)의 오랜 화두로, 그동안 효과적인 성찰을 도울 수 있는 디자인 전략들이 다수 연구되어 왔지만 대부분 다양한 사용자 데이터 수집 전략을 통해 과거 회상 및 해석을 돕는 데 그쳤다. 최근 소위 '챗봇 상담사'가 등장하여 심리상담과 치료 분야에 적용되고 있지만, 이 또한 성찰을 돕기보다는 효율적인 처치 도구에 머무르고 있을 뿐이다. 즉 기술은 치료 수단이거나 성찰의 대상이 되지만, 그 과정에 개입하는 경우는 제한적이라고 할 수 있다.

이에 본 연구는 '성찰 동반자'로서 대화형 에이전트인 챗봇을 디자인할 것을 제안한다. 이 챗봇의 역할은 사용자의 부정적인 감정적 경험 또는 트라우마에 대해 이야기할 수 있도록 도울 뿐 아니라, 그 과정에서 반추를 통제하여 건설적인 내러티브를 이끌어 내는 가이드를 제공하는 것이다. 이러한 챗봇을 설계하기 위해, 선행 연구를 기반으로 사용자의 자기노출(user self-disclosure)과 챗봇 가이드(guidance)를 두 축으로 한 디자인 공간(design space)을 정의하였다. 그리고 자기노출과 가이드의 정도에 따른 네 가지 자아성찰 경험을 분류하였다: 자기노출과 가이드가 최소화된 '회상' 공간, 자기노출이 위주이고 가이드가 최소화된 '설명' 공간, 자기노출과 챗봇이 이끄는 가이드가 혼합된 '탐색' 공간, 가이드를 적극 개입시켜 자기노출을 높이는 '변화' 공간이 그것이다.

본 연구의 목표는 상술된 디자인 공간에서의 성찰 경험과 과정을 돕는 챗봇을 구현하고, 사용자 실험을 통해 성찰 경험과 디자인 전략에 대한 반응을 수집 및 분석함으로써 챗봇 기반의 자아 성찰 인터랙션을 새롭게 제시하고 이에 대한 실증적 근거를 마련하는 것이다. 현재까지 많은 성찰 기술은 '회상'에 집중되어 있기에, 나머지 세 공간에서의 성찰을 지원하는 보노봇과 기본형·반응형 일기봇을 디자인하였다. 또한, 사용자 평가를 바탕으로 도출한 연구결과를 통해 도래한 인간-인공지능 상호작용(human-AI interaction)의 맥락에서 성찰 동반자로서의 챗봇 기술이 갖는 의미와 역할을 탐구한다.

보노봇과 일기봇은 인간중심상담과 대화분석의 이론적 근거를 바탕으로 한 정서지능(emotional intelligence)과 절차지능(proecedural intelligence)을 핵심 축으로, 대화 흐름 제어(flow manager)와 발화 생성(response generator)을 핵심 모듈로 구현하였다. 먼저, 보노봇은 동기강화상담(motivational interviewing)을 기반으로 고민과 스트레스에 대한 내러티브를 이끌어내어, 이에 대한 해결을 위한 가이드 질문을 통해 '변화'를 위한 성찰을 돕는다. 챗봇의 구현을 위해, 동기강화상담의 네 단계 대화를 설정하고 각 단계를 구성할 수 있는 상담사 발화 행동을 관련문헌에서 수집 및 전처리 과정을 거쳐 스크립트화하였다.

또한, 사전 전처리된 문장이 맥락을 유지할 수 있는 대화에 쓰일 수 있도록, 대화의 주제는 대학원생의 어려움으로 한정하였다.

보노봇과의 대화가 사용자의 성찰에 미치는 영향과 이에 대한 인식을 탐색하기 위해 질적 연구방법을 사용하여 30명의 대학원생과 사용자 실험을 진행하였다. 실험결과, 사용자는 변화 대화를 유도할 수 있는 다양한 탐색 질문을 선호하였다. 또한, 사용자의 맥락에 정확히 들어맞는 질문과 피드백은 사용자를 더욱 적극적인 자기 노출로 이끌게 할 수 있음을 발견하였다. 그러나 챗봇이 마치 상담사처럼 대화를 이끌어갈 경우, 높아진 사용자의 기대 수준으로 인해 일부 사용자가 변화에 대한 동기를 표출하였음에도 불구하고 변화에 대한 자율성을 챗봇에 양도하려는 모습 또한 나타남을 분석하였다.

