Trust and Perceived Control in Burnout Support Chatbots

MSc Thesis

Chadha Degachi

Trust and Perceived Control in Burnout Support Chatbots

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Chadha Degachi

Interactive Intelligence Research Group
Department of Intelligent Systems
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
`www.ewi.tudelft.nl`

© 2022 Chadha Degachi.

Trust and Perceived Control in Burnout Support Chatbots

Author: Chadha Degachi
Student id: 5189241

Abstract

Increased levels of user control and feedback incorporation in learning systems is commonly cited as good AI development practice. However, the evidence as to the exact effect of perceived control over trust in these systems is mixed. This study investigates the relationship between different dimensions of trust and perceived control in postgraduate student burnout support chatbots. We present an in-between subject controlled experiment using simulated therapy-goal learning to study the effects of goal editing and feedback incorporation on perceived agent benevolence and competence. Our results showed that perceived control was moderately positively correlated with benevolence ($r = 0.448, BF_{10} = 7.150$), and weakly correlated with competence, and general trust.

Thesis Committee:

Chair:	Dr. W. P. Brinkman, Faculty EEMCS, TU Delft
University supervisor:	Dr. M. Tielman, Faculty EEMCS, TU Delft
Committee Member:	Dr. H. Hung, Faculty EEMCS, TU Delft

Preface

I would like to extend my thanks to my supervision team; Dr. M. Tielman, Pei-Yu Chen, and Mohammed Al-Owayyed, for their support and feedback throughout this project, to Drs. W. P. Brinkman and H. Hung for their contributions as thesis committee, and to the TU Delft student psychology team for their role as expert consultants on this project. I am further grateful for the participants who have contributed to this study, both through their input and through their recruitment of fellow participants. Lastly, I would like to thank all those who have contributed to this project through brainstorming, proofreading, and debugging.

Chadha Degachi
Delft, the Netherlands
June 27, 2022

Contents

Preface **iii**

Contents **v**

List of Figures **ix**

1 Introduction **1**

1.1 Research Questions 3

2 Theoretical Background **5**

2.1 Burnout 5

2.2 Goals & Mobile Behavioural Change Support Systems 7

2.3 Language in Therapy and Chatbots 8

2.4 Human Centred Design 10

3 Related Work **13**

3.1 Chatbots in Mental Healthcare 13

3.2 Goal Solicitation 14

3.3 Trust & Perceived Control 15

4 Prototype Development **19**

4.1 Requirements Seeking 19

4.2 Expert Reviews 20

4.3 User Trials 21

4.4 Design Iterations 30

5 Implementation **31**

5.1 Intervention Design 32

6 Experiment Design **37**

6.1 Hypotheses 37

6.2 Participants . . . 37
6.3 Measures . . . 38
6.4 Procedures . . . 39

7 Results **43**
7.1 Manipulation Check . . . 43
7.2 Perceived Control and Trust . . . 46
7.3 User Feedback . . . 54
7.4 Discussion . . . 56
7.5 Threats to Validity . . . 59

8 Conclusions **61**
8.1 Contributions . . . 61
8.2 Limitations . . . 62
8.3 Future work . . . 63
8.4 Chatbots in Burnout Support . . . 64
8.5 Trust and Perceived Control . . . 65

Bibliography **67**

A List of Burnout Exercises **87**
A.1 CBT . . . 87
A.2 MAV . . . 88

B Requirement Gathering **89**
B.1 Psychological Background . . . 89
B.2 Therapy Barriers . . . 89
B.3 eHealth . . . 89

C First Expert Review **91**
C.1 Introduction . . . 91
C.2 Burnout Interventions . . . 91
C.3 Goal Solicitation & Alignment . . . 92
C.4 Prototype Design . . . 92

D Application Screens (First Prototype) **93**

E Mock Conversations **101**
E.1 The Casual vs The Formal Agent . . . 101
E.2 The Introverted vs the Extroverted Agent . . . 101

F Design Study Survey **105**
F.1 Design . . . 105
F.2 Language . . . 106
F.3 Demographic . . . 110

G Design Study Codebook **111**

H Pilot Screenshots **117**

I Pilot Survey **121**
I.1 Prototype 121
I.2 Demographic 122

J Pilot Codebook **123**

K Functional Requirements **127**

L OLBI-S Abbreviated **129**

M Goal - OLBI Mapping **131**

N Pilot Dialogue Flow **133**

O Pretest Survey **135**
O.1 Demographic 136
O.2 Attitude Towards AI 136

P Post Test Survey **137**
P.1 Goals 137
P.2 Perceived Control 137
P.3 Trust 137
P.4 Secondary Factors 138
P.5 User Experience 138

Q Experiment Codebook **139**

List of Figures

2.1 Our Human Centred Design Process 10

4.1 Example Screen: Intervention Plan Proposal 24
4.2 Features Ranked The Most Important 25
4.3 Features Ranked The Least Important 26
4.4 Requested Features 26
4.5 Predicted Closeness Across Four Linguistic Styles. 27
4.6 Keywords Participants Associated with Prototype 28
4.7 Perceived Chatbot Personality 28
4.8 Perceived Closeness to Chatbot 29
4.9 Properties of Humanness 29

5.1 Dialogue Flow Part 1 34
5.2 Dialogue Flow Part 2 35

6.1 Distribution of Educational Background of Sample Population 39
6.2 Experiment Procedures 41

7.1 Prior vs Posterior Distribution of Perceived Control By Experimental Group 44
7.2 Bayes Factor in T-test across Several Priors: Perceived Control By Experimental Group 45
7.3 Raincloud Plots of Perceived Control Means by Experimental Group . . 45
7.4 Bayesian Pearson Correlation Robustness Check: Perceived Control & Benevolence 47
7.5 Raincloud Plots of Benevolence grouped by Experimental Group 47
7.6 Bayesian Pearson Correlations: Perceived Control & Benevolence 49
7.7 Bayesian Pearson Correlations Robustness Check: Perceived Control & Benevolence 50
7.8 Raincloud Plots of Benevolence grouped by Perceived Control 50
7.9 Bayesian Pearson Correlation: Competence & Trust 54
7.10 Feedback on perceptions of model accuracy 56

7.11 Feedback on trust influences in system interaction 57

8.1 Google Cloud Platform VM Instance Memory Use 63

D.1 Sign Up 93
D.2 Privacy and Data Use 94
D.3 Avatar Selection 94
D.4 Avatar Selection Alt 95
D.5 Home Screen 95
D.6 OBLI Survey 96
D.7 Chatbot Interface 96
D.8 Plan Proposal 97
D.9 Goal Modification (A) 97
D.10 Goal Modification (B) 98
D.11 Goal Modification (B2) 98
D.12 Goal Modification (B3) 99
D.13 Goal Modification Success 99
D.14 Goal Overview 100
D.15 Push Notification 100

F.1 Conversation 1 - Alex (White) and M (Blue) 106
F.2 Conversation 2 - Alex (White) and M (Blue) 107
F.3 Conversation 1 - Alex (White) and M (Blue) 108
F.4 Conversation 2 - Alex (White) and M (Blue) 108
F.5 Classical Navigation Goal Editing Flow 109
F.6 Inclusion-of-the-Self-in-the-Other (Closeness) 109

H.1 Prototype Chat Screen: Mood recording (Screenshots by Author) 118
H.2 Prototype Chat Screen: Breathing Exercise (Screenshots by Author) . . 119

N.1 Pilot Study Dialogue Flow 133

Chapter 1

Introduction

Recent research has seen a rising trend in mental health issues, such as depression and anxiety, among university students [93, 113, 128]. Specifically, burnout, generally defined as a psychological syndrome in which an individual suffers emotional exhaustion, depersonalization, and reduced personal accomplishment [107], is seeing much wider prevalence [129, 12, 160]. Moreover, these trends have been compounded by the outbreak of the COVID-19 virus and subsequent pandemic. This pandemic not only further distressed and exhausted students [158, 102, 30], but also created an unprecedented level of demand for fully remote mental health services [14]. Barriers to accessing in-person therapy and mental health services, including finance, location, demand outstripping available resources, and social stigma [17, 14], had already been well documented previous to the pandemic but have now come into much sharper relief against the backdrop of a global crisis. At TU Delft, we can see these barriers implicitly in the discrepancy, discussed by the counselling team, between the reported numbers of students with mental health complaints, and the number of students who contact the counselling service. Among the counselling team, this discrepancy has been attributed to fear of judgement, an aversion to burdening others, and the lack of awareness of the university's mental health services and their effectiveness. There are many reasons to be concerned by these trends. Academic burnout and poor mental health has been associated with poor academic performance [162] and higher dropout rates [105], but most importantly, there is a link between prolonged mental stress and adverse physical and psychological side effects in students [91].

eHealth solutions, both mobile and web-based, have appeared in recent years with the hope of bridging this gap between individuals and mental health services, reaching a population not currently served by in-person support. Unlike traditional approaches, eHealth solutions offer greater scalability, constant availability, lower cost, anonymity, tailored content, and resource equity [170, 20]. In theory, such solutions can serve those populations that may be otherwise hesitant, unwilling, or unable to seek traditional care. In a survey of four US-based psychiatric clinics, 70% of participants were interested in using a mobile 'app' for monitoring their mental health [166]. However, there are two main challenges in this area. First, and this is common to many behaviour-change-support systems [35, 23], is attrition, i.e., the loss of user engagement over time [126]. Second, is the lack of medical validation and evidence-based solutions. Many products can be found in application stores listed under depression, anxiety, sleep problems, and other such terms, but few are those that have also been validated in clinical trials [170]. Nonetheless, of those systems that have undergone such studies, many do, in fact, show empirical evidence towards efficacy [20].

One possible method of increasing user engagement with these systems is the use of chatbots [126]. Chatbots simulate human conversation by way of natural language text, and more recently, images and emoticons. They have a long history in the field of mental health care, where they can mimic the support of healthcare professionals, thereby fostering a stronger sense of accountability in users, and promoting better engagement [126] in accordance with Mohr's 'Model of Supportive Accountability' [115]. Modern consumer mental health chatbots include Woebot[1], Wysa[2], Mina [3], Replika [4], and Youper[5], of which only Wysa has been used to address burnout [177]. Mina, currently in beta testing, intends to include burnout specific modules in the future.

In this study, we prototype a chatbot which tackles burnout in post-graduate student populations. We focus on human-centred development processes of chatbots and increasing trust in conversational agents. Trust is a key aspect not only of human-chatbot interactions, but also of human-therapist interactions [139, 89]. It has received much interest as AI and Machine Learning (ML) technologies have grown both in prowess and prevalence but not transparency [53], creating contentious relationships between user and technology where the dimensions of trust are disrupted [45]. In the field of human-computer interaction, the most common, and most relevant, dimensions are benevolence (the confidence that one's wellbeing or values will be protected by the trustee [61, 52]) and competence (confidence in the trustee's skill level [61, 52]). Improvement along those dimensions creates more positive and meaningful interactions with technology [112], while maintaining steady relationships with intelligent systems [171]. In relation to chatbots, lack of trust on the part of the user may be a large driver behind its low adoption rate [153, 180]. Trust further influences perceived ease of use and intention to use in potential adopters [47].

Many factors influence trust; explainability [53], brand reputation [40, 173], and even the user's own personality [117]. We choose to investigate trust as affected by degree of perceived control and feedback incorporation. The evidence as to the exact effect of control over an interaction with an intelligent system is mixed. In some cases, allowing users to correct mistakes made by the system was seen to improve user trust [54, 156, 157], but in others, the opposite is true [59]. Nonetheless, allow-

---

[1] woebothealth.com

[2] wysa.io

[3] minabot.ai

[4] replika.ai

[5] youper.ai

ing the user some degree of control over intelligent systems remains a recommended 'best-practice' in industry standards [50, 51, 7]. We choose user goals as the locus of exercising said control. Goal-setting and goal-alignment are vital activities when establishing trust in client-therapist relationships [60, 123], as well as when motivating and sustaining behavioural change in users [95], thus users may view this part of the system as more personally relevant to them.

1.1 Research Questions

Given this introduction, we propose the following research questions:

1. **RQ1**: How does allowing the user to edit the agent's learned model of their goals affect their perception of the agent's benevolence?
2. **RQ2**: How does allowing the user to edit the agent's learned model of their goals affect their perception of the agent's competence?

To answer these questions, we designed a prototype of a goal-soliciting empathetic chatbot. While the chatbot was presented to participants as intelligent and able to learn user goals from conversation, goal formulation was actually accomplished using a clinical burnout inventory.

In the remainder of this document, we discuss the theoretical and psychological background underpinning our design choices (Chapter 2) and similar works in the field of mental health chatbot and human-agent trust (Chapter 3). We detail our prototype development process in Chapters 4 & 5, and our experiment design in Chapter 6. Results, and our contributions, are discussed in Chapters 7 & 8, respectively.

Chapter 2

Theoretical Background

2.1 Burnout

Occupational 'burnout' emerged as a concept in the 1970s [42]. Initially, it was described in terms of gradual emotional depletion, loss of motivation, and reduced commitment in human-facing professional settings, such as care giving, psychiatry, and legal services [146]. Then, in 1981, Maslach and Jackson introduced what quickly became the golden standard model of burnout [106]. The model conceptualized burnout along three dimensions; the aforementioned emotional exhaustion, depersonalization (negative, cynical attitudes regarding one's clients [107]), and reduced personal accomplishment (a tendency to evaluate oneself negatively [107]). Studies also linked this emerging concept to higher job turnover, lower morale, and possibly poorer quality of care, as well as higher personal distress, physical exhaustion, increased use of alcohol and drugs, and family problems [107, 42].

Within a few years, the definition had expanded, becoming *"...a state of exhaustion in which one is cynical about the value of one's occupation and doubtful of one's capacity to perform"* [107] as empirical evidence showed burnout to be domain-independent [86]. Eventually, the term came to include students as well with researchers arguing that, with assignments, exams, and so on, students also perform a kind of work [145] and can also become exhausted and withdraw from their studies [144]. Research has since shown that burnout symptoms are common in all students regardless of their study subject, and can possibly follow them as they begin their career [138]. For PhD students in the Netherlands, an estimated 47% suffered from increased mental health complaints, almost 60% indicated their workload was too high, and 41.6% have considered quitting their program [129]. Across all students in higher education in the Netherlands, 51% suffered psychological complaints [30]. Similar trends are seen in other countries [159, 118].

While the Maslach burnout model continues to dominate research in this area, other measures for burnout have been proposed and verified, such as the Copenhagen Burnout Inventory [77] or the Stanford Professional Fulfilment Index [1]. In this study,

---

[1] wellmd.stanford.edu/about/model-external.html

we use the Oldenburg burnout model and associated inventory (OLBI) proposed in 2003 [28], specifically the student OLBI-S adapted and verified by Reis *et al.* (2014) [138]. In this model, the dimensions of burnout are: exhaustion (physical, affective, or cognitive) and disengagement (the distancing of oneself from their work and developing negative attitudes towards it) [28].

Predictors and drivers of burnout have been much researched. Of interest here are student dissatisfaction with study topic [44], lower perceptions of social support and leisure opportunities [44], higher perceptions of stress [172, 26], lower self-efficacy (perception of one's own capabilities) [134], and lack of community [110]. These factors also go on to erode students' resilience and exacerbate existing burnout [127]. While we focus on the internal drivers of burnout, as those are the ones we can best tackle, we would be remiss not to point out the important external factors that contribute to this condition as well. In the literature, two core drivers emerge; service demand outstripping worker resources, and employer-employee value conflict [146, 27].

2.1.1 Burnout Interventions

The stability of the burnout syndrome makes it unlikely that its symptoms would naturally reverse or lessen over time [85]. Strategies for addressing burnout are varied, and no one evidence-based regimen exists [74], though it is worth noting that studies have seen some success in longer term treatment plans [85], an approach which eHealth is particularly suited for. Here, these strategies were used to guide the design of the therapeutic exercises recommended by our chatbot.

Cognitive Behavioural Therapy (CBT) emerged as a psychiatric intervention in the 1980s [133]. It is centred around unlearning maladaptive patterns of thinking and acting, and acquiring better coping skills [133]. CBT-based approaches have been shown to alleviate symptoms of burnout [143, 74, 116]. Some specific CBT exercises, most suitable for a chatbot-based interface, are seen in Appendix A.

Mindfulness-, Acceptance-, and Value-Based Therapy (MAV) has also seen some success in reducing burnout [73]. Mindfulness is an attentive, non-judgemental experience of the present and the awareness that comes from such [70]. Acceptance is the experience of internal and external events as they are, without evaluation or avoidance [70]. Mindfulness and acceptance practices purport to reduce the power of a person's evaluative models, allowing them to function more flexibly and be more accepting of themselves and others [56]. Meanwhile, value-based actions help cement more long-lasting change in behaviour [56]. Some mindfulness and MAV exercises are shown in Appendix A.

Lifestyle can also heavily mediate the degree of burnout. Sleep quality is especially noted as a possible predictor of future burnout and a central feature of clinical burnout [31, 8]. Other behaviours such as physical exercise and engaging in social behaviour, especially when combined with positive reinforcement, can also have a restorative function for the individual. Like depression, burnout can limit a person's behavioural repertoire, and a *behavioural activation* approach (which relies partly on CBT theory) can help users expand it again [109, 3]. Encouraging exercise and similar ’wellness’ improving behaviour in users is, in fact, a common application of mobile technology [98, 97, 4], and some best practice recommendations already exist for achieving this goal [124, 111]. Moreover, behavioural activation approaches have been positively linked to user engagement in mobile systems [4].

2.2 Goals & Mobile Behavioural Change Support Systems

Behavioural Change Support Systems (BCSSs) are persuasive technologies that aim to influence user behaviour, especially in relation to their health. The underpinning design theories of which are based on social and cognitive psychology [124].

One such theory is Self-Efficacy, which explains an individual’s capacity for action in relation to their perception of their own competence. Theoretically, said capacity can be strengthened through social persuasion, altering negative emotion patterns, and vicarious experiences [101]. As mentioned, low self-efficacy is a well-known component of burnout [162]. Common design techniques in applications aiming to improve self-efficacy include gain-framed feedback messaging for fitness apps [90], and gamification for smoking cessation [135], can thus be transferred to our domain.

The other theory of relevance here is Goal Setting Theory, where conscious goals are acknowledged as a direct influence over one’s actions [95]. Over the years, empirical evidence has shown that goal difficulty and specificity are positively correlated with performance [95], largely due to goals’ capacity to act as action-directing, energizing, persistence-prolonging, arousal-raising motivators [95]. Goals have been used in behavioural change systems to improve application compliance [94], user depression [181], and ’information quality’ [11]. Goal-setting is linked to an improved sense of autonomy in users [169]. Lastly, goals are an important factor of person-therapist interaction. The concept of therapeutic alliance, which describes the working relationship of patient and therapist, is built around not only their rapport, but also their ability to agree on goals and goal-priority in treatment [60, 123]. This alliance can impact treatment outcome directly (improved relationship leading to stronger impact) or indirectly (enabling better engagement in treatment from the patient) [60].

It is also worth noting that self-efficacy and goal setting are tied [95]. Simply being assigned higher and more difficult goals to employees can improve their self-efficacy, as it is often perceived as an implicit ’vote of confidence’ in their ability [95]. Moreover, self-efficacy can mediate the effect of gain vs loss framed feedback on goal attainment and future goal setting [95]. Lastly, self-efficacy correlates significantly with goal commitment [95].

2.3 Language in Therapy and Chatbots

In most therapeutic practices, language and conversation is the tool through which the mental health of clients is improved. Thus, therapeutic conversation can exhibit unique characteristics not common elsewhere. Pawelczyk (2011) [125] identifies three norms of therapeutic social activities which necessitate said characteristics; transparency of meaning, self-disclosure, and communication of emotion.

Transparency of meaning relies on explicit interactional work to clarify the significance and frame of reference behind phrases used by the client and therapist, in a manner that often reads as overtly confrontational in other contexts. Such work includes probing questions (e.g., 'What do you mean by that?'), overt continuers (e.g., 'Keep going...'), and even moving non-verbal actions to verbal expressions (e.g., 'Why are you shaking your head?').

As for self-disclosure, it is characterized as the voluntary process in which a person reveals personal information, thoughts, opinions, or feelings to another [6]. Therapist views on self-disclosure vary both in theory, and in practice (where dyad gender configuration can affect therapist preference) [37]. Nonetheless, it is a key part of intimate discourse and interpersonal relation development [6]. Unlike participants in other forms of intimate discourse, therapists are actively involved in encouraging and facilitating client self-disclosure. [125] noted that significant moments of client self-disclosure were often surrounded by the discourse markers 'you know' and 'I don't know', indicating an appeal to be understood. In turn, therapists encouraged clients to resume exploration of these revelations with responses such as 'no, I don't know', 'what don't you know?', or possibly by repeating parts of client speech as questions. On the other hand, therapist self-disclosure, which fosters empathy and support, is often accomplished though mirroring (See Table 2.1)

Context	Self-Disclosure by Mirroring
Client	[...]. I think it sort of spills over into my need of trying to control things which are uncontrollable.
Therapist	Well. I have memories of training myself, with that voice inside, being on a bus going to school, I trained myself with that kind of voice.
Client	Yeah, and even coming here I packed almost everything [...].

Table 2.1: Mirroring by Drawing on Similar Personal Experiences [125]

Lastly, therapeutic conversation is one of the few contexts in which both negative and positive emotion construction is highly encouraged and expected [125]. In therapist-client interactions, therapists will often refocus fact-oriented event descriptions to more emotion-oriented ones using questions such as 'how do you feel about this?', or 'were you happy to hear that?'. Further, therapists will also assist clients in constructing less socially acceptable emotions, and express emotional support and empathy via emotive extension, emotive reaction, or validation (e.g., 'that's an understandable reaction').

2.3.1 Language As Personality

For chatbots, text, and occasionally images or emoticons, are the only means of communication with users and thus, the only medium though which emotion, empathy, listening skills, and personality can be conveyed. Personality is a surprisingly important aspect of chatbot's presentation, it adds stability to the way the bot is perceived, creates a more consistent user experience, and can even improve the over all user experience [155, 114].

In their effort to develop a user-based personality-adaption framework, Mairesse *et al.* (2010) [103] catalogued 67 psychologically-motivated adjustable parameters for personality variation. They then tagged a wide array of linguistic features with the Big-Five personality dimensions; extraversion, emotional stability, agreeableness, conscientiousness, and openness [69]. In this instance, we are most interested in extraversion, which describes a person's talkativeness, assertiveness, and energy [69].

The researchers used these parameters to automatically generate restaurant recommendations with target perceived personalities (see Table 2.2). Human evaluators were able to consistently identify said target personality, especially where extraversion was high. Human evaluators also rated generated utterances as moderately natural, with high extraversion utterances being most natural.

Context	Low Extraversion	High Extraversion
Chatbot	Chimichurri Grill isn't as bad as the others.	I am sure you would like Chimichurri Grill, you know. The food is kind of good, the food is tasty, it has nice servers, it's in Midtown West and it's a Latin American place. Its price is around 41 dollars, even if the atmosphere is poor.

Table 2.2: Generated Responses under Low vs High Extraversion [103]

Mairesse *et al.* do successfully point to informality as an aspect of extraverted linguistic behaviour, but the changes in conversational style affected by formality go beyond the few identified there. Liebrecht *et al.* (2020) [88], catalogued eight verbal and non-verbal cues which affected perceived chatbot formality. The researchers found informal communication styles resulted in higher perceived social presence, and thus better interaction quality and brand attitude.

Interestingly, Fadhil *et al.* (2018) [36], investigated the effect of emoticons specifically. There, participants scored chatbots with emoticons higher than plain text chatbots on enjoyment, confidence, and attitude, but only when conversations were on the topic of mental wellbeing, and not physical wellbeing.

2.4 Human Centred Design

Human- or User- Centred Design (UCD), is a theoretical design approach built around incorporating end-user's feedback and interests into iterative design processes [2]. In this study, we relied on the ISO standard 9241-210:2019 for interactive systems [65] to inform our prototype development. This means development was initiated by specifying the context of prototype use, then specifying user requirements, producing design solutions, and evaluating designs with primary (students), and secondary (university counsellors) stakeholders at each step. This process is visualized in Figure 2.1.

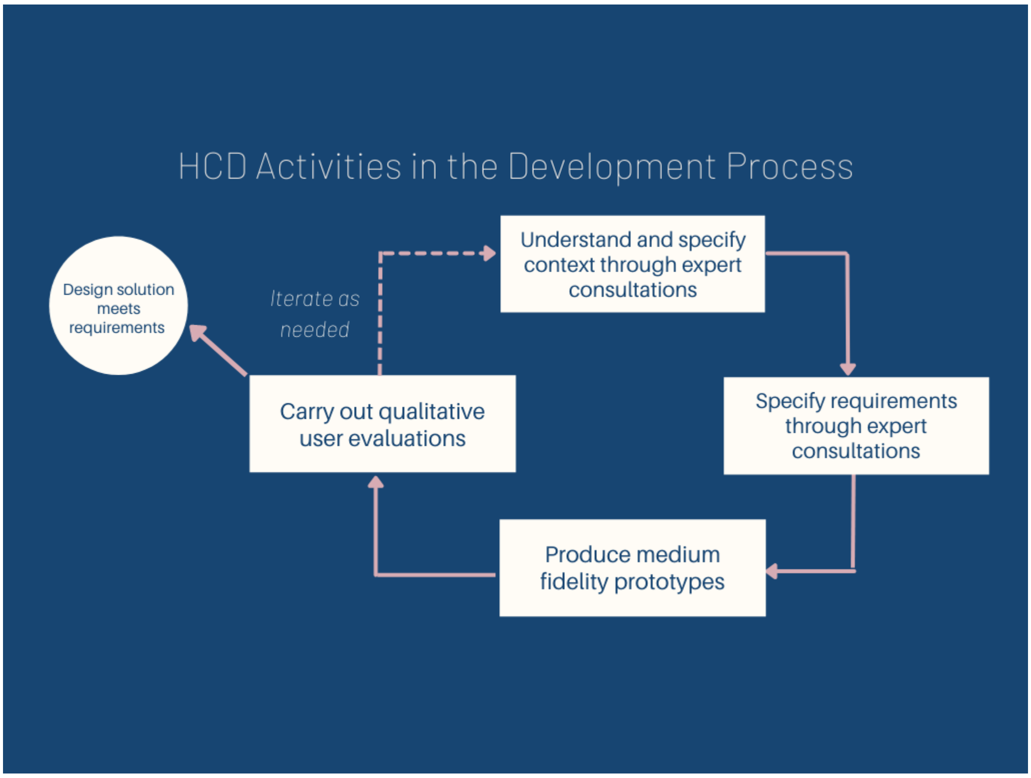

HCD Activities in the Development Process
Design solution meets requirements
Iterate as needed
Understand and specify context through expert consultations
Specify requirements through expert consultations
Produce medium fidelity prototypes
Carry out qualitative user evaluations

Figure 2.1: Our Human Centred Design Process

Context and requirements can be gathered in several ways, including performance measures, interviews, questionnaires, expert evaluation, and collaborative design [64]. They serve to define 1) target users and stakeholders, 2) their characteristics, 3) their goals and tasks, 4) their environments, 5) the intended context of use, and 6) user needs. In this process, requirements act as the measure against which evaluation is performed. Design evaluations are similarly conducted and are needed to 1) provide feedback about design strengths and weaknesses, 2) asses whether requirements have been met, and 3) establish baselines to allow for design comparison. They can be performed over prototypes of varying levels of fidelity, as well as final outcomes.

By and large, UCD is motivated by the desire to reduce product misuse and misalignment, as well as reducing the need for technical support during product use [96, 65]. By using UCD metrics, such as usability, we can use these design processes and feedback sessions as further trust-building exercises with our target user-base [176, 100]. Similarly, we know that communicability and aestheticism in online platform design also influence brand value perception [46], while usability is indirectly correlated with brand loyalty through its direct positive effect on brand trust and user satisfaction [80]. Given the sensitive subject matter of our design, it is integral to establish user trust early on. Moreover, by eliminating reputation and design as a source of mistrust, we can better focus on the issue of perceived control and feedback.

Chapter 3

Related Work

3.1 Chatbots in Mental Healthcare

Eliza, the first conversational agent, was developed in 1966 to emulate a Rogerian style of psychotherapist; encouraging users to talk about themselves and their problems by 'reflecting' their statements back at them [174]. Eliza used a concise set of decomposition-reassembly rule pairs triggered by keywords to gain minimal context, and give the impression of listening and understanding. Partly, Eliza also relied on the assumptions of the user, as well as the expected style of interaction with a Rogerian therapist, to maintain this image of intelligence [174].

Though less common in other domains, social chatbots like Eliza, meant to be virtual companions to their users [154], have become fairly common in wellness and mental health domains. In a scoping review of 53 studies on mental health chatbots [1], only 28.3% of surveyed studies used simple task-oriented chatbots. 32% used chatbots for therapy (as opposed to training or screening), of which 58.8% used CBT approaches, and 30.2% focused on tackling depression.

Wysa, described as an emotionally intelligent chatbot [63] with a focus on tackling depression, is one such social bot. Wysa uses CBT, behavioural reinforcement, mindfulness, and guided tools to help users with issues relating to anxiety, focus, sleep, conflicts, and other such topics [63]. Wysa's conversational style varies between task-oriented coaching and the more therapeutic, empathetic, ELIZA-style listening. In a mixed-method study of Wysa's effectiveness, researchers concluded that high-engagement users improved in self-reported mood significantly more than low-engagement, proxy-control-group, users [63]. Furthermore, 67.7% of the feedback collected indicated that Wysa was helpful and encouraging [63].

As well as emotion, disclosure is an important factor of relationship building in social chatbots. Lee *et al.* (2020) [83] investigated the possibility of promoting long-term deep self-disclosure in users. In our study, self-disclosure is a key part of our ability to extract user goals from conversation and establish therapeutic relationships. At baseline, chatbots already facilitate a higher-quality of self-disclosure from users than web surveys [71]. Over the three-week period of this study, chatbots which exhibited self-disclosure features, such as small talk with highly personal and emotional content, had a reciprocal effect on participants. The level at which the chatbot exhibited self-disclosure (none, low, high) matched the level at which the participant practised it. Researchers also found improved perceived intimacy, trust, and enjoyment over the study period in those participants working with self-disclosing bots. Such bots are also seen in Ravichander *et al.'s* work (2018) [137] where researchers, using a chatbot that engages with Amazon's Alexa users, found that instances of self-disclosure on the agent's part correlated positively with increased disclosure on the part of the user. They also found that users who chose to self-disclose early in their conversation with Alexa also went on to have significantly longer conversations with the bot overall.

In relation to burnout, one chatbot, Vicki (2021) [62], appears. Vicki, a rule-based chatbot with gamified story and feedback features, was created to encourage employees to complete workplace mental health assessments concerning depression, anxiety, and burnout. Viki's avatar, language, and conversational style were all iteratively moulded around user needs as inferred from focus groups and telephone interviews. Of the 120 participants of the study, 98 started their assessment form, and 77 completed it. The authors concluded that Viki seems to be 'highly engaging' and 'effective' with response rates comparable to face-to-face interviews.

3.2 Goal Solicitation

As discussed, conversation is the main tool of therapeutic practice, and this includes patient-goal solicitation and setting. In a qualitative analysis of client-goal-exploration styles [123], researchers found that explicit goal consensus was not common in therapist-client conversation, instead therapists worked to *"clarify the key characteristics of the change projects that the client wished to pursue"* [123]. This was accomplished by one of two action categories, the ensuring of relevant, meaningful, therapeutic activity, or the factoring of complexity and client resistance to change. In the first category, therapists would establish a focus, elucidate on expectations for change, and characterize practice as a means to an end. Patient and therapist both traded back and forth to 1) clarify problem dynamics (e.g., by verifying their impression of the other's utterance) and 2) explore the variety of changes with which resolution could be achieved (e.g., by prompting the imagination of different possible futures). The second category included actions such as pointing out ambivalence and contradictions in the client's meaning system, acknowledging unpleasant aspects of practice, and proposing objections to their own suggestions (e.g., 'this may be too time-intensive, and you are not obligated, but…').

To our knowledge, this technique of goal extraction from natural language conversation has not been discussed in machine learning publications. The closest studies attempt something similar in requirements engineering, where natural language is processed with regular expressions to discern stakeholder goals and preferences [5]. However, this technique is not dissimilar to intent classification, which is very common to chatbot development. In chatbot domains, intents and goals are sometimes interchangeable [57], referring to the objective which the user seeks through natural language utterances, for example to book a flight, or find out the weather [57]. This is a much more explicit and concrete understanding of goals than is usually seen in therapeutic practice, but some concepts are still relevant. Developing such complex systems is not always possible, however, or even necessary. In some studies, Wizard of Oz (WoZ) approaches are used instead [18]. In WoZ systems, systems are presented as autonomous learners, when, in fact, their behaviour is controlled by a human. This facilitates testing research questions while minimizing cost and development time [18]. Theoretically, it is also possible to achieve such advantages by simulating ML behaviour, not with human intervention, but with a programmatic solution to a simplified subset of the problem.

3.3 Trust & Perceived Control

There are a few different definitions of trust across the literature of Human-Computer Interaction, as well as Social Psychology. Interpersonal, i.e., human-human, trust is generally defined as *"a trustor's willingness to be vulnerable to a trustee's actions based on the expectation that the trustee will perform a particular action important to the trustor"* [108]. Perceived competence (confidence in the trustee's skill level [61]) and benevolence (the confidence that one's wellbeing or values will be protected by the trustee [61]) are the two most commonly used dimensions of trust, but others appear depending on domain. Integrity [178], understandability [58], and openness [61] are all such dimensions in interpersonal trust. In human-computer trust, factors such as reliability and utility [58] also emerge. Thus, we can define trust in interactive computer-based systems as *"[the belief] that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability"* [82].

Some outcomes linked to increased trust were already discussed in Chapter 1, but of further interest here are its effects on attrition and therapeutic effectiveness. Interpersonal trust between patient and doctor has been repeatedly found to affect said patients' adherence to their medication regimen [119, 34, 84]. Like goal-setting, trust is an important aspect of the therapist-patient therapeutic alliance which significantly mediates treatment outcome both offline [60], and online [39]. In fact, trust was highlighted as a key issue in the future development of psychiatric eHealth [120]. In human computer relationships, trust, like engagement, can combat attrition with its positive effect on continued use intention in users [22].

Like trust, perceived control also has a number of definitions. Where providing a level of control over ML- and AI-based features is recommended as best practice, it is defined as *"the ability to specify [...] preferences, make corrections when the system does not behave as [...] expect[ed], and give [...] opportunities to provide feedback"* [51]. In this study, we similarly define user control as *"the ability to affect or override machine-generated settings, models, or outcomes which directly affect the user* *experience of a system via user feedback"*. This definition derives from the field of interactive or human-in-the-loop ML (IML). In IML, users directly influence model behaviour, most commonly by labelling the data points presented to, or selected by, them. Data points are presented either with (coactive) or without (active) their corresponding predicted label. Alternative feedback incorporation methods also exist, such as giving users control over the model's feature space and their associated weights [21]. Largely, models which use IML can improve their performance over time by incorporating corrected or newly labelled data points into their training set [59]. They can further address the problem of 'concept drift' in which the user interests and the model's initial understanding of the user grow apart over time [59].

Trust in IML is not extensively studied, and studies that do address this issue occasionally contradict in their findings.

Gutzwiller *et al.* (2021) [54], Smith *et al.* (2018) [156], and Sollner *et al.* (2012) [157] investigated this relationship in unmanned vehicles, topic modelling, and recommender systems respectively, concluding that control via feedback incorporation was positively correlated with trust. In the first study, three levels of interactivity between evolutionary algorithm and user were compared; none, low, and high. Study participants with control over the systems were allowed to select which behaviour a given vehicle was to evolve further given their goal of searching outlined areas. Researchers found that participants not only trusted and preferred the IML regardless of interaction level, but were also able to correctly recognize them as such when compared to non-interactive systems. In [156], participants used an array of interactive topic modelling features (e.g., removing incorrect topics) to train an ML model, after which their experiences were collected via semi-structured interviews. After the interactive training session, participants felt the system was fairly controllable and trustworthy. In fact, researchers noted that participants seemed to overtrust the system, or lack confidence in their own judgement, leading them to feeling these lower levels of frustration when the system didn't respond to their modifications or performed poorly. In the last study, no learning or improvement was derived from allowing users to edit their machine-generated restaurant recommendations, however, the ability to simply override unsatisfactory recommendation outcomes was regarded as an important feature by participants. The control feature significantly affected both trust and intention to use the product, though it is difficult to determine to what extent exactly, as the effect was confounded by other design choices as well.

On the other hand, using a simulated face detection model, Honeycutt *et al.* (2020) [59] found correcting the system to be negatively correlated with trust and perception of accuracy, regardless of whether the model improved in accuracy after feedback.

[59] and [156] choose to select non-expert end-users as their participant pool, both contrasting their choice with the more common choice of developers or annotators as the ones interacting with the system in IML research. [59] points out that users themselves could wish for the ability to provide feedback on system outcomes by which they are affected [163, 173], and have in fact shown a higher willingness to

use imperfect systems when they were able to correct them [29]. We further note that users may also value systems they participated in the training of more highly [122]. Moreover, in human teams, feedback provision and incorporation improves the perceived fairness of team decisions, and improves trust in the decision-making process, while ignoring provided feedback decreases trust [75, 168]. For these reasons, we were also motivated to involve the real-world use case of student burnout in answering our research questions. Of course, comparison across these studies is somewhat hindered by the fact that they do not employ the same scale for measuring trust, and changes thereof.

An emerging framework aims to make trust more central to IML; Explanatory Interactive Machine Learning (XIL). XIL, proposed by Teso *et al.* (2019) [165] expands on existing IML techniques by also allowing users to correct the *explanation* associated with a given label as well the label itself. The primary motivation given was that *"the predictions were not explained to [the user]. So, why should users trust models learned interactively?"*. Indeed, both Teso and other researchers [48, 149] confirm this approach can improve performance, explanatory power, and, of course, trust in annotators. However, Ghai *et al.* (2021) also point out that domain knowledge, need for cognition (tendency towards complex cognitive activities), and experience with AI do significantly affect the annotators' relationship with, and reliance on, the ML model. These factors are likely to be further exasperated in an end-product user-base as is used in this study. Moreover, given the dearth of research on interactivity and control, and the already established strong effect of explainability on trust [53], we think it is important in this use case to avoid confounding the two variables before establishing a foundational understanding of the effects of the first.

Chapter 4

Prototype Development

4.1 Requirements Seeking

Establishing need and collecting requirements for a burnout support application among postgraduate student populations was accomplished by consulting with the university's counselling team. A set of questions was drawn up and semi-structured interviews were conducted with staff to answer them. See Appendix B for said questions. The consultation was conducted online and was not recorded, but extensive meeting notes were taken by the corresponding researcher. The aim of this meeting was discussing whether a burnout problem existed among the postgraduate student population, specifically in TU Delft, what barriers exist between students and the support of the counselling team, and whether experts found eHealth solutions to be a desirable support function in their practice.

The counselling team confirmed that burnout was prevalent among postgraduate students in the Netherlands, especially PhDs, citing the Promovendi Netwerk Nederland (PNN) 2020 survey [129] in their discussions. However, they also noted that the number of students reaching out to their services did not reflect this trend, implying a barrier between in-need students and mental health support services. The team discussed many issues which they saw as driving students' hesitancy in seeking help, including cultural values, shame, low motivation, fear of burdening others, fear of judgement, lacking awareness of the effectiveness of psychological support, and not feeling as though they struggle enough to be in need of support.

eHealth solutions were discussed largely in the context of support tools which were used in concert with face to face interventions. Woebot, a CBT-based wellbeing-support chatbot mentioned earlier in this document, and Healthyboel[1] were particularly cited. Healthyboel is a self-management tool offered by TU Delft for its students which equips them to deal with a variety of wellbeing issues (e.g., self-compassion, choice stress, etc.) using programs of video instructions and exercises. The counsellors had positive reviews of these systems, but mentioned that, at least in the case of Healthyboel, low adherence presented a common problem. Further-

---

[1] tudelft.nl/studenten/begeleiding/awareness-self-management

more, discussion arose regarding the difficulty of determining intervention success or effectiveness. Mental health indexes and measures were noted as sometimes failing to account for cases where students' symptoms had remained the same, or worsened, since they began therapy despite having become better equipped to deal with their problems. Instead, the team encouraged us to consider student satisfaction and emotional state as measures of success.

Overall, we were able to successfully establish our target user group, and learned some of their common characteristics and motivations. Furthermore, we were able to understand how eHealth solutions behaved within the context of professional therapeutic practice.

4.2 Expert Reviews

Expert opinion from the counselling team was again incorporated throughout the design process. The aim was to vet the list of therapeutic exercises collected through literature review with regard to their potential as part of our chatbot's tool set. Further, we discussed our experts' personal experience of the process of goal alignment in therapeutic practice, and how well that aligns with the literature's view of the topic. See Appendix C for the full survey.

In the same vein as the previous consultation, this meeting was carried out as a semi-structured, unrecorded interview, where the corresponding researcher was both interviewer and notetaker.

In this discussion, context was heavily emphasized. It was seen as key to establish, from the start of an intervention, (1) how external factors can lead to burnout, (2) to discuss the prevalence of the syndrome with the student, and (3) to highlight the role of our system within the larger picture; to address their emotional state in the here-and-now, not to solve overarching institutional issues. Using this approach, we planned to negotiate appropriate expectations with our users, while encouraging them to commit to the work of improving their emotional resilience and wellbeing. They moreover emphasized the importance of avoiding diagnosing users, and its associated stigma. Instead, they suggested relying on the OLBI [138] as a tool for locating problem areas (and possible goals) for students using the system.

Moreover, some trade-offs of the CBT-style expert-therapist vs the more person-centred, patient-as-expert (in their own life), approaches to therapeutic practice were discussed. Person-centred therapy, derived from Rogerian approaches, highlights congruence, unconditional positive regard, and empathetic understanding on the part of the therapist [140]. Meanwhile, CBT, which is more suited to tackle longer term issues in the patient's psyche, demands more upfront cognitive effort from patients. Our consulted experts noted that, in their experience, students with burnout responded better to connection-focused styles such as person-centred therapy. However, experts also noted that such decisions can vary depending on the therapist, their practice, their client, and the relationship they build over time. Experts noted that they are open to adapting their approach and communication-style over time to better suit different students' needs. A few therapeutic exercises were selected in this discussion as particularly useful for an introductory session tackling burnout symptoms. First, was breathing manipulation, which is detailed in Appendix A. Second, was the stress thermometer, which acts as a visual indicator that helps users pinpoint their current stress level, and identify its associated stressors and coping mechanisms [41]. Lastly, was the value-compass, which is another visual tool that encourages value inventorying, as seen in Appendix A.

These consultations motivated several design decisions throughout the product lifetime, largely on the conceptual and theoretical level. Said decisions are detailed in Section 4.4.

4.3 User Trials

As discussed in our review of the Human Centred Design process (Chapter 2), the next step of prototype development, after establishing requirements, is collecting user input on initial designs.

4.3.1 Participants

Design Study

Initially, ten participants were recruited from the pool of university students, including local and international students, with the aim of reflecting the target audience as closely as possible. Two further participants were recruited after initial data analysis, for a total of 12 participants. Participants were not screened for clinical burnout, but assumed to experience some level of stress in their day-to-day life as students, which would lead them to be interested in our prototype. The inclusion criteria demanded participants be over 18, currently attending a university, or a recent graduate, and comfortably proficient with the English Language.

Our participant pool had the characteristics seen in Table 4.1.

Table 4.1: Design Study: Characteristics of the Surveyed Population

Gender	83 % Female and 17% Male
Age Group	8% 18-21, 50% 21-24, 25% 24-27, and 17% 27-30.
Status	8% Undergraduate, 83% Postgraduate, and 8% Recent Graduate.
Occupation	50% Art, Design, and Entertainment, 17% Mechanical, Electrical, and Aerospace Engineering, 8% Law, Politics, and History, 25% Computer Science and Mathematics.
Geographic Area of Origin	25% Southern Europe, 8% South Asia, 25% Western Europe, 25% East Asia, 8% North America, and 8% Southeast Asia.

Had Previous Experience with Wellbeing Apps	42% Yes, and 58% No.
Technical Skills	17% Far above average, 33% Somewhat above average, 33% Average, and 17% Somewhat below average.

Of the participants that described themselves as having used wellbeing apps in the past, 8% described the experience in a negative light, while 33% described it positively. Moreover, 8% of participants explicitly discussed expense as a drawback of these apps, and 8% discussed disengagement and low motivation as a part of their experience.

Pilot Study

The second study recruited from the same pool of possible participants, with a total of 5 participants recruited. One participant was dropped since they did not answer survey questions. Our participant pool had the characteristics seen in Table 4.2.

Table 4.2: Pilot Study: Characteristics of the Surveyed Population

Gender	75 % Female and 25% Male.
Age Group	75% 21-24, 25% 24-27.
Status	50% Postgraduate, 50% Recent Graduate.
Occupation	25% Art, Design, and Entertainment, 75% Mechanical, Electrical, and Aerospace Engineering.
Geographic Area of Origin	50% Southern Europe, 25% South Asia, 25% East Asia.
Had Previous Experience with Wellbeing Apps	100% No.

4.3.2 Measures

Design Study

The initial user study, being qualitative in nature, did not rely heavily on questionnaire tools. We did, however, measure perceived closeness when discussing chatbot personality and language styles. Closeness describes a sense of social intimacy and being together with another. It is a part of therapeutic alliance and may be linked to user-agent rapport and user adherence [76]. We measured this using the Inclusion-of-the-Other-in-the-Self (IOS) [9] scale, a single-item, pictorial, measure. The IOS depicts seven sets of circles of varying degrees of overlap, correlating to degrees of relationship intimacy.

Pilot Study

This study made use of the IOS as well, and further employed the humanness scale. Humanness or Mindless Anthropomorphism refers to the unconscious attribution of human characteristics to artificial agents, such as likeability or friendliness. This metric was of relevance to us since some research has positively linked perception of humanness to willingness to establish common ground with agents [25], human-agent rapport [99], and human-agent trust [40]. Here, it was measured using the scale proposed by Kim *et al.* (2012) [72]. This scale consists of a 4-item, 10-point (*very poorly* to *very well*) survey, in which a participant rates how well each adjective (*likeable, sociable, friendly, and personal*) fits the chatbot.

Qualitative short-answer questions were used to investigate concepts of trust and usability in both studies.

4.3.3 Procedures

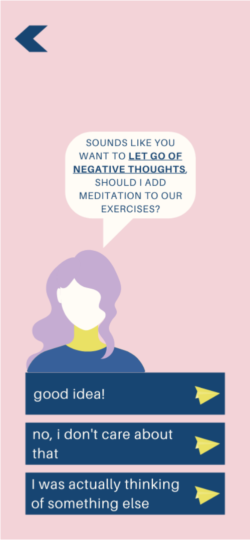
Design Study

This design trial had two parts. In the first part, participants were asked to study our medium-fidelity horizontal prototypes (Figure 4.1 & Appendix D) by performing a simple task of editing a goal, and answering questions about usability and design. While low-fidelity prototypes are used more often at this stage of development, we elected to begin collecting feedback at a later point in the design process, allowing us the time to conduct the pilot high-fidelity vertical prototype study with a functional chatbot. In the second part, participants studied mock conversations simulating human-chatbot conversation (See Appendix E), and answered questions around the themes of perceived agent humanness, benevolence (trust), competence (trust), and closeness. See Appendix F for the full list of questions.

The aim of this study was to answer the following (sub)research questions:

1. **sRQ1**: What features are most and least important to students when using a burnout support system?
2. **sRQ2**: Would students prefer more human-presenting or non-human presenting avatars for our chatbot?
3. **sRQ3**: Would students prefer more chatbot-centred or navigation-centred goal-editing workflows?
4. **sRQ4**: What personality would students prefer from our chatbot?

Data Analysis: Collected qualitative feedback was analysed using grounded theory approaches [49]. An initial codebook was developed using open-coding on a third of collected responses ($n = 3$), then refined and finalized using the remaining responses ($n = 7$). Further participants were recruited to ensure information saturation ($n = 2$), meaning that no new codes or concepts appeared in the analysis of



SOUNDS LIKE YOU WANT TO LET GO OF NEGATIVE THOUGHTS. SHOULD I ADD MEDITATION TO OUR EXERCISES?

good idea!
no, i don't care about that
I was actually thinking of something else

Figure 4.1: Example Screen: Intervention Plan Proposal

the new responses [49]. Selective coding was used to cluster and categorize granular codes. The codes, their descriptions, and categories can be seen in Appendix G.

Pilot Study

In our second study, we presented participants with a working chatbot prototype that they were able to converse with. The chatbot implemented a subset of the feature set it would eventually possess; it collected the user's mood, it encouraged the user to 'vent' about the circumstance, and suggested therapeutic exercises. It did not broach the topic of goals or goal editing. See Figure Appendix N for a visualization of the pilot dialogue flow. After speaking with the bot regarding what was on their mind that day, participants filled out a short survey once again centred around the themes of usability, humanness, benevolence (trust), competence (trust), and closeness. See Appendix I for the full list of questions.

The aim of this study was to answer the following (sub)research questions:

5. **sRQ5**: How was the chatbot personality perceived by users?
6. **sRQ6**: What attitude did users exhibit towards the chatbot?

This study also functioned as a small pilot, allowing us to find and correct issues with the system, as pointed out by users, before moving on to experiments.

Data Analysis: Study two, was analysed using the same grounded theory approach so that $n = 2$ responses were used to draft the initial codebook, $n = 2$ further responses were used to finalize it. The second study's codebook can be seen in Appendix J.

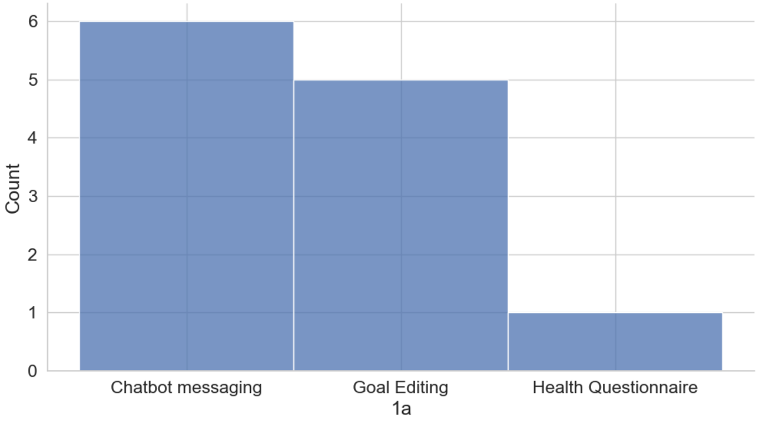


Figure 4.2: Features Ranked The Most Important

Demographic data were also collected including age range, gender, geographic area of origin, and technical skill level for both studies. All data were collected via an online survey hosted on Qualtrics [2]. Study procedures were approved the TU Delft Human Research Ethics Committee.

In both studies, quantitive feedback and demographic information were only aggregated, except for the IOS data which is processed using density estimators for higher interpretability.

4.3.4 Results

Design Study

Sub-research question **sRQ1** was addressed in a number of our survey's questions. First, we looked at most vs least important features as ranked by the participants. In Figure 4.2 we can see 'Chatbot Messaging' was ranked as the most important for 50% of participants. Participants who chose this feature cited a perceived great capacity for support, as well as higher feelings of engagement and social presence. On the contrary, 41.67% of participants ranked 'Avatar Selection' as the least important (See Figure 4.3). Participants saw two concerns related to this feature, 1) was the aesthetic design of the avatars, and 2) was the lack of useful functionality. One participant felt the avatars increased their enjoyment of the system, saying: *"It's a nice delighter but it does not enhance the main function/goal."*.

Next, in Figure 4.4 we see the different features suggested by participants as expected and desired in a burnout support system. Mood-tracking and Sleep-tracking constitute the most popular tools. Other interesting features suggested by users included 'emergency support', meaning immediately effective exercises for panic

---

[2] qualtrics.com

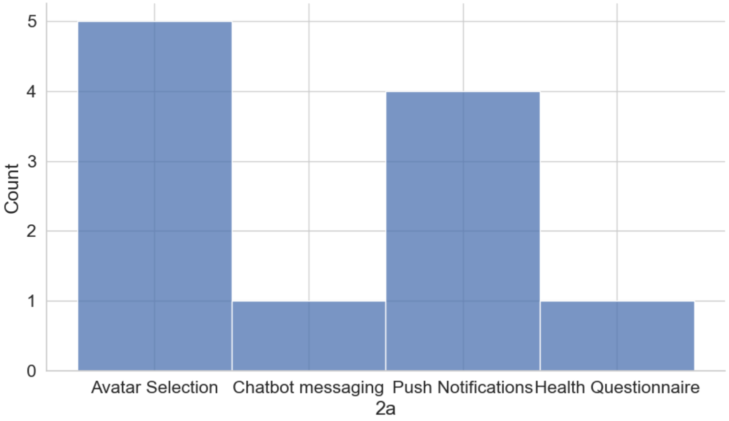


Figure 4.3: Features Ranked The Least Important

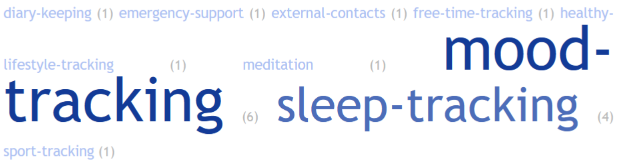


Figure 4.4: Requested Features

attacks, and 'external support' meaning the ability to highlight selected trusted external contacts in the app as potential support vectors.

sRQ2 yielded interesting results. In terms of the aesthetic design, 75% of participants preferred human-presenting over nonhuman-presenting avatars, largely citing increased perceptions of social presence. However, proponents of nonhuman avatars also put forth compelling arguments, where some felt that nonhuman-presenting avatars were 'cute' but completely inappropriate within system context (Quote: *"Burnout is a serious problem, talking to cute icons would make me feel like if I was in a game."*), they felt said avatars were 'cute' and more supportive of system goals than human avatars (Quote: *"The characters brings happy thoughts :) and the person one feels therapist-y"*). Furthermore, several participants expressed concerns about the lack of inclusivity in the avatar design choices and hoped for more customization in this area.

As for **sRQ3**, 58.33% of participants preferred a chatbot-focused style of goal editing. Those who preferred the chatbot emphasized ease-of-use, social presence, and communication. Navigation based editing flows were discussed as more editable and providing better goal overviews.

sRQ4 tested two dimensions of personality as expressed through language; ex-

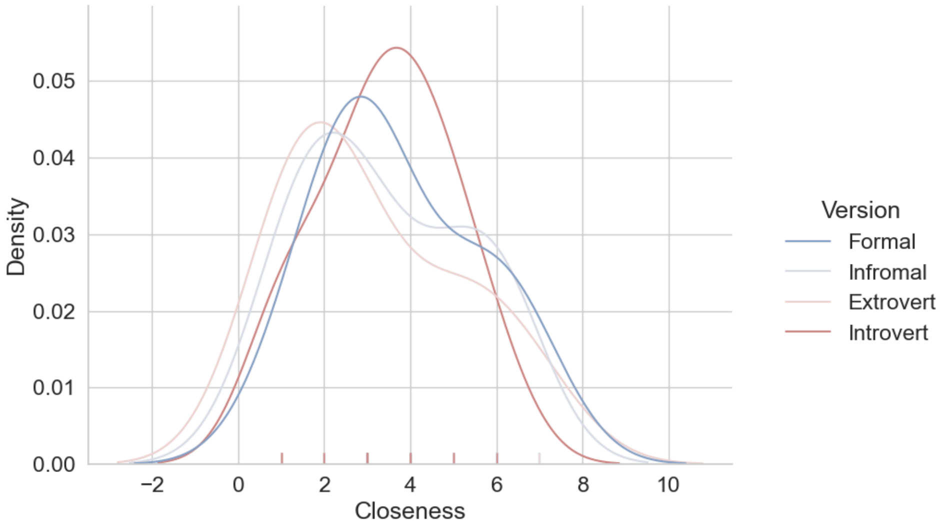


Figure 4.5: Predicted Closeness Across Four Linguistic Styles.

traversion – introversion and formality – informality. 66.67% of participants preferred a formal — extroverted chatbot. Some participants described the informal linguistic style as ’ingenuine’ and ’inappropriate’, while some actually read the formal speech as more ’casual’ citing the ’directness’ as an aspect of ’friendliness’. Similar aspects of ’friend-like behaviour’ and ’closeness’ were discussed in relation to introverted-linguistic-style use. This unexpected outcome is reflected in Figure 4.5, where formality and introversion seem to correlate with higher perceived closeness. In this analysis, we map the circles of the ISO to a numerical value from 1 (not close at all) to 7 (extremely close), then aggregate participant responses using kernel-based density estimators [141]. The estimators return a probability distribution of perceived closeness over four linguistic styles. We used a Gaussian kernel and the bandwidth smoothing method ’Scott’s Rule’ [151], such that data was smoothed by a factor of $n \times (\frac{-1}{d+4})$, for this analysis. Here n is the number of data points and d the number of dimensions.

Nonetheless, the extraverted bot was preferred by participants due to its greater perceived levels of warmth, care, and consideration (Quote: *“It feels professional but still friendly. Option 2 is too abrupt, to the point and cold. It seems as if something is wrong in the feelings of the person.”*).

Lastly, we wish to note that many participants expressed a sense of disorientation in their responses. In fact, we can see in Figure 4.6 that ’confusing’ was among the common keywords participants associated with the system (though the keywords ’easy’ and ’simple’ also occur the same amount). In some cases, this sense arose from the design of the study itself (Quote: *“I could not understand some instructions”*). In some others, it flowed from the structure of the system, where both push notifications

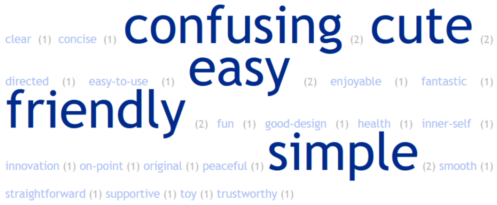


Figure 4.6: Keywords Participants Associated with Prototype

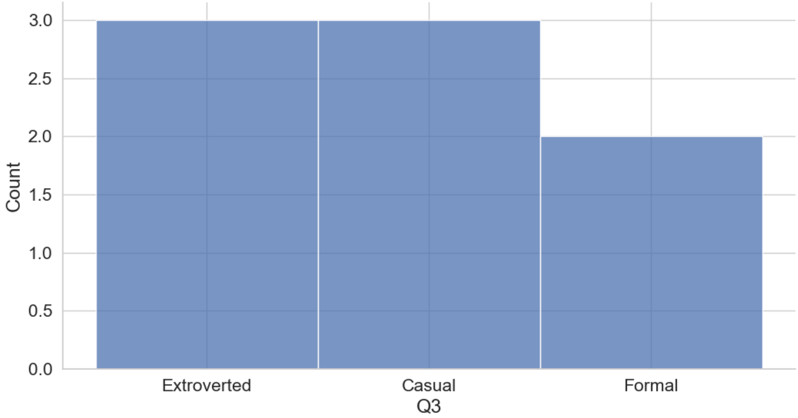


Figure 4.7: Perceived Chatbot Personality

and navigation order were points of confusion for users (Quote: *"[...] sometimes it was hard for me to keep track where I am in the app"*). Other notes included the desire for a more aesthetically neutral design.

Pilot Study

Sub-research question **sRQ5** was addressed in question three of our survey. The users described the personality of M as seen in Figure 4.7. These responses indicated we were able to successfully convey an extraverted communication style in our chatbot, though formality was seemingly not conveyed as clearly.

As for sub-research question **sRQ6**, it was addressed across several questions. Overall, participants were comfortable working with the bot (75% Yes, 25% Somewhat), though they did not feel particularly close to it (see Figure 4.8). Participants did, however, feel the bot was effective at providing emotional support (average Likert score $avg = 4, std = 0.71$), and interested in their needs ($avg = 4, std = 1.22$). In terms of humanness, the chatbot scored an average of 31.75 with $std = 2.49$, suggesting our bot conveyed a strong sense of social presence. See Figure 4.9 for a breakdown of humanness scores across their four dimensions.

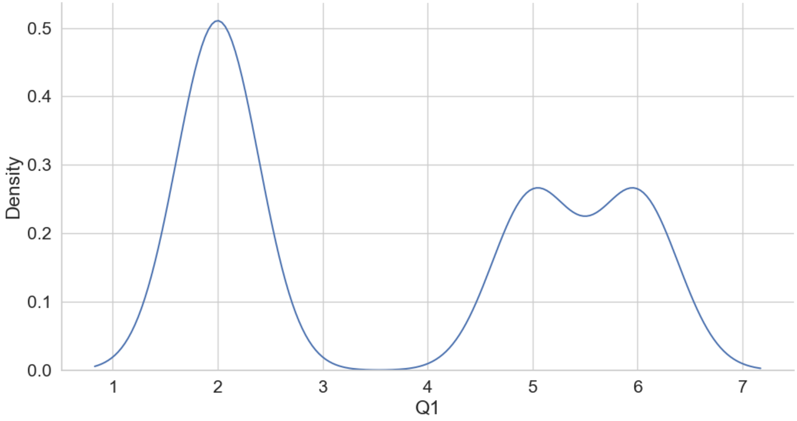


Figure 4.8: Perceived Closeness to Chatbot

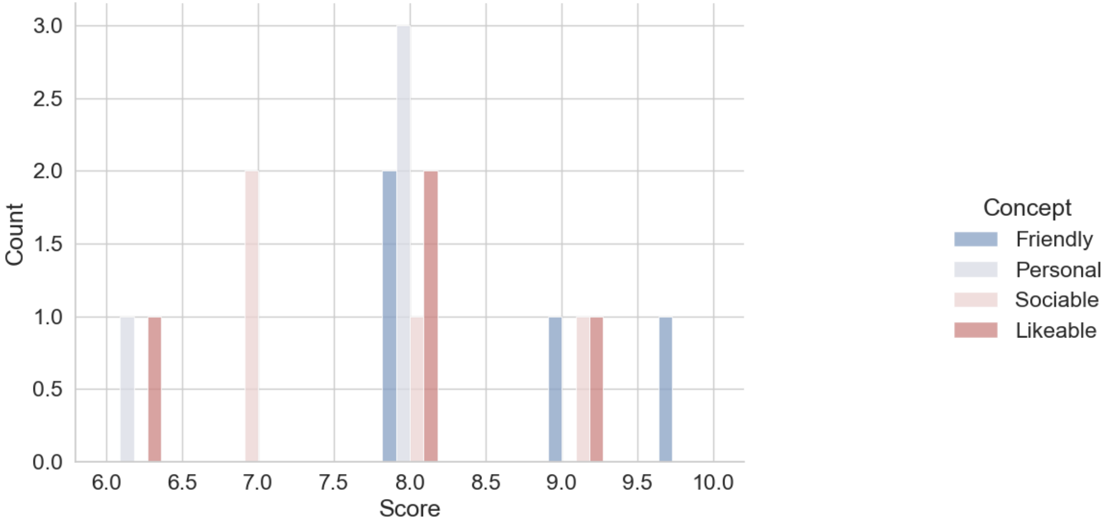


Figure 4.9: Properties of Humanness

Participants also noted that the prototype overall was informative (*"[...] it was giving me practical tips"*) and easy to use. However, some issues with bot behaviour were detected. Learning the user's name was a particular weak point, with one user pointing out the bot repeated their name in sentences, and another encountering an issue with the bot not learning their name at all. Moreover, it was not always clear to participants whether the bot's behaviour at a given moment was correct (*"It did require me to 'tell me more...' I'm guessing it couldn't comprehend my reply the first time."*).

4.4 Design Iterations

Based on these reviews and studies, we effected many changes throughout the project's lifetime.

Healthyboel, discussed by our expert consultants, eventually became the main fallback support for our system, and the point to which we redirect students seeking more help. Drawing from this same expert advice, we also elected to design our short-term interventions around mindfulness-, acceptance-, and value-based approaches with a focus on person-centred practice. Cognitive approaches would be reserved for users further along on their journey with the system. We thus built our introduction session milestones around context-awareness, relationship-building, and breath-work.

Mood-tracking features were introduced based on participants' suggestions. Similarly, the 'Avatar Selection' feature, present in the first prototype iteration, was dropped. In the second iteration of our system design, users interacted with a chatbot represented by a fixed human-presenting androgynous, ethnically ambiguous avatar instead. Moreover, chatbot-based goal editing flows were chosen over classic navigation styles, and non-core features were dropped to increase ease-of-use and design simplicity. This second iteration design was also made more aesthetically neutral.

Initial user response also led us to select a formal (or professional) but extraverted style of language for our chatbot's interactions.

Post pilot study, we further implemented some changes to our conversation design. Firstly, to increase the sense of social presence in chatbot interaction, message typing delay was changed from a fixed value to one dependent on string length. Moreover, some bot lines, such as *"tell me more about that..."*, were reworded to make it more clear when that response is appropriate, becoming *"I see what you are saying, tell me more..."*. Lastly, the inconsistent and often incorrect behaviour around name learning was addressed by removing this feature from the system.

Chapter 5

Implementation

Given user initial studies, and our review of similar chatbots in this field, we drew up a list of functional requirements, seen in Appendix K. From these requirements, the following technologies were selected for implementation: HTML and JavaScript to develop the overall system, and Rasa (Version 3.0) [16] to develop the chatbot specifically.

Rasa is an open source infrastructure platform for developing conversational AI based, as of Version 1.8, on the Dual Intent and Entity Transformer (DIET) intent-classification and entity-recognition architecture [104, 19]. The DIET architecture is modular and compact, able to work with other pre-trained word embeddings like BERT and GloVe, though by default it depends on more spare featurizers. It achieved an F1 score of 90.18% on intent classification, and 86.04% on entity recognition [19]. In our architecture, the DIET was also augmented with the pretrained English language NLP model from SpaCy [1], to improve named entity recognition. Training data for Rasa is provided in the form of developer-generated stories, intents, and actions which are formatted as YAML files (See Listing 1). Training data could also be collected from test conversations. Rasa was selected for this study due to its flexibility as a platform, its emphasis on data privacy and GDPR compliance [136], and the active development community surrounding it [2].

The prototype was made publicly accessible on ii.tudelft.nl/burnout-support/ using a Google Compute Engine[3] instance, a Socket.io [4,5] channel, and the Rasa interface widget, Botfront [6] (See Listing 2).

---

[1] spacy.io/usage/models

[2] forum.rasa.com

[3] cloud.google.com/compute

[4] socket.io/

[5] rasa.com/docs/rasa/connectors/your-own-website#websocket-channel

[6] github.com/botfront/rasa-webchat

```
stories:
  - story: greet
    steps:
      - intent: greet
      - action: utter_hello
      - action: utter_iamm
      - action: utter_whatisname
      - action: name_form
      - action: utter_nicetomeetyou
      - action: utter_askmood
      - checkpoint: moods
```

Listing 1: An Example Greeting Story for Model Training

5.1 Intervention Design

Structure-wise, we borrow the implemented design of our conversation from the clear intervention design structure of Addis *et al.* (2004) [3] as well as our discussions with experts, wherein the first step of a given intervention is the understanding of how 'depression', or in this case 'burnout', works and how it is related to the user's behaviour and experiences. This introduction was followed by the participant's sharing of said experiences and behaviours with the chatbot before moving on to the simple therapeutic exercise of breathing manipulation. Figures 5.1 & 5.2 illustrate the dialogue flow in the prototype.

In the upcoming subsections, we detail how different aspects of this flow were developed.

5.1.1 Administrating the OLBI

The OLBI, introduced in Chapter 2 consists of sixteen items answered using a five Likert point scale from *strongly agree* to *strongly disagree*. To work within the constraints of Rasa, and as we would not be performing diagnostic functions, the scale was abbreviated to five items scored on a two point scale. The abbreviated question set is seen in Appendix L. It was administered to users using Rasa's button response functionality. This reduces errors resulting from incorrect intent classification and speeds up the form-filling process, while maintaining the conversational style of the chatbot.

5.1.2 Listening to the User

The bot encouraged user self-disclosure and emotional expression by asking leading questions, using overt continuers, and expressing empathy, as discussed in Chapter 2. The bot also practised self-disclosure through emotional and personal utterances, encouraging reciprocal user disclosure.

```
const queryString = window.location.search;
const urlParams = new URLSearchParams(queryString);
const userid = urlParams.get('userid');

!(function () {
  let e = document.createElement("script"),
    t = document.head || document.getElementsByTagName("head")[0];
  (e.src =
    "https://cdn.jsdelivr.net/npm/rasa-webchat@1.0.1/lib/index.js"),
    (e.async = !0),
    (e.onload = () => {
      window.WebChat.default(
        {
            customData: { language: "en" , "userid": userid},
            selector: "#webchat",
            socketUrl: "http://<url>:5005",
            socketPath: "/socket.io/",
            fullScreenMode: true,
            embedded: true,
            title: "M: Burnout Support",
            inputTextFieldHint: "Say hi!",
            isChatOpen: true,
            showCloseButton: true,
            profileAvatar: "http://<url>/icon_bot_1.png",
            tooltipDelay: 500,
            initPayload: "/greet",
            hideWhenNotConnected: false,
            storage: "session"
        },
        null
      );
    }),
    t.insertBefore(e, t.firstChild);
})();
```

Listing 2: Abbreviated Botfront Chatbot JavaScript Interface

Throughout the conversation, the bot would prompt the user to speak about whatever is on their mind two to three times before proposing each goal.

5.1.3 Creating Goals

Goals were created for the user based on their answers to the OLBI, and presented to them as though learned from conversation.

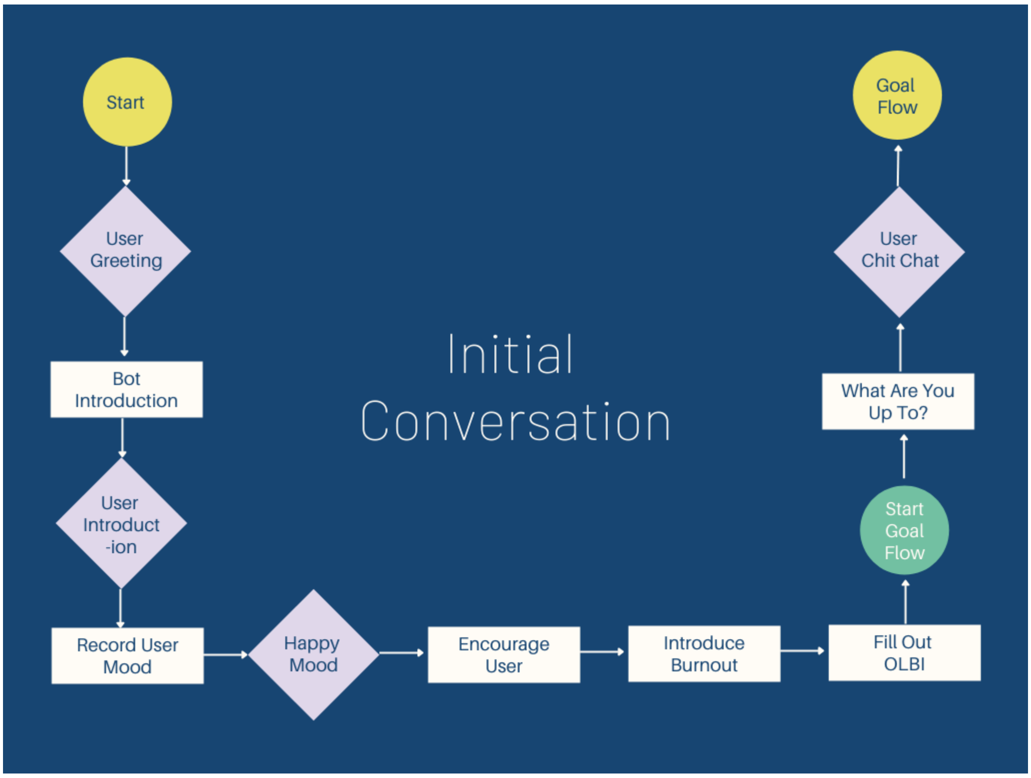


Figure 5.1: Dialogue Flow Part 1

Using Rasa's action server and custom slots, user answers were recorded as they worked through the OLBI. Once the form was complete, a score was calculated for each goal, equal to the score of its associated OLBI question (accounting for the inverted scores of negatively worded questions). The goals with the lowest score would be those most relevant to the given user. Purposefully irrelevant goals were also used to augment this goal list so that it was always possible to select blatantly incorrect goals and allow users the opportunity to correct the system.

The association between goals and OLBI questions was created based on psychological background theory, for example the goal 'being more present' was associated with question eleven. See Appendix M for the full mapping.

Whether the system proposed a correct or incorrect goal at a given point in the conversation was dependent on the experiment design. Proposed goals were not to be proposed again. Once a goal was proposed, a user could accept, deny, or change the goal. On choosing to change a goal, the user was prompted with two possible alternatives to choose from, those being the next most relevant goal, and an irrelevant goal.

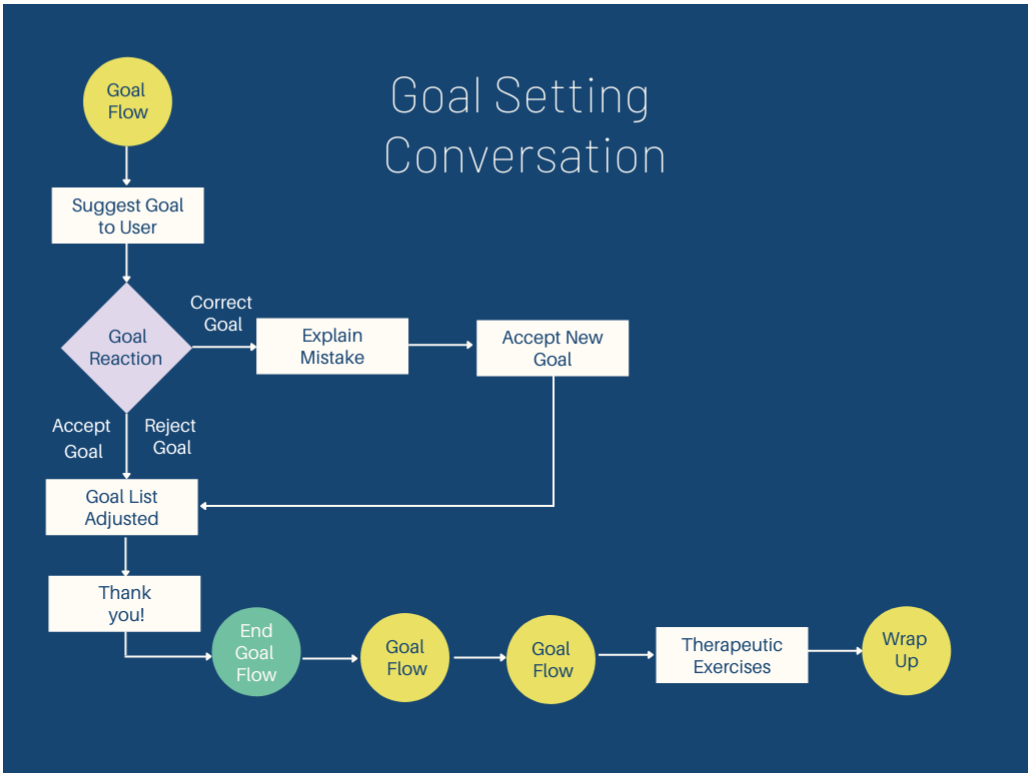


Figure 5.2: Dialogue Flow Part 2

5.1.4 Exercise Coach

Based on the aforementioned consultations with the TU Delft student psychologists, breathing manipulation was selected as the most relevant therapeutic exercise to the largest set of users. The exercise was implemented using an animated image which illustrates the 4-4-4-4 box breathing technique [121], and allows the user to sync their breathing to the animation. The image was sent along with a text message from the chatbot using Rasa's image message features.

Chapter 6

Experiment Design

6.1 Hypotheses

Previous studies in trust and perceived control have not employed granular, multidimensional, measures of trust. Thus, it is not possible for us to ground our hypotheses fully in the literature. However, Honeycutt et al. [59] do speculate in their discussion whether the negative impact on trust brought on by the interactive learning features was due to the action of system correcting embedding system mistakes more deeply in the users' memory than system successes. We thus expect our findings to conform with the following:

1. **H1**: Variance in control level has a positive correlation with perceived benevolence.

2. **H2**: Variance in control level has a negative correlation with perceived competence.

6.2 Participants

Over the course of two weeks, 109 participants were recruited for this experiment, the inclusion criteria demanded participants be over 18, currently attending a university, or a recent graduate, and comfortably proficient with the English Language. Of those participants, 35 completed the whole study, for a follow-through rate of 32.11%

Participants were not screened for clinical burnout, but assumed to experience some level of stress in their day-to-day life as students, which would lead them to be interested in our prototype.

Study participants had the characteristics summarized in Table 6.2 and Figure 6.1.

Gender	63.33% Female, 33.33% Male, and Other 3.33%
Age Group	3.33% 18-21, 56.67% 21-24, 23.33% 24-27, 13.33% 27-30, and 3.33% 35+.
Status	36.67% Undergraduate, 30% Postgraduate, and 33.33% Recent Graduate.
Occupation	3.33% Art, Design, and Entertainment, 3.33% Bio and Chemical Engineering, 10% Business, Finance and Management, 26.67% Mechanical, Electrical, and Aerospace Engineering, 6.67% Healthcare, 6.67% Law, Politics, and History, 3.33% Social Sciences, 40% Computer Science and Mathematics.
Geographic Area of Origin	16.67% Southern Europe, 6.67% South Asia, 36.67% Western Europe, 30% East Asia, 3.33% Middle East, 3.33% North America, and 3.33% Northern Europe.
Had Previous Experience with Wellbeing Apps	23.33% Yes, and 76.67% No.
Technical Skills	20% Far above average, 43.33% Somewhat above average, 26.67% Average, 6.67% Somewhat below average, and 3.33% Far below Average.

Table 6.2: Characteristics of the Surveyed Population

6.3 Measures

A few metrics for user experience are used to gather user feedback and perform evaluations:

- **Trust**, via the Human-Computer Trust Scale [52]. The Human-Computer Trust scale uses four dimensions; benevolence and competence, which are common to most trust scales, as well as perceived risk, replacing the more common integrity, and general trust. It is a 12 item, 5-point Likert scale. Items measuring perceived risk were excluded from this study as it was not an investigated concept.

- **Perceived Control**, via an adapted version of the scale developed by Yu [179]. This scale was developed to measure control in intelligent-system-user interaction. The scale was modified as to better suit our experiment task. Where Yu [179] used the consistency of interaction rhythm in mixed initiative chatbots to manipulate sense of control, we used goal editing. Therefore, the statement *I was controlling the pace* was rewritten as *I was controlling M's understanding of me.* Moreover, while the original version used the statement *The software intended to challenge me* to convey feelings of frustration with an inconsistent system, we used *M intended to complete its own task* to better

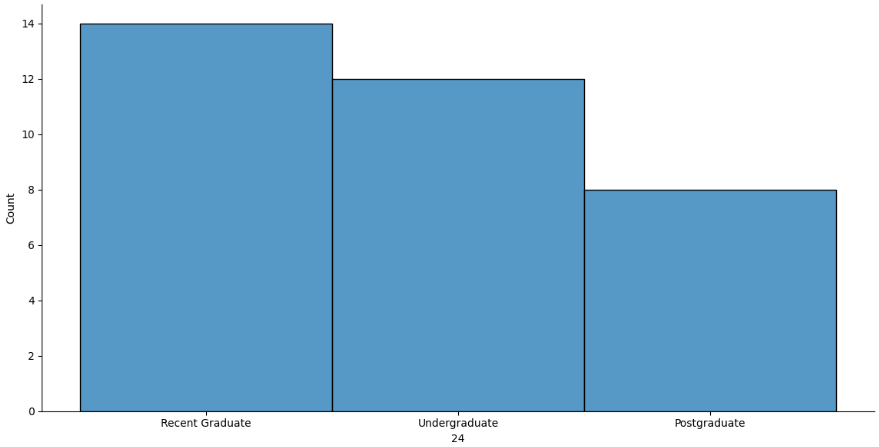


Figure 6.1: Distribution of Educational Background of Sample Population

align with the kind of impression a negative interaction with a supposedly adaptive system would invoke. The scale consists of 5 bipolar items.

- **Usability**, via the UMUX-LITE [87]. Usability is generally defined around the effectiveness and efficiency of an artefact, as well as the user's satisfaction with their product experience [1], and is generally the core metric used in UCD [64]. The UMUX-LITE is a 2-item, 7-point, Likert scale designed to concisely measure subjective usability.
- **Humanness or Mindless Anthropomorphism**, as introduced in Chapter 4 from Kim *et al.* (2012) [72].
- **Closeness**, as detailed in Chapter 4 from Aron (1992) [9].
- **Attitude Towards AI**, via two, five-point Likert scale, items used to categorize users based on their pre-existing Attitude Towards AI. We borrow these items from the 20 item questionnaire proposed by [147].

All measures and questionnaires can be found in Appendices O and P.

6.4 Procedures

First, all participants completed a pre-test qualitative survey establishing their baseline attitude towards interactive intelligent systems. Users also provided demographic data including age range, gender, geographic area of origin, and technical skill level. Participants were then randomly split into control and experimental groups, where they conversed with the chatbot as it suggested therapeutic goals to

---

[1] usability.gov/what-and-why/usability-evaluation.html

them. The control group were unable to correct the chatbot when it was mistaken, while the experimental group were. The bot proposed goals to users in the following pattern: irrelevant goal, most relevant goal, irrelevant goal. Thus, the system showed some inconsistent signs of improvement after it has been corrected, but did not give the unrealistic impression that it could learn perfectly from one piece of feedback. The in-between subject experiment design was chosen to avoid biasing the user's impression of the chatbot.

During the post-test, participants completed the human-computer trust, perceived control, usability, closeness, and humanness surveys. Users also answered a few short-answer questions regarding the reasoning behind their survey answers, allowing them to delve deeper into their experience.

All data were collected via an online survey hosted on Qualtrics. Study procedures and relevant concepts are illustrated in Figure 6.2. Study procedures were approved the TU Delft Human Research Ethics Committee (application number: 2005).

We note here that over the course of the two weeks of data collection, we did enact one change to the behaviour of the chatbot. The conversational branches were pruned so that the model behaved in a more stable manner and classified intention more consistently. Issues with intent misclassification did in fact persist after this change, but seemed to be less severe. Before this change we had recruited 19 participants, the remaining participants were recruited afterwards. We expect this change made it more likely for participants to have been able to complete the requisite bot interaction before moving on to the post-test survey, but would not have affected their perception of system usability too drastically, since misunderstandings were not completely eliminated.

6.4.1 Data Analysis

Data were processed so that Likert scale answers were converted into numerical scores (inverting for negatively valenced questions) and then aggregated and normalized. Attitude towards AI and Skill with Technology were not normalized. As survey questions did not force participants to answer, some items had missing answers. In those cases, respondent answers on the inventory associated with said missing item were dropped from the analysis.

Quantitative Analysis

The first step of our analysis was to perform a manipulation check. By comparing perceived control across our experimental and control groups, we verified whether withholding the ability to edit goals significantly affected participants' sense of control. Testing was carried out using a Bayesian t-test [78]. Bayesian analysis is the distribution of degree of belief over a range of values given a set of input data [78], allowing us to form a more probabilistic understanding of our hypotheses [132]. We performed this analysis using JASP [66], calculating first the posterior distribution

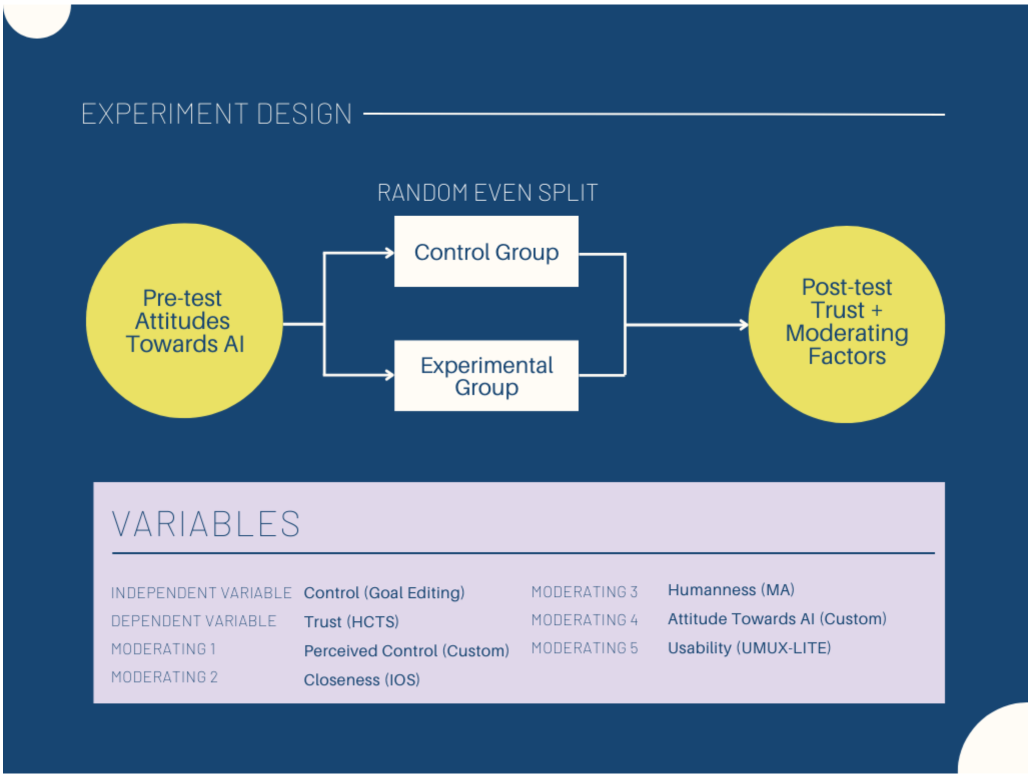


Figure 6.2: Experiment Procedures

and the Bayes factor. Bayes factor is the ratio between marginal likelihoods of null and alternative models, and is used to quantify the extent to which evidence is likely under a model [132]. A Bayes factor of $3 - 10$ is generally considered moderate evidence, and a factor of $10 - 30$ strong evidence [67]. Moreover, moderation analysis with Bayesian Linear Regression was used to investigate whether a significant relationship exists between participant demographic information and perceived control.

Next, we investigated the effect of perceived control variance on competence, benevolence, and general trust in a two-level between subject design, using Bayesian correlation. For our prior distribution, we used a Cauchy distribution centred around $d = 0.45$ [148], since based on prior literature [59, 54], we expected to observe a medium-size $Cohen's\ d$ effect in our data. The Cauchy distribution is similar to the normal distribution, albeit with less central mass [142], and is a common choice for Bayesian priors [132]. To further test the robustness of our evidence, we also repeat our analysis over a range of prior values [132]. Lastly, we again use moderation analysis, this time using both Bayesian and frequentist Linear Regression, to study the interaction effects of secondary factors on the relationship of perceived control and trust (components).

Qualitative Analysis

Collected qualitative feedback was analysed using grounded theory approaches [49]. An initial codebook was developed using open-coding on a third of collected responses ($n = 10$), then refined and finalized using ($n = 21$) responses.

Chapter 7

Results

7.1 Manipulation Check

In this analysis, users who had trouble completing the conversational flow with the bot, and therefore only saw one or no goal(s) throughout the interaction, were excluded from the data set. This is because they did not interact enough with the goal setting and modification mechanism for their sense of control to be influenced by the manipulation. Of our $n = 35$ participants, this excluded $n = 11$, for a total of $n = 24$ responses, with a further two responses excluded for missing data. However, it is worth noting that users may have underreported how many goals they saw throughout the conversation when answering the survey, though since we did not log user conversations we cannot correct for this.

We split our data on experimental group, so that we had $n = 12$ participants in the Control group, and $n = 11$ participants in the Experimental one. We then compared the two group's mean perceived control using an independent samples t-test with the default prior of 0.707. From Figure 7.1 we see the effect was anecdotal [67]. With a Bayes Factor of 0.985, observing the data we do is 0.985 time more likely to occur under our model (where perceived control was higher in the experimental group) than the alternative, null, model. In Figure 7.2 we can see that our evidence remained in the anecdotal range across all possible prior values.

Our experimental group showed only a marginally higher average perceived control ($M_{exp} = 0.185, M_{control} = -0.278$), further suggesting goal modification to not have contributed strongly to manipulating the user's sense of control. See Figure 7.3.

7.1.1 Other Factors and Perceived Control

Next, we use Bayesian linear regression to ascertain if other factors, specifically participant demographic information, significantly moderated the effect of experimental group on perceived control. Using Bayesian linear regression, with default JASP model prior values (a beta binomial prior with $a = 1, b = 1$), we analyse the model in Equation 7.1, where the interaction of the newly introduced factors and

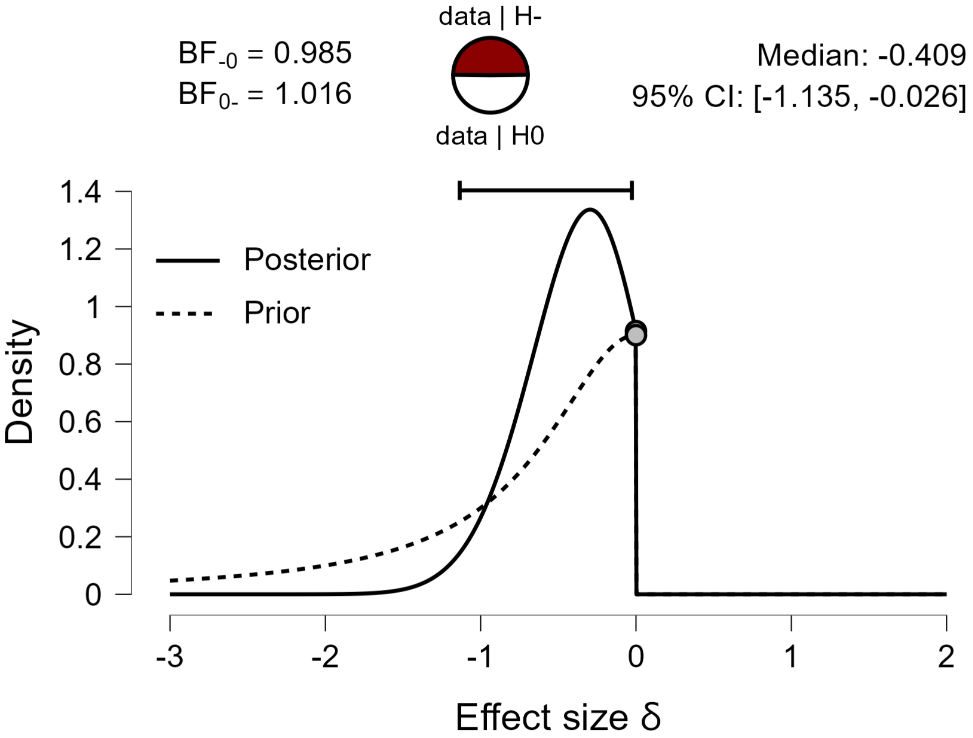


Figure 7.1: Prior vs Posterior Distribution of Perceived Control By Experimental Group

the experimental group represents these factors' moderating effect [15]. Different models were automatically generated from this baseline formula and their predictive power was compared, the six best models are listed in Table 7.1.

$$\begin{aligned} precivied - control = \beta_0 \cdot group + \beta_1 \cdot gender + \\ \beta_2 \cdot age + \beta_3 \cdot origin + \beta_4 \cdot tech - skill + \\ \beta_5 \cdot (group * gender) + \beta_6 \cdot (group * age) + \\ \beta_7 \cdot (group * origin) + \beta_8 \cdot (group * tech - skill) + \epsilon_i \end{aligned} \tag{7.1}$$

We can see that participant demographic information had no predictive power over their perceived control. For all models except the model relying solely on experimental group to predict perceived control, the predictive Bayes Factor (BF_{10}) was < 1.00. Again, we can see the weak relationship between experimental group and perceived control, with the experimental group model being marginally more effective at perceived control prediction than the null model ($BF_{10} = 1.515 > 1.00$).

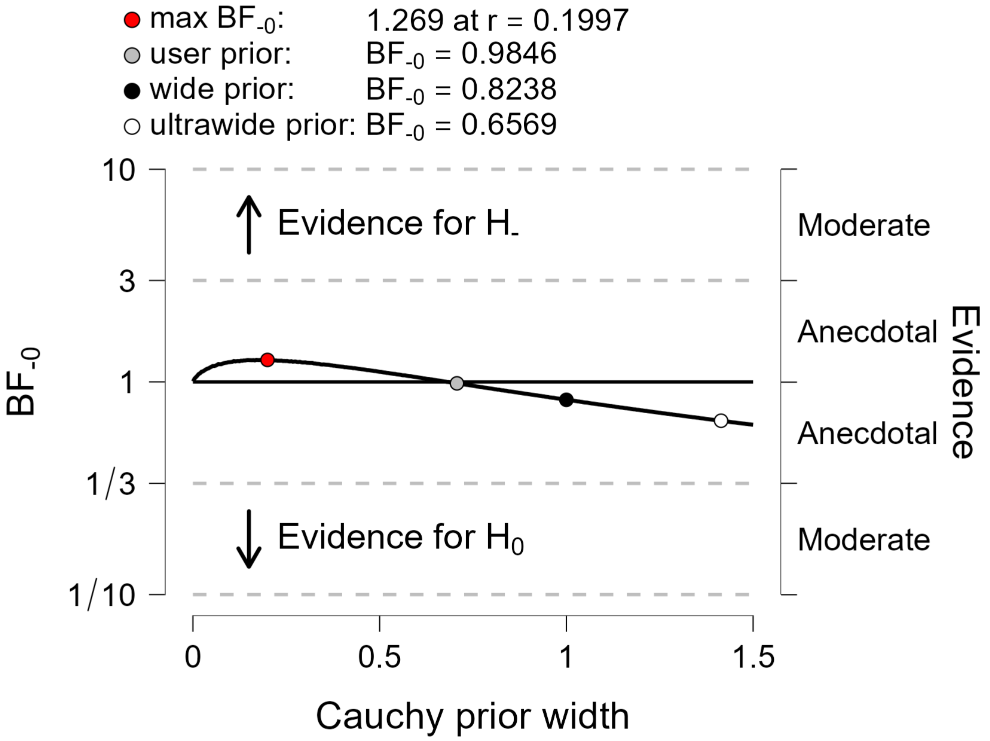


Figure 7.2: Bayes Factor in T-test across Several Priors: Perceived Control By Experimental Group

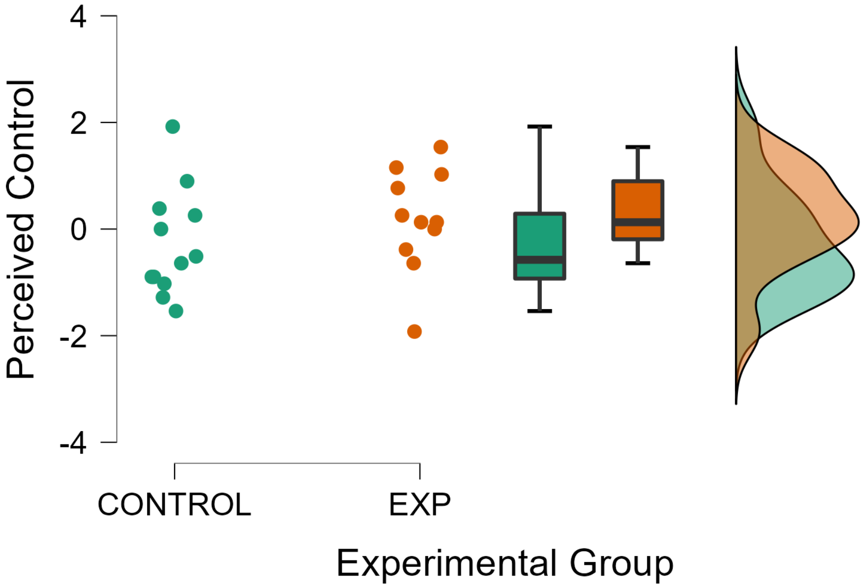


Figure 7.3: Raincloud Plots of Perceived Control Means by Experimental Group

Table 7.1: Comparison of the six best models of the perceived control—experimental group relationship

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
Null model	0.238	0.443	2.540	1.000	0.000
Group	0.026	0.075	2.962	1.515	0.116
Gender + Age + Origin + Skill + Group + Group * Gender + Group * Age + Group * Origin + Group * Skill	0.238	0.058	0.198	0.132	0.324
Age	0.026	0.038	1.464	0.779	0.068
Origin	0.026	0.031	1.160	0.622	0.050
Skill	0.026	0.030	1.133	0.607	0.049

7.2 Perceived Control and Trust

7.2.1 Initial Analysis

Since the relationship between experimental group and perceived control is weak, we did not analyse differences in trust between these two groups of participants. Instead, we used Bayesian correlation to investigate the relationship of perceived control with benevolence, competence, and general trust directly, with a prior width of $\frac{1}{3}$ [148]. However, we did, again, exclude the $n = 11$ participants who encountered only one or no goal(s) in the conversation. From Table 7.2.1 we can see that in the cases of general trust and competence, this correlation was anecdotal, while benevolence showed moderate to anecdotal (see Figure 7.4) correlation with perceived control ($r = 0.453, BF_{10} = 3.016$). See Figure 7.5. This evidence then goes some way towards supporting our first hypothesis (**H1**), but not our research question (**RQ1**), since the variance in perceived control was not necessarily caused by the ability to edit goals.

			n	Pearson's r	BF_{10}	Lower 95% CI	Upper 95% CI
Perceived Control	-	Comp.	23	0.185	0.601	−0.227	0.484
Perceived Control	-	Ben.	23	0.453	3.016	0.003	0.652
Perceived Control	-	Trust	23	0.055	0.461	−0.322	0.398

Table 7.2: Bayesian Pearson Correlations: Perceived Control & Trust Dimensions

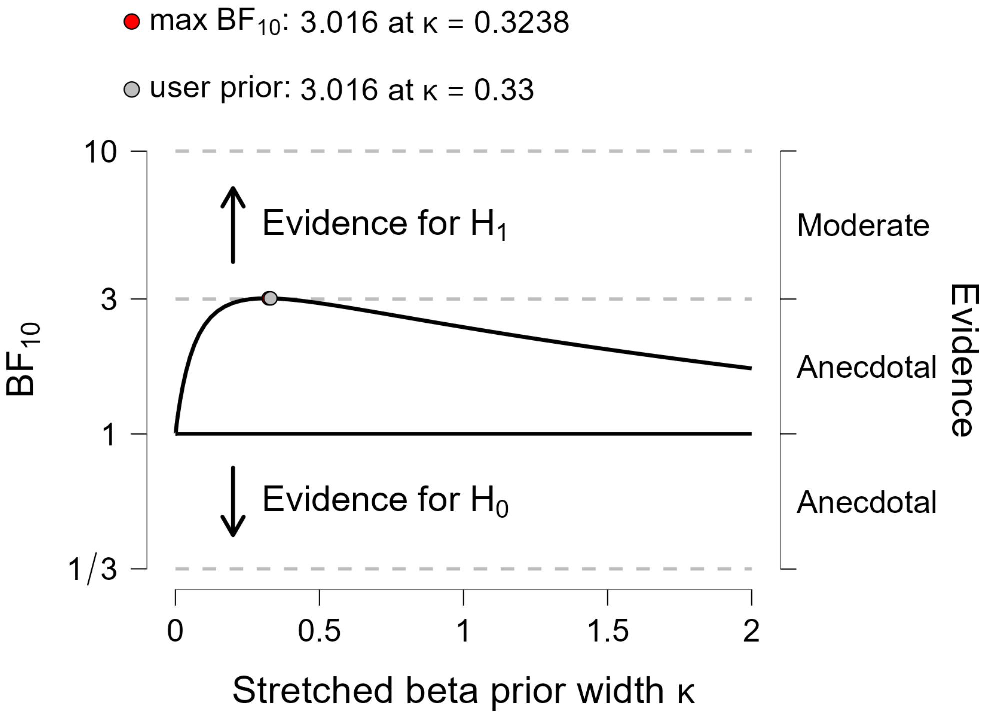


Figure 7.4: Bayesian Pearson Correlation Robustness Check: Perceived Control & Benevolence

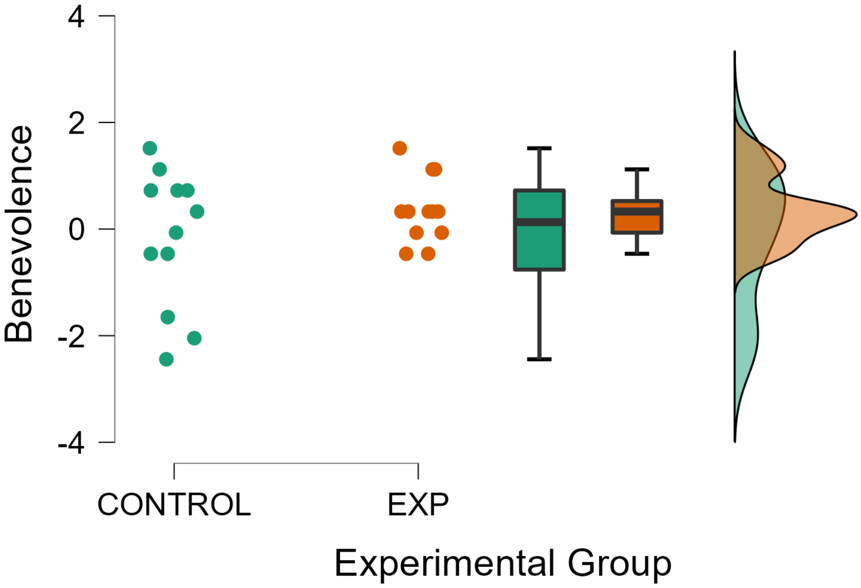


Figure 7.5: Raincloud Plots of Benevolence grouped by Experimental Group

7.2.2 Expanded Analysis

Here, the previously excluded participants, who had been removed on the basis of the depth of their interaction with the bot, were re-introduced to the data set. In this analysis, we again cannot claim that the variance in perceived control, and whatever effect said variance had on trust, were results of goal editing ability. However, since we are interested in perceived control variance in general, whether it arises from technical issues, design choices, or from our experimental setup, it follows that we should investigate the effects of this variable on benevolence, competence, and general trust across all participants. Of these participants, two were excluded for missing data. In this analysis, we saw moderate evidence for an influence of perceived control over benevolence (See Figure 7.6) with $r = 0.448, BF_{10} = 7.150$. Evidence for this relationship remained in the moderate range across all priors (see Figure 7.7). If we group participants by their perceived levels of control (High: $x \in x \geq 17, n = 15$. Low: $x \in x < 17, n = 20$), we can also see this relationship in the difference in mean perceived benevolence between the two subsets ($M_{high} = 0.381, M_{low} = -0.2860$) (See Figure 7.8). Nonetheless, the exact effect size in this relationship remains fairly uncertain; bound with 95% confidence between 0.082 – 0.630. Meanwhile, competence and general trust again exhibited only anecdotal evidence towards a correlation with perceived control. Thus, we can say our data offers support towards **H1**, but not **H2**, or **RQ1** and **RQ2**.

Table 7.3: Bayesian Pearson Correlations: Perceived Control & General Trust

			n	Pearson's r	BF_{10}	Lower 95% CI	Upper 95% CI
Perceived Control	-	Ben.	33	0.448	7.150	0.082	0.630
Perceived Control	-	Trust	33	0.222	0.749	−0.131	0.475
Perceived Control	-	Comp	33	0.291	1.232	−0.070	0.523

Other Factors and Trust

We then studied the effects of moderating factors on the relationship between perceived control and trust, much as we have done for experimental groups and perceived control. Instead of demographic information, we are now interested in other aspects of the system interaction measured during the experiment, namely usability, humanness, closeness, and attitude towards AI. Using Bayesian linear regression, we analysed the model in Equation 7.2, using a beta binomial model prior of $a = 3, b = 3$ [148] [15].

Since benevolence and perceived control were moderately correlated in the previous analysis, we began by investigating this relationship further. In Table 7.4 we can

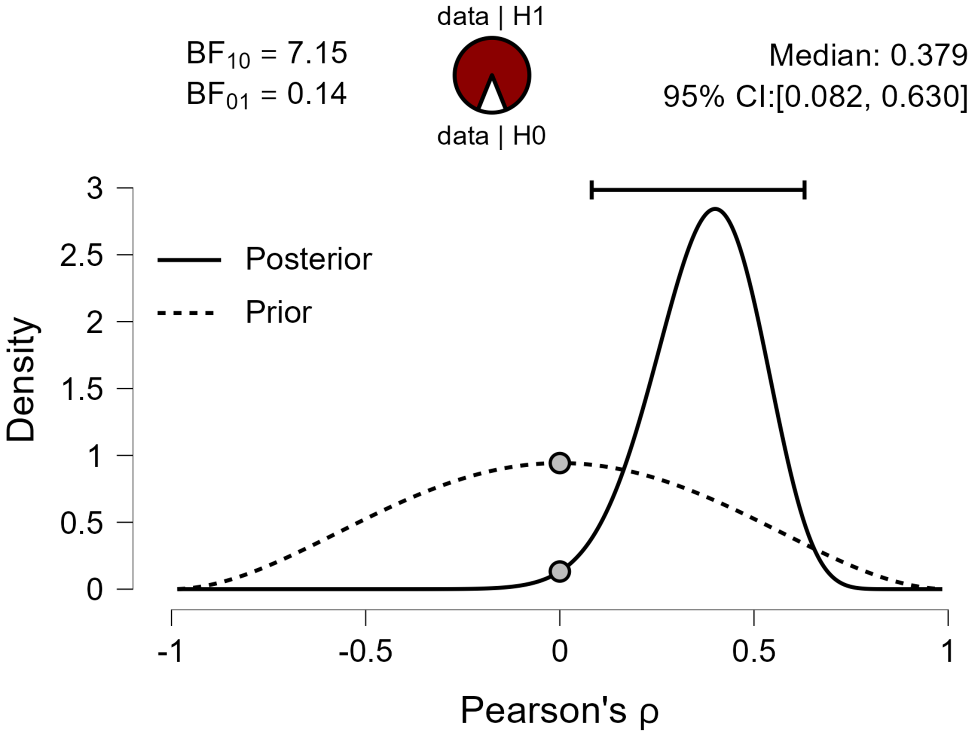


Figure 7.6: Bayesian Pearson Correlations: Perceived Control & Benevolence

see that usability, humanness, and closeness had little moderating effect on the relationship between perceived control and benevolence, all exhibiting less likelihood to be included in a predictive model of benevolence than perceived control. Meanwhile, the interaction effect of perceived control and attitude towards AI exhibited a higher likelihood of inclusion than perceived control alone ($BF_{inclusion} = 1.117 > 0.745$). This interaction was also the only one which was included within the top six predictive models of benevolence (See Table 7.5).

Interestingly enough, the best models of benevolence in this analysis mostly did not use perceived control at all, and the top model relied on closeness alone as predictor with posterior odds of $P(M|Data) = 0.084$. However, the model which does contain perceived control (as well as humanness and attitude towards AI), though of lower posterior odds, has 3.093 times the likelihood of co-occurring with our observed data, than a model containing only closeness [167]. If we use classic frequentist multiple linear regression to compare these two models, we can see the latter achieves higher R^2 and adjusted R^2, as well as lower root mean squared error (RMSE) (See Table 7.6). Thus, though both models achieve statistical significance (See Table 7.7), the model containing perceived control, humanness, and attitude

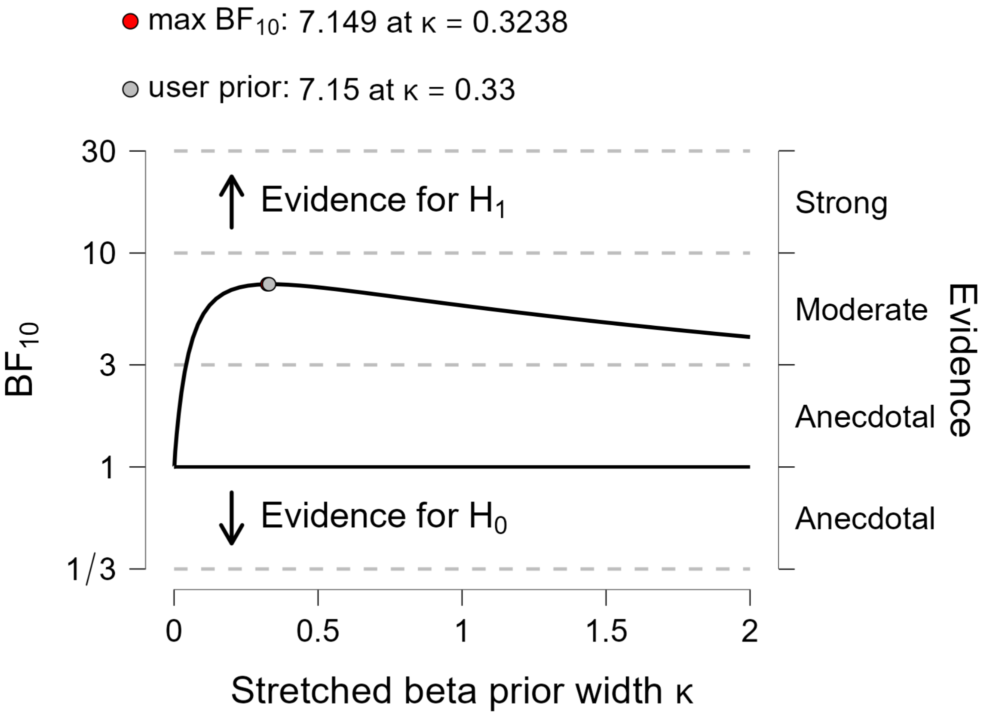


Figure 7.7: Bayesian Pearson Correlations Robustness Check: Perceived Control & Benevolence

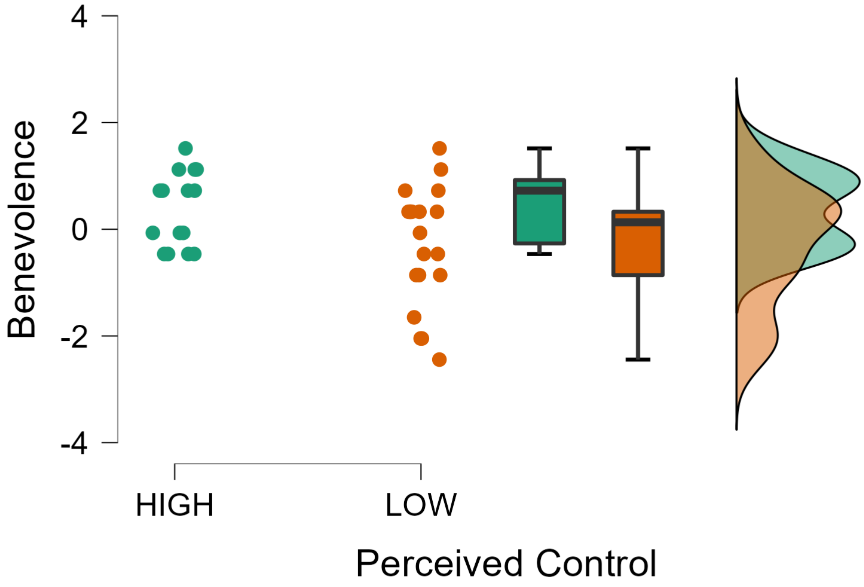


Figure 7.8: Raincloud Plots of Benevolence grouped by Perceived Control

towards AI is a better fit over our data [55].

$$\begin{aligned} benovalnce = \beta_0 \cdot control + \beta_1 \cdot closeness + \beta_2 \cdot humanness \\ +\beta_3 \cdot attitude + \beta_4 \cdot usuability + \beta_5 \cdot (control * closeness) \\ +\beta_6 \cdot (control * humanness) + \beta_7 \cdot (control * attitude) \\ +\beta_8 \cdot (control * usuability) + \epsilon_i \end{aligned} \tag{7.2}$$

Table 7.4: Posterior Summaries of Coefficients: Benevolence & Perceived Control

Coefficient	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BF$_{inclusion}$
Intercept	1.000	0.000	1.000	0.000	1.000
Perceived Control	0.699	0.301	0.634	0.366	0.745
Humanness	0.598	0.402	0.711	0.289	1.653
Attitude	0.598	0.402	0.668	0.332	1.349
Closeness	0.598	0.402	0.721	0.279	1.738
Usability	0.598	0.402	0.378	0.622	0.408
Per. Cont. * Humanness	0.334	0.666	0.251	0.749	0.670
Per. Cont. * Attitude	0.334	0.666	0.359	0.641	1.117
Per. Cont. * Closeness	0.334	0.666	0.180	0.820	0.438
Per. Cont. * Usability	0.334	0.666	0.152	0.848	0.357

Table 7.5: Comparison of the six best models of the benevolence-perceived control relationship

Models	P(M)	P(M|data)	BF$_M$	BF$_{10}$	R^2
Closeness	0.027	0.084	3.302	1.000	0.428
Closeness + Attitude	0.011	0.053	5.089	1.568	0.511
Humanness + Closeness	0.011	0.048	4.595	1.423	0.506
Per. Cont. + Humanness + Attitude + Per. Cont. * Attitude	0.005	0.043	10.006	3.093	0.633
Humanness + Closeness + Attitude	0.006	0.042	7.224	2.240	0.576
Humanness	0.027	0.034	1.268	0.405	0.383

Table 7.6: Linear Regression Model Comparison: Benevolence

Model	R	R^2	Adjusted R^2	RMSE
$H_{per.cont.0}$	0.000	0.000	0.000	0.951
$H_{per.cont.1}$	0.803	0.645	0.588	0.610
$H_{closeness.0}$	0.000	0.000	0.000	1.000
$H_{closeness.1}$	0.601	0.361	0.341	0.812

Table 7.7: ANOVA of two benevolence predicting models

Model		Sum of Squares	df	Mean Square	F	p
$H_{per.cont.}$	Regression	16.925	4	4.231	11.354	$< .001$
	Residual	9.317	25	0.373		
	Total	26.242	29			
$H_{closeness}$	Regression	12.262	1	12.262	18.614	$< .001$
	Residual	21.738	33	0.659		
	Total	34.000	34			

We repeat this analysis with general trust and competence as our dependant variables. Moderating effects on perceived control remained minimal. In the case of competence, the best predictors were closeness and usability (see Table 7.8). As for general trust, it was best predicted by attitude towards AI and usability. Tables 7.9 and 7.11 show both models to also have statistically significant predictive powers. Our model of competence achieved an $R^2 = 0.591$, an adjusted $R^2 = 0.564$, and an $RMSE = 0.666$. As for trust, this predictive model achieves an $R^2 = 0.570$, an adjusted $R^2 = 0.543$, and an $RMSE = 0.685$.

Table 7.8: Comparison of the six best models of the competence—perceived control relationship

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
Closeness + Usability	0.011	0.238	28.524	1.000	0.654
Usability	0.027	0.127	5.232	0.214	0.551
Attitude + Closeness + Usability	0.006	0.089	16.092	0.671	0.686
Closeness	0.027	0.057	2.158	0.095	0.520
Attitude + Usability	0.011	0.052	4.980	0.217	0.606
Per. Cont. + Closeness + Usability	0.006	0.050	8.672	0.377	0.669

Lastly, we analysed the difference in effect between competence and benevolence on general trust. Once again, we used Bayesian linear regression with default JASP prior settings. Here we can see perceived competence emerged as a more strongly predictive factor of general trust than perceived benevolence (See table 7.12). When

Table 7.9: ANOVA of Competence's best predictive model

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	19.285	2	9.643	21.717	$< .001$
	Residual	13.320	30	0.444		
	Total	32.606	32			

Table 7.10: Comparison of the six best models of the trust—perceived control relationship

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
Usability	0.027	0.232	10.860	1.000	0.511
Attitude + Usability	0.011	0.147	15.707	1.580	0.586
Closeness + Usability	0.011	0.119	12.311	1.279	0.578
Attitude + Closeness + Usability	0.006	0.074	13.239	1.438	0.630
Closeness	0.027	0.036	1.328	0.154	0.429
Humanness + Usability	0.011	0.032	3.039	0.347	0.527

Table 7.11: ANOVA of General Trust's best predictive model

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	19.316	2	9.658	20.586	$< .001$
	Residual	14.544	31	0.469		
	Total	33.860	33			

correlating competence and general trust, we could see the strength of this relationship with $r = 0.868, BF = 3.151e + 8$, in comparison to the $r = 0.515, BF = 26.065$ of the benevolence—general trust correlation (see Figure 7.9, as well as Table 7.13). This implies that, in our use case, participants prioritized competence over benevolence in their general trust judgements.

Table 7.12: Comparison of all models of trust based on competence and benevolence.

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
Comp.	0.167	0.792	19.008	1.000	0.754
Ben. + Comp.	0.333	0.208	0.526	0.132	0.754
Ben.	0.167	$7.784e - 8$	$3.892e - 7$	$9.831e - 8$	0.263
Null model	0.333	$8.498e - 9$	$1.700e - 8$	$5.367e - 9$	0.000

Table 7.13: Bayesian Pearson Correlations: Trust & Competence vs Benevolence

			n	Pearson's r	BF_{10}	Lower 95% CI	Upper 95% CI
Ben.	-	Trust	35	$0.515*$	26.065	0.205	0.710
Comp.	-	Trust	34	$0.868***$	$3.151e+8$	0.728	0.929

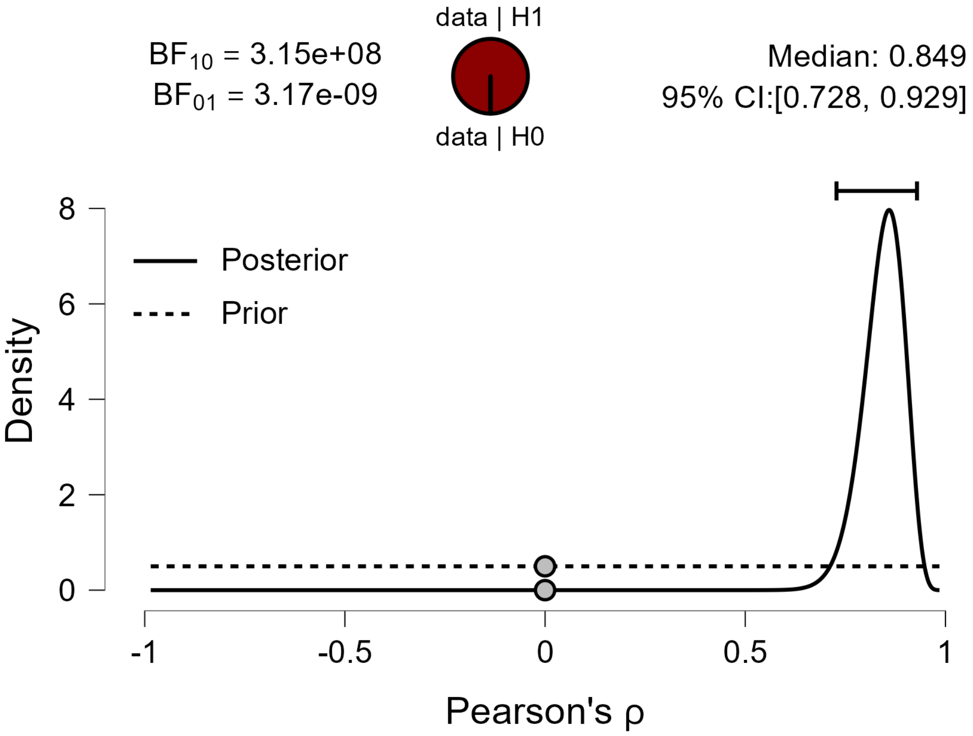


Figure 7.9: Bayesian Pearson Correlation: Competence & Trust

7.3 User Feedback

Of our $n = 35$ participants, $n = 31$ answered the optional qualitative question on perceived changes in accuracy, $n = 28$ answered the question on factors which affected their sense of trust, and $n = 29$ addressed the importance of goal editing capability. We analysed the first third of these responses to create our initial codebook, then refined and finalized said codebook with the remaining responses. Selective coding was used to cluster and categorize granular codes. The codes, their descriptions, and categories can be seen in Appendix Q.

7.3.1 Accuracy and Improvement

Of the $n = 25$ participants who indicated whether they felt the chatbot’s recommendations improved over the course of the conversation when answering this question, 52% said Yes, 44% No, and 4% were Neutral. This is an interesting mix of results, as we know from our experimental setup that recommendation relevance was fixed throughout the conversation, and did not take into account user feedback even though the user was informed it would. Nonetheless, more users felt their feedback was effectively incorporated into the chatbot than not, citing factors such as supportiveness (*“Yes, it seemed to understand what I was feeling/needed.”*), adaptation (*“Yes, it went from completely generic to somewhat personalized”*), and competence (*“Yes, M came up with right conclusions”*). On the other hand, participants who disagreed also cited competence as a key influence over their perception of model accuracy (45.5% of all participants who felt the model did not improve during the conversation discussed this perception in relation to concepts of competence) (*“No. I didn’t mention anything related to M’s recommendations”*). Other influences were disorientation, a concept somewhat related to poor competence but with a greater focus on the system and study usability and less on the chatbot’s suitability to its tasks, and lack of explainability (*“No, I did not understand why things w~~[h]~~ere suggested”*). Similarly, the remaining neutral response referenced both competence and adaptation, such that the participant felt the system did adapt to some of their responses but did not prioritize its recommendations well. Participants who did not address the question of accuracy fluctuations in their answers instead discussed disorientation and technical issues they encountered with the system, including the skipping of conversation segments and the failing to understand user input. See Figure 7.10.

7.3.2 Goal Editing and Trust

Participants cited many factors as impacting their trust in the system. 25% discussed elements related to benevolence (*“The friendly tone and words M used, which made me feel safe[...]”*). 39.3% mentioned competence related factors (*“it is still sometimes not understanding what I[’]m telling[it,] so feels a bit dumb”*). Three participants explicitly discuss the automated nature of the system, both negatively and positively (*“[...], but because it is rather transparently a chatbot. I don’t generally trust systems like these because they can break easily.”*), and two the academic context in which the conversation was taking place (exclusively positive) (*“M itself didn’t assure me that our conversation would be private. If we removed the context of this being a study[,] that would definitely affect my sense of trust for it”*). Lastly, lack of explainability was discussed as a deterrent to trust (*“I don[’]t know what information is in M and what the recommendations are based on”*). See Figure 7.11.

In terms of the capacity to edit goals, 76.9% or (20 out of 26) of participants wished for, or appreciated, this feature, while 19.2% were not interested, and 3.8% were neutral. This was somewhat unexpected, as our quantitive results indicate

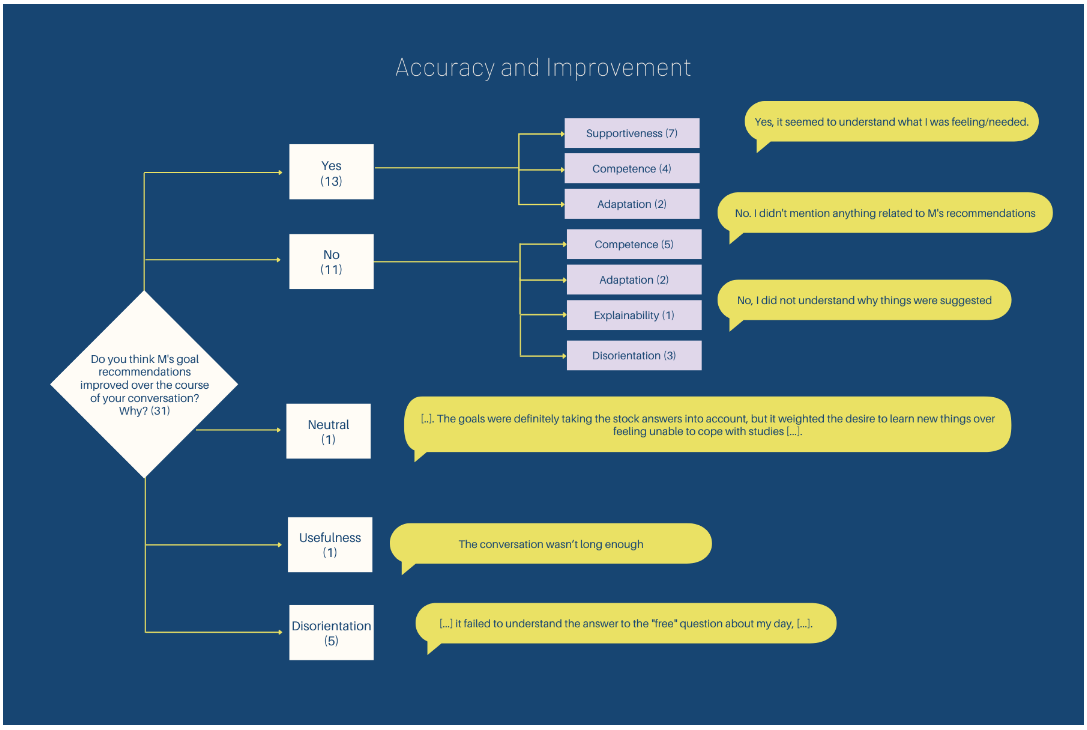

Accuracy and Improvement
Do you think M's goal recommendations improved over the course of your conversation? Why? (31)
Yes (13)
Supportiveness (7)
Competence (4)
Adaptation (2)
Yes, it seemed to understand what I was feeling/needed.
No. I didn't mention anything related to M's recommendations
No (11)
Competence (5)
Adaptation (2)
Explainability (1)
Disorientation (3)
No, I did not understand why things were suggested
Neutral (1)
[..]. The goals were definitely taking the stock answers into account, but it weighted the desire to learn new things over feeling unable to cope with studies [...].
Usefulness (1)
The conversation wasn't long enough
Disorientation (5)
[...] it failed to understand the answer to the "free" question about my day, [...].

Figure 7.10: Feedback on perceptions of model accuracy

goal editing did not have a strong positive effect on users' perceived control, and yet it seems that users' nonetheless valued this feature greatly. Of the participants who were not interested in goal editing, only one explicitly discusses this feature as *"[...] too much trouble"*. Otherwise, participants were explicitly interested in this opportunity for increased autonomy *"Yes, it would make it more customisable and relevant to me"* and possibly for improved chatbot performance *"Yes, this gave me the feeling of stirring th[e] chatbot into a more useful direction"*.

7.4 Discussion

The goal of our research questions, **RQ1** and **RQ2**, was to investigate whether providing feedback in the form of goal editing to an "ML" system would affect a user's perceived benevolence, competence, and general trust. We expected that participants who were able to change goals proposed to them by the system, instead of simply accepting or rejecting them, would feel a greater sense of system benevolence, but have a lower perception of its competence. In this chapter, we have only tangentially addressed these questions, since our manipulation of perceived control via goal editing and (simulated) feedback incorporation was not very effective, and thus limited our ability to speak to the effects of goal editing on the user experience. Given the prominence of disorientation in the participant feedback we received, we

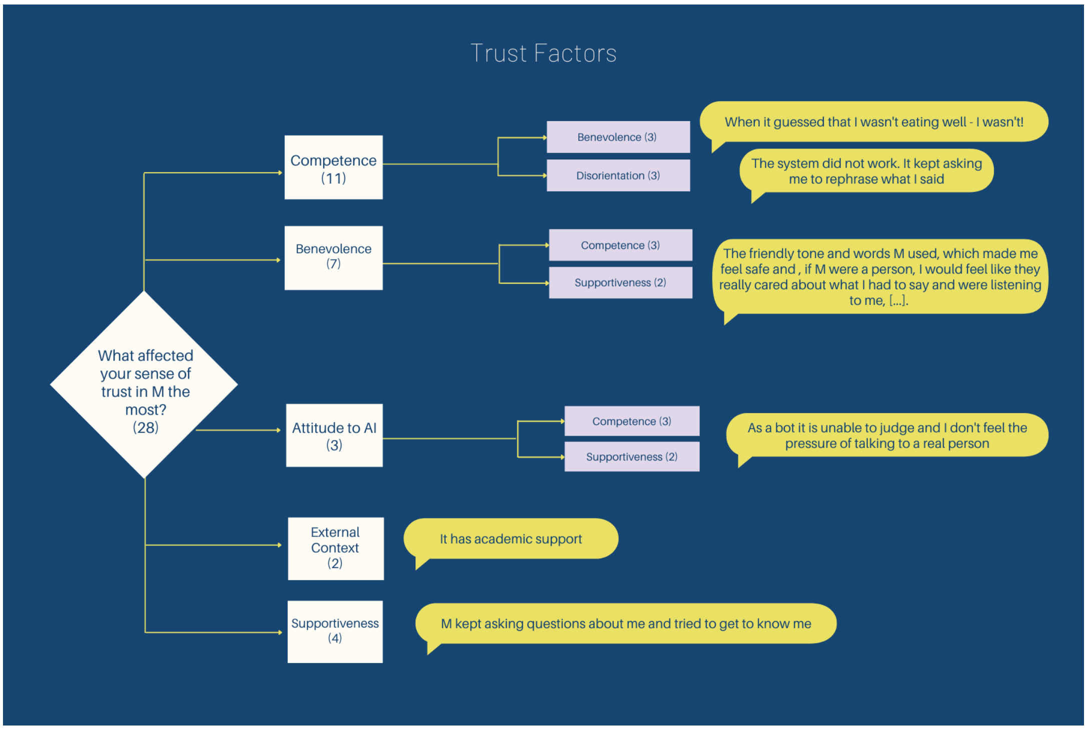

Trust Factors
What affected your sense of trust in M the most? (28)
Competence (11)
Benevolence (3)
Disorientation (3)
When it guessed that I wasn't eating well - I wasn't!
The system did not work. It kept asking me to rephrase what I said
Benevolence (7)
Competence (3)
Supportiveness (2)
The friendly tone and words M used, which made me feel safe and , if M were a person, I would feel like they really cared about what I had to say and were listening to me, [...].
Attitude to AI (3)
Competence (3)
Supportiveness (2)
As a bot it is unable to judge and I don't feel the pressure of talking to a real person
External Context (2)
It has academic support
Supportiveness (4)
M kept asking questions about me and tried to get to know me

Figure 7.11: Feedback on trust influences in system interaction

expect this lack of effect was due to perceived control being too strongly affected by other aspects of the system, such as intent misclassification and failure to understand the users' input, to be effectively manipulated through goal editing.

However, by divorcing the method through which we achieve variance in perceived control from that selfsame variance, we did learn much about perceived control in general, and its relationship to perceived benevolence, competence, and general trust. We observed that perceived control was moderately, positively, correlated with benevolence with a Pearson's $r = 0.448$. This effect size, which falls above the median reported effect size in human research [150], lends support for hypothesis **H1** that increased control would co-occur with increased perceptions of benevolence. We now wonder if this relationship between perceived control and benevolence is largely the reason studies have seen a positive correlation between trust and interactivity in machine learning? While incorporating user feedback into a system may or may not communicate ability on behalf of said system, it does communicate a certain consideration of user needs on behalf of both the system and its developers (to quote one participant: *"I trust M in the sense that it will always want the best for me, [...]"*), elements which are closely related to benevolence.

Though perceived control exhibited only a weak correlation with general trust and competence within our study, it is possible such a relationship could emerge given a larger dataset. When investigating a medium size effect using an in-between subject design, frequentist power analysis [38] and Bayesian sample-size literature [43], both indicate a population in the range of $154 < n < 216$ would best enable its observation, a population much larger than recruited in this study. Furthermore, it is unlikely we would see a closer relationship between general trust and perceived control without observing a similar relationship between perceived control and competence. This is because, within our study setting, competence and general trust seem to be much more strongly correlated than general trust and benevolence. Such a relationship could indicate that competence is generally a more important factor in predicting user trust, but it could also follow from perceived competence (non-normalized) being lower ($M_{competence} = 9.212, SD = 3.150$) than perceived benevolence ($M_{benevolence} = 11.353, SD = 2.321$) in our setting, and thus more of a deciding factor in user trust. If this were the case, we would then expect a more competent system with lower benevolence to exhibit a stronger relationship between trust and benevolence. Of course, it is possible no actual relationship exists between general trust, competence and perceived control, and the effect detected on trust by previous studies was entirely due to the effect on benevolence. However, if only benevolence is tied to control via feedback incorporation, then it does not follow that trust should decrease with user feedback as was seen in [59], thus we venture that the relation does exist, but was not visible in our dataset.

As for our secondary factors, we saw evidence for a mixture of expected and unexpected relationships.

In the case of benevolence, the strong influence of closeness and humanness is most expected. Though all three scales have different goals, they do measure concepts in relation to friendliness, intention, and intimacy, thus their correlation is apt. Perceived control and benevolence, though it may the principal relationship under investigation here, is a less obvious connection, and it is an indicator of the strength of this correlation that perceived control is predictive of benevolence even in comparison to these more related concepts. Attitude towards AI is then the most interesting case, as it is the only secondary factor which expressed not only direct influence over benevolence, but also moderated the relationship between benevolence and perceived control. We can see how such a connection came to be; if a user is naturally mistrusting of intelligent systems, even gestures of benevolence such as feedback solicitation may be viewed as manipulative. If a user is overly positive towards automation, they may place too much emphasis on this interactivity and attribute unearned benevolence to the agent.

The other surprising relationship is the direct correlation of competence and perceived closeness in the agent-user dyad. It is possible that users who felt closer to the agent were more willing or likely to overlook mistakes or technical issues during the interaction with the chatbot, leading to higher perceived competence and vice versa. We can see how such a relationship can then lead to misaligned trust in the system, much in the same way as attitude towards AI can. Either the user is overestimating the system's capability due to an overly intimate reading of their relationship with it, or underestimating it because they do not feel close enough. While this is already an issue as over- and under- trust can lead to mis-

and dis- use of intelligent systems [81], it is also an issue because it may allow developers to maliciously drive up user-system closeness without improving actual system trustworthiness or capability.

The issue with separating manipulation method from data variance then becomes the difficulty of comparing our results to those we have seen in previous studies from which we have taken inspiration. Though previous studies have also used the manipulation of feedback incorporation into model development to affect sense of control in the user, they did not (in all but one study [156] which did not measure trust) actually measure perceived control. Thus, such studies analyse the relationship between a tangential but not identical independent variable (feedback incorporation) and trust, where we study perceived control and trust (components). We therefore largely limit our conclusions to the scope of our study. Nonetheless, one point of comparison is possible in relation to perceived accuracy improvement. In both our study and [59] a surprising level of confidence is shown on the part of participants that their feedback was incorporated into the model. [59] interpret this increased perception of feedback usage as following from increased capacity to provide feedback. We suspect the same applies in our case, where even simply allowing the users to reject goals may have impacted their sense of model adaption over time.

7.5 Threats to Validity

Possible threats to validity followed as a consequence of some of our design choices.

Firstly, we acknowledge our sample size is small, which in turn limits our ability to generalize our findings to other settings and populations. Further, it limits our ability to observe small-to-medium sized effects from our dataset, and so may be preventing us from commenting on relationships that do in fact exist in the general population.

Second, we note the subjectivity involved in detecting chatbot mistakes in this setting. Though intent misclassification is easy enough for the user to observe, goal recommendations, though we may have formulated them as irrelevant or relevant, will not always be read thus by the user. Therefore, not every user, even controlling for number of goals seen, will have observed the same ratio of hits and misses in their subjective experience of the system. Since we chose to perform most of our analysis directly on the correlations of perceived control and trust components, this variance was not as concerning as it would have otherwise been. If we choose in the future to analyse this data via independent group means testing on experimental group, we would also have to be careful to control for how many goals a user accepts (assuming that accepted goals are perceived as correct by the user).

Lastly, we acknowledge the non-deterministic nature of our chatbot, which was developed via machine learning over conversational data, which can act as an uncontrolled source of mistrust at points. Since there was no way for us to guarantee the chatbot will behave as expected at every conversational turn, the user then has to handle issues of misunderstandings and failures, which greatly affects their view of the system and their capacity to use it to complete the study. In some cases, the bot may fail without the notice of the user such that important parts of the conversation are missed without knowledge of the user, which again would affect their relationship greatly. As user-agent conversations are not recorded, we can only estimate when these failures occur based on user answers to our survey, and control for this variance accordingly (hence the two-pronged analysis earlier in this chapter).

Chapter 8

Conclusions

In this thesis project, we have focused on building a chatbot which can be used for supporting postgraduate students experiencing burnout, as well as enabling the investigation of the relationship between perceived control (through feedback incorporation) and different dimensions of trust. Though similar chatbots existed which support a variety of mental health issues [177, 131], none had focused on burnout specifically. Moreover, while some studies had investigated the effect of interactive machine learning on trust, none had used granular, dimensional, measures of trust [59, 54, 156]. Our two research questions (**RQ1** and **RQ2**) directed our investigation in the direction of simulated goal learning, goal editing, and its effects on benevolence and competence. Though we were not able to address these research questions throughout the course of our research, we were able to investigate the correlation of general perceived control and trust components directly. In this analysis, we found a moderate positive correlation between perceived control and benevolence. Furthermore, we were able to design, develop, and deploy a chatbot which both supported students, and enabled us to conduct our experiment.

8.1 Contributions

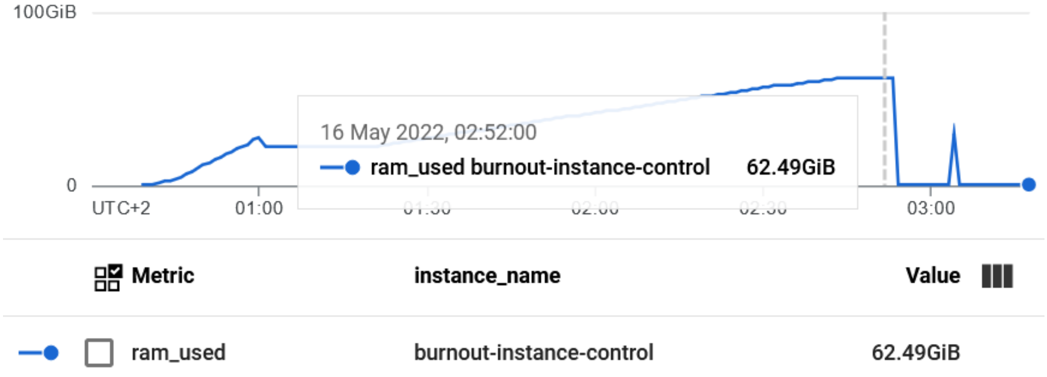
Some implications arise from our prototype evaluation. We know the incorporation of user feedback into model behaviour and system operation can be a complex process, does the effect this process has on trust make it a worthwhile investment? Based on our results, we would highly recommend this design choice. Though this feedback incorporation mechanism had only a small effect on perceived control, and thus its effects on trust (components) cannot be ascertained from this study alone, the users' feedback clearly states that such interaction was important to them. Users were clear that these interactions would help them understand learning systems better and improve their sense of autonomy, and should help them improve the system itself as well. Of our participants, only one felt that this level of interactivity would be too much work. Thus, the recommending of increased user control over intelligent systems [50, 51, 7] is a justifiable design strategy.

In our investigation, we did also uncover some moderating, and directly influential, factors within our trust (components) models. Attitude towards AI especially emerged as an important moderating variable in relation to trust and benevolence, and while pre-test trust and propensity to trust questions are a common part of many trust studies, pre-existing user conceptions and expectations of AI are less so. Though it was not a moderator, usability, as expected [47], emerged as a strong direct influence over competence and general trust. Unexpectedly, user-agent closeness also directly influenced competence. We thus encourage future studies to be mindful of these elements when designing experiments which involve perceived control so as not to misattribute variance in their data. We also advocate for the further development of concise measures for attitude towards AI in order to enable research to work with this variable more easily within their studies. Lastly, we want to emphasize the importance of aligning signals of user-agent closeness in chatbots, such as linguistic style and personalization, with actual agent competence to avoid encouraging inappropriate trust in the system.

8.2 Limitations

A few limitations cropped up throughout this project's lifetime. Firstly, the issues of project funding and project timescale, which intertwined to mostly affect our evaluation phase, namely our capacity for participant recruitment. With our low follow-through rate, and inability to secure small financial incentives for participation or extend our recruitment phase, we have limited our study to a small population size. Though this population may be fairly representative of our end goal user, this did limit the extent to which we can generalize our findings and our capacity to observe small-to-medium effects, as discussed in Chapter 7. It is even possible that, with more incentives for completing the study and a more robust system, we would have attained a higher follow through rate and increased our sample size within the same two-week timeframe.

Second, was the issue of hardware and virtual machine (VM) memory use upper bounds. During the implementation phase of this project, we quickly became aware that past a given conversation length or number of conversation branches, it would become impossible to train our models on any hardware or virtual machine we had access to. This is because these large models would then create extremely large training examples which could not be contained in memory and training would fail. See Figure 8.1 where one model was consuming $62.49GB$ of memory at one point before crashing. Such a limitation then forced us to be more cautious with the size of our conversation; fallbacks which were unlikely to be triggered had to be removed, fine-grained mood detection was likewise cut. In turn, this affected the robustness of our chatbot, the realism of its conversation, and presumably, the follow-through rate of participants.



Metric	instance_name	Value
☐ ram_used	burnout-instance-control	62.49GiB

Figure 8.1: Google Cloud Platform VM Instance Memory Use

8.3 Future work

In the future, a longer term interaction with study participants would be an exciting venture. The reasoning for this is two-fold. First, a longer study would allow us to study the effects of trust (components) variance on intervention adherence and effectiveness over time, which would then address a key motivator behind our choice of study context. Second, should we implement a functional goal-learning interactive machine learning model, it would also to present the user with a more realistic model improvement curve as affected by their feedback. Therefore, we would also have a closer estimate of users' sense of control in the real world when such systems are unpredictable or incorrect. A longer interaction would more closely resemble how users would rely on such support systems in their day-to-day life, and repeated interaction would likely affect how their sense of trust in the system forms and changes throughout the study.

The reproduction of this study with a larger sample size and a more robust chatbot could be a worthwhile pursuit. In terms of robustness, such development could possibly necessitate the migration of our prototype away from end-to-end chatbot development frameworks, and towards models which allow for finer-grain control on the part of developers, and which can handle longer conversations. Alternatively, the robustness of our bot could be improved using Conversation Driven Development [1]. In this process, users would have early access to test our prototype, and their conversations would be collected, reviewed, and annotated by the development team to be fed back to the conversational model as training data. Of course, the ability to collect private user conversations, especially in relation to a sensitive topic such as their mental health, is not a given, and should be something that is carefully negotiated between users, developers, and relevant research ethics bodies. Nonetheless, given a more consistent interaction with the system, and a larger dataset, we

---

[1] rasa.com/blog/conversation-driven-development-a-better-approach-to-building-ai-assistants/

would be able to more closely approximate the real effect size in the relationship between perceived control and trust (components) and thus offer more concrete design recommendations for the incorporation of user feedback in learning systems.

8.3.1 Expert Opinion

If we consider our prototype not as an experimental tool, but as a potential mental health support system, further improvements become evident. We drive these largely from our discussion with the TU Delft counselling team, who have been involved throughout the design process. Through a semi-structured interview, we concluded a few recommendations. Though our experts were approving of the linguistic style and conversational structure used in our system, they also highlighted the frustration of intent misclassification and unexpected bot behaviour. This frustration was seen as particularly detrimental to students who are likely already tense and frustrated, thus future development of this product would need to prioritize robustness and predictability.

Our experts also recommended the future addition of a tension tracking feature along with the existing mood tracker. As stress and tension are large components of the burnout syndrome, encouraging students to become more aware of how tense they may be, could then help them better tackle these issues. This feature expansion would also be accompanied by the implementation of more therapeutic exercises highlighted in our literature and expert discussions to further empower users with appropriate coping tools. Lastly, our experts recommended we provide our users with more information explaining how and why our therapeutic exercises work, as to better motivate them to follow along with said exercises, as well as recommending the correction of some terminology used in relation to these exercises for clarity.

8.4 Chatbots in Burnout Support

Over the course of this project we have worked through all development stages of human centred design as discussed in Chapter 2. In requirement elicitation, we relied on experts and literature to ascertain appropriate specifications for such a system. In the design stages, we conducted user studies, expert interviews, and iterative design adjustments. We implemented our final prototype and evaluated the system as a whole with a final user study. Thus, in this document, we present an example of not only a chatbot focused on the burnout domain, but one which was also developed using a user-first approach which has become increasingly popular in recent years [10, 161, 130]. Given the sensitivity of our context. we are glad to have chosen this style of prototype development. Both expert and user feedback was key to the calibration of this systems design, without which we would not have been to a chatbot which was sufficiently supportive, mindful of user-agent boundaries, and rooted in evidence-based literature on burnout and mental health support.

8.5 Trust and Perceived Control

In the evaluation phase of our project, we studied the relationship of the user-agent dyad in relation to trust and trust components. We used three subscales from the Human-Computer Trust scale [52] along with a modified version of the perceived control scale proposed by [179] to measure these variables. Our Bayesian analysis confirmed one hypothesis (**H1**) though we were unable to confirm **H2, RQ1, or RQ2**. Using frequentist linear regression, we also created more complex models of trust components, the most salient of which here is that of benevolence which included humanness, attitude towards AI, and perceived control, as well as the interaction of those two latter variables. Also noteworthy, was the model of competence, which included usability and closeness.

These findings allowed us to make several design and research recommendations for future work. We were able to confidently support the incorporation of user feedback in learning systems based on user feedback. We highlighted attitude towards AI as a possibly important moderator of the relationships of trust components and perceived control, and encouraged further research of this variable. Lastly, we noted the possibility for malicious manipulation of perceived system competence on the part of developers through the communication of inappropriate user-agent closeness.

Bibliography

[1] Alaa A. Abd-alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M. Bewick, Peter Gardner, and Mowafa Househ. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978, 2019. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2019.103978. URL `https://www.sciencedirect.com/science/article/pii/S1386505619307166`.

[2] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, 37(4):445–456, 2004.

[3] Michael E. Addis and Christopher R. Martell. *Overcoming depression one step at a time: The new behavioral activation approach to getting your life back.* New Harbinger Publications, 2004.

[4] Abby Aizenstros, David Bakker, Stefan G. Hofmann, Joshua Curtiss, and Nikolaos Kazantzis. Engagement with smartphone-delivered behavioural activation interventions: A study of the moodmission smartphone application. *Behavioural and Cognitive Psychotherapy*, 49:569–581, 9 2021. ISSN 14691833. doi: 10.1017/s1352465820000922.

[5] Fatima Alabdulkareem, Nick Cercone, and Sotirios Liaskos. Goal and preference identification through natural language. In *IEEE International Conference on Requirements Engineering*, pages 56–65. Ieee, 8 2015. ISBN 978-1-4673-6905-3. doi: 10.1109/re.2015.7320408.

[6] Irwin Altman and Dalmas A Taylor. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston, 1973.

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. *Guidelines for human-AI*

interaction, page 1–13. Association for Computing Machinery, 5 2019. ISBN 9781450359702. doi: 10.1145/3290605.3300233.

[8] Galit Armon, Arie Shirom, Itzhak Shapira, and Samuel Melamed. On the nature of burnout–insomnia relationships: A prospective study of employed adults. *Journal of Psychosomatic Research*, 65(1):5–12, 2008. ISSN 0022-3999. doi: 10.1016/j.jpsychores.2008.01.012.

[9] Arthur Aron, Elaine N Aron, and Danny Smollan. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology*, 63(4):596, 1992.

[10] Mohammed Bahja, Rawad Hammad, and Gibran Butt. A user-centric framework for educational chatbots design and development. In Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones, editors, *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, pages 32–43, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60117-1.

[11] Marco Bardus, Samantha B. van Beurden, Jane R. Smith, and Charles Abraham. A review and content analysis of engagement, functionality, aesthetics, information quality, and change techniques in the most popular commercial apps for weight management. *International Journal of Behavioral Nutrition and Physical Activity*, 13(1), 2016. doi: 10.1186/s12966-016-0359-9.

[12] Erin T. Barker, Andrea L. Howard, Rosanne Villemaire-Krajden, and Nancy L. Galambos. The rise and fall of depressive symptoms and academic stress in two samples of university students. *Journal of Youth and Adolescence*, 47: 1252–1266, 6 2018. ISSN 15736601. doi: 10.1007/s10964-018-0822-9.

[13] Aaron T. Beck. *Cognitive therapy of depression.* Guilford Press, 1987.

[14] Charles B. Bennett, Camilo J. Ruggero, Anna C. Sever, and Lamia Yanouri. ehealth to redress psychotherapy access barriers both new and old: A review of reviews and meta-analyses. *Journal of Psychotherapy Integration*, 30:188–207, 6 2020. ISSN 15733696. doi: 10.1037/int0000217.

[15] Jeremy C. Biesanz, Carl F. Falk, and Victoria Savalei. Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research*, 45(4):661–701, 2010. doi: 10.1080/00273171.2010.498292.

[16] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management, 2017. URL `https://arxiv.org/abs/1712.05181`.

[17] Soo Borson, Anna Korpak, Pedro Carbajal-Madrid, Denise Likar, Gretchen A. Brown, and Romilla Batra. Reducing barriers to mental health care: Bringing evidence-based psychotherapy home. *Journal of the American Geriatrics Society*, 67:2174–2179, 10 2019. ISSN 15325415. doi: 10.1111/jgs.16088.

[18] Jacob T. Browne. Wizard of oz prototyping for machine learning experiences. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Chi Ea '19, page 1–6, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359719. doi: 10.1145/3290607.3312877. URL `https://doi-org.tudelft.idm.oclc.org/10.1145/3290607.3312877`.

[19] Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. Diet: Lightweight language understanding for dialogue systems, 2020. URL `https://arxiv.org/abs/2004.09936`.

[20] Pooja Chandrashekar. Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. *mHealth*, 4:6–6, 3 2018. doi: 10.21037/mhealth.2018.03.02.

[21] Minseok Cho, Gyeongbok Lee, and Seung-won Hwang. Explanatory and actionable debugging for machine learning: A tableqa demonstration. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sigir'19, page 1333–1336, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331404.

[22] Kijung Choi, Ying Wang, and Beverley Sparks. Travel app users' continued use intentions: it's a matter of value and trust. *Journal of Travel & Tourism Marketing*, 36(1):131–143, 2019. doi: 10.1080/10548408.2018.1505580.

[23] Helen Christensen, Kathleen M Griffiths, and Louise Farrer. Adherence in internet interventions for anxiety and depression: Systematic review. *J Med Internet Res*, 11(2):e13, Apr 2009. doi: 10.2196/jmir.1194.

[24] Susannah C. Coaston. Self-care through self-compassion: A balm for burnout. *The Professional Counselor*, 7:285–297, 10 2017. ISSN 2164-3989. doi: 10.15241/scc.7.3.285.

[25] Kevin Corti and Alex Gillespie. Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior*, 58:431–442, 2016. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2015.12.039.

[26] Rodrigo Marques da Silva, Carolina Tonini Goulart, Luis Felipe Dias Lopes, Patrícia Maria Serrano, Ana Lucia Siqueira Costa, and Laura

de Azevedo Guido. Hardy personality and burnout syndrome among nursing students in three brazilian universities–an analytic study. *BMC nursing*, 13(1):1–6, 2014. doi: 10.1186/1472-6955-13-9.

[27] Evangelia Demerouti, Friedhelm Nachreiner, Arnold B. Bakker, and Wilmar B. Schaufeli. The job demands-resources model of burnout. *Journal of Applied Psychology*, 86:499–512, 2001. ISSN 00219010. doi: 10.1037/0021-9010.86.3.499.

[28] Evangelia Demerouti, Evangelia Demerouti, Arnold B. Bakker, Ioanna Vardakou, and Aristotelis Kantas. The convergent validity of two burnout instruments. *European Journal of Psychological Assessment*, 19(1):12–23, 2003. doi: 10.1027//1015-5759.19.1.12.

[29] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64:1155–1170, 3 2018. ISSN 15265501. doi: 10.1287/mnsc.2016.2643.

[30] JM Dopmeijer, J Nuijen, MCM Busch, NI Tak, and A Verweij. Monitor mentale gezondheid en middelengebruik studenten hoger onderwijs. deelrapport i. mentale gezondheid van studenten in het hoger onderwijs, 11 2021. URL `https://www.rivm.nl/en/news/concerns-about-mental-health-and-substance-use-among-students-justified`.

[31] Mirjam Ekstedt, Marie Söderström, Torbjörn Åkerstedt, Jens Nilsson, Hans-Peter Søndergaard, and Perski Aleksander. Disturbed sleep and fatigue in occupational burnout. *Scandinavian Journal of Work, Environment & Health*, 2:121–131, 4 2006. ISSN 0355-3140. doi: 10.5271/sjweh.987.

[32] Albert Ellis. Rational psychotherapy and individual psychology. *Journal of individual psychology*, 13(1):38, 1957.

[33] Albert Ellis. *Reason and emotion in psychotherapy.* Citadel Press, 1962.

[34] Adriana Espinosa and Selma Kadić-Maglajlić. The role of health consciousness, patient–physician trust, and perceived physician's emotional appraisal on medical adherence. *Health Education & Behavior*, 46(6):991–1000, 2019. doi: 10.1177/1090198119859407.

[35] Gunther Eysenbach. The law of attrition. *J Med Internet Res*, 7(1):e11, Mar 2005. doi: 10.2196/jmir.7.1.e11.

[36] Ahmed Fadhil, Gianluca Schiavo, Yunlong Wang, and Bereket A. Yilma. The effect of emojis when interacting with conversational interface assisted health coaching system. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '18, page

378–383, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450364508. doi: 10.1145/3240925.3240965.

[37] Barry Alan Farber. *Self-disclosure in psychotherapy.* Guilford Press, 2006.

[38] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007.

[39] Leon Fletcher-Tomenius and Andreas Vossler. Trust in online therapeutic relationships: The therapist's experience. *Counselling Psychology Review*, 24 (2):24–34, 2009. URL `http://oro.open.ac.uk/17204/`.

[40] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. What makes users trust a chatbot for customer service? an exploratory interview study. In *Internet Science*, volume 11193 Lncs, pages 194–208, Cham, 2018. Springer Verlag. ISBN 9783030014360. doi: 10.1007/978-3-030-01437-7_16.

[41] Ohio Center for Autism and Low Incidence. Stress thermometer. URL `https://www.ocali.org/project/resource_gallery_of_interventions/page/StressThermometer`.

[42] Herbert J. Freudenberger. Staff burn-out. *Journal of Social Issues*, 30(1): 159–165, 1974. doi: 10.1111/j.1540-4560.1974.tb00706.x.

[43] Qianrao Fu, Herbert Hoijtink, and Mirjam Moerbeek. Sample-size determination for the bayesian t test and welch's test using the approximate adjusted fractional bayes factor. *Behavior Research Methods*, 53(1):139–152, 2021.

[44] Maria José Quina Galdino, Júlia Trevisan Martins, Maria do Carmo Fernandez Lourenço Haddad, Maria Lucia do Carmo Cruz Robazzi, and Marcela Maria Birolim. Burnout syndrome among master's and doctoral students in nursing. *Acta Paulista de Enfermagem*, 29:100–106, 2016.

[45] Shuqing Gao, Lingnan He, Yue Chen, Dan Li, and Kaisheng Lai. Public perception of artificial intelligence in medical care: Content analysis of social media. *J Med Internet Res*, 22(7):e16649, Jul 2020. ISSN 1438-8871. doi: 10.2196/16649.

[46] Franca Garzotto, Fabio Sorce, Davide Bolchini, and Tao Yang. Empirical investigation of web design attributes affecting brand perception. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pages 188–197, 2010.

[47] David Gefen, Elena Karahanna, and Detmar W. Straub. Trust and tam in online shopping: An integrated model. *Mis Q.*, 27:51–90, 2003. doi: 10.2307/30036519.

[48] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable active learning (xal). *Proceedings of the ACM on Human-Computer Interaction*, 4:1–28, 1 2021. doi: 10.1145/3432934.

[49] William J. Gibson and Andrew Brown. *Working with Qualitative Data*. SAGE, 2009. ISBN 9781412945653.

[50] Google. Feedback + control, . URL `https://pair.withgoogle.com/chapter/feedback-controls/`.

[51] Google. Explainability + trust, . URL `https://pair.withgoogle.com/chapter/explainability-trust/`.

[52] Siddharth Gulati, Sonia Sousa, and David Lamas. Design, development and evaluation of a human-computer trust scale. *Behaviour and Information Technology*, 38:1004–1015, 10 2019. ISSN 13623001. doi: 10.1080/0144929x.2019.1656779.

[53] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai&#x2014;explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, 2019. doi: 10.1126/scirobotics.aay7120.

[54] Robert S. Gutzwiller and John Reeder. Dancing with algorithms: Interaction creates greater preference and trust in machine-learned behavior. *Human Factors*, 63(5):854–867, 2021. doi: 10.1177/0018720820903893.

[55] Curt Hagquist and Magnus Stenbeck. Goodness of fit in regression analysis–r 2 and g 2 reconsidered. *Quality and Quantity*, 32(3):229–245, 1998.

[56] Steven C. Hayes. Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. *Behavior Therapy*, 35(4):639–665, 2004. ISSN 0005-7894. doi: 10.1016/s0005-7894(04)80013-3.

[57] Yulan He. Goal detection from natural language queries. *International Conference on Application of Natural Language to Information Systems*, pages 157–168, 2010. doi: 10.1007/978-3-642-13881-2_16. URL `http://link.springer.com/10.1007/978-3-642-13881-2%5F16`.

[58] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. Trust in automation. *IEEE Intelligent Systems*, 28:84–88, 2013. ISSN 15411672. doi: 10.1109/mis.2013.24.

[59] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 63–72, 2020.

[60] Adam O. Horvath. The alliance in context: Accomplishments, challenges, and future directions. *Psychotherapy: Theory, Research, Practice, Training*, 43: 258–263, 9 2006. ISSN 1939-1536. doi: 10.1037/0033-3204.43.3.258. URL `http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-3204.43.3.258`.

[61] Wayne K. Hoy and Megan Tschannen-Moran. Five faces of trust: An empirical confirmation in urban elementary schools. *Journal of School Leadership*, 9(3): 184–208, 1999. doi: 10.1177/105268469900900301.

[62] Ines Hungerbuehler, Kate Daley, Kate Cavanagh, Heloísa Garcia Claro, and Michael Kapps. Chatbot-based assessment of employees' mental health: Design process and pilot implementation. *JMIR Formative Research*, 5, 4 2021. ISSN 2561326x. doi: 10.2196/21678.

[63] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 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. ISSN 22915222. doi: 10.2196/12106.

[64] Iso. Ergonomics of human-system interaction-usability methods supporting human-centred design, 6 2002. URL `https://www.iso.org/standard/31176.html`.

[65] Iso. Ergonomics of human-system interaction-human-centred design for interactive systems, 7 2019. URL `https://www.iso.org/standard/77520.html`.

[66] JASP Team. JASP (Version 0.16.2)[Computer software], 2022. URL `https://jasp-stats.org/`.

[67] Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

[68] Ravinder Jerath, Molly W. Crawford, Vernon A. Barnes, and Kyler Harden. Self-regulation of breathing as a primary treatment for anxiety. *Applied Psychophysiology and Biofeedback*, 40:107–115, 6 2015. ISSN 1090-0586. doi: 10.1007/s10484-015-9279-8. URL `http://link.springer.com/10.1007/s10484-015-9279-8`.

[69] O.P. John and S. Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality: Theory and Research*, volume 2. Guilford Press, 1999.

[70] Jon Kabat-Zinn. Mindfulness-based interventions in context: Past, present, and future. *Clinical Psychology: Science and Practice*, 10(2):144–156, 2003. doi: 10.1093/clipsy.bpg016.

[71] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. *Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality*, page 1–12. Association for Computing Machinery,

New York, NY, USA, 2019. ISBN 9781450359702. doi: 10.1145/3290605.3300316.

[72] Youjeong Kim and S. Shyam Sundar. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, 28:241–250, 1 2012. ISSN 0747-5632. doi: 10.1016/j.chb.2011.09.006.

[73] Sanna M. Kinnunen, Anne Puolakanaho, Anne Mäkikangas, Asko Tolvanen, and Raimo Lappalainen. Does a mindfulness-, acceptance-, and value-based intervention for burnout have long-term effects on different levels of subjective well-being? *International Journal of Stress Management*, 27:82–87, 2020. ISSN 10725245. doi: 10.1037/str0000132.

[74] Dieter Korczak, Monika Wastian, and Michael Schneider. Therapy of the burnout syndrome. *GMS health technology assessment*, 8:Doc05, 2012. ISSN 1861-8863. doi: 10.3205/hta000103.

[75] M. A. Korsgaard, David M. Schweiger, and Harry J. Sapienza. Building commitment, attachment, and trust in strategic decision-making teams: The role of procedural justice. *Academy of Management Journal*, 38(1):60, 02 1995. doi: 10.5465/256728.

[76] Tobias Kowatsch, M. Nißen, Dominik Rüegger, and Mirjam Stieger. The impact of interpersonal closeness cues in text-based healthcare chatbots on attachment bond and the desire to continue interacting: an experimental design. In *Ecis*, 2018.

[77] Tage S Kristensen, Marianne Borritz, Ebbe Villadsen, and Karl B Christensen. The copenhagen burnout inventory: A new tool for the assessment of burnout. *Work & Stress*, 19(3):192–207, 2005. doi: 10.1080/02678370500297720.

[78] John K Kruschke and Torrin M Liddell. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychon Bull Rev*, 25:178–206, 2018. doi: 10.3758/s13423-016-1221-4.

[79] Robert L. Leahy. *Cognitive therapy: Basic principles and applications.* Rowman & Littlefield Pub., 