보노봇 연구를 바탕으로 일기봇은 챗봇 대신 사용자가 보다 적극적으로 성찰 내러티브를 전개할 수 있도록 디자인하였다. 일기봇은 트라우마에 대한 표현적 글쓰기를 지원하는 챗봇으로, 기본형 또는 반응형 대화를 제공한다. 기본형 대화는 트라우마에 대해 자유롭게 '설명'할 수 있는 대화 환경을 제공하고, 반응형 대화는 사용자가 작성한 내러티브에 대한 후속 인터랙션을 통해 과거의 경험을 '재탐색'하도록 하였다. 또한, 후속 인터랙션의 발화 행동은 다양한 상담치료에서 발췌하되 유저의 내러티브에서 추출한 감정어 및 인간관계 키워드를 활용하도록 하였다.

각 일기봇에 대한 반응을 비교·분석하기 위해, 챗봇 없이 도큐먼트에 표현적 글쓰기 활동만을 하는 대조군을 설정하고 30명의 사용자를 모집하여 각 조건에 랜덤으로 배정, 설문과 면담을 동반한 4일간의 글쓰기 실험을 진행하였다. 실험결과, 사용자는 일기봇과의 인터랙션을 통해 보이지 않는 가상의 청자를 상상함으로써 글쓰기를 대화 활동으로 인지하고 있음을 알 수 있었다. 특히, 반응형 대화의 후속 질문들은 사용자로 하여금 상황을 객관화하고 새로운 관점으로 생각해 볼 수 있는 효과를 거두었다. 반응형 대화에서 후속 인터랙션을 경험한 사용자는 일기봇의 인지된 즐거움과 사회성, 신뢰도와 재사용

의향에 대한 평가가 다른 두 조건에서보다 유의하게 높았다. 반면, 기본형 대화 참여자는 다른 두 조건에서보다 감정적 표현의 용이성과 글쓰기의 어려움을 각각 유의하게 높게, 그리고 낮게 평가하였다. 즉, 챗봇은 많은 인터랙션 없이도 청자의 역할을 수행할 수 있었지만, 후속 질문을 통한 인터랙션이 가능했던 반응형 대화는 더욱 적극적인 유저 참여(engagement)를 이끌어낼 수 있었다. 또한, 실험이 진행됨에 따라, 사용자가 반응형 일기봇의 알고리즘에 자신의 글쓰기 주제와 단어 선택 등을 맞게 바꾸어 가는 적응적(adaptive) 행동이 관찰되었다.

앞선 연구결과를 통해, 다양한 챗봇 디자인 전략을 바탕으로 사용자의 내러티브가 다르게 유도될 수 있으며, 따라서 서로 다른 유형의 성찰 경험을 이끌어낼 수 있음을 발견하였다. 또한, 자율적인 행위인 자아성찰이 기술과의 상호작용으로 호혜적 성질을 갖게 될 때 사용자의 자율성, 상호작용의 예측가능성과 디자인 투명성에서 발생할 수 있는 갈등관계(tensions)를 탐색하고 인공지능 에이전트의 알고리즘 어포던스(algorithmic affordances)를 논의하였다.

보이지 않는 챗봇 알고리즘에 의해 사용자의 성찰이 유도될 수 있다는 것은 기존의 인간-컴퓨터 상호작용에서 강조되는 사용자 제어와 디자인 투명성에서 전복을 초래하는 것처럼 보일 수 있으나, 상징적 상호작용(symbolic interaction)의 맥락에서 오히려 사용자가 알고리즘에 의해 지나간 과거에 대한 새로운 의미를 적극 탐색해나가는 과정이 될 수 있다. 본 연구는 이것을 새로운 디자인 메타포, 즉 '의미-만들기'(meaning-making)로 제안하고 알고리즘의 '넛지'(nudge)에 의한 사용자의 주관적 해석 경험(interpretive process)을 강조한다. 이것은 하나의 챗봇 알고리즘이라 할지라도 서로 다른 사용자의 다양한 성찰 경험을 유도해낼 수 있다는 것을 의미하며, 이러한 맥락에서 인공지능은 기존의 '블랙 박스'를 유지하면서도 사용자의 자율성을 보장할 수 있다.

본 연구는 우리와 협업하는 인공지능 챗봇 기술의 디자인에 대한 경험적 이해를 높이고, 이론을 기반으로 한 챗봇을 구현함으로써 디자인 전략에 대한 실증적 근거를 제시한다. 또한 자아 성찰 과정에 동행하는

동반자(companion)로서의 기술로 새로운 디자인 메타포를 제시함으로써 인간컴퓨터상호작용(HCI)의 이론적 확장에 기여하고, 사용자의 부정적 경험에 대한 의미 추구를 돕는 관계지향적 인공지능으로서 향후 현대인의 정신건강에 이바지할 수 있는 사회적, 산업적 의의를 갖는다.

**주요어:** 자아성찰; 챗봇 디자인; 정신건강; 인공지능; 인간컴퓨터상호작용

**학번:** 2015-31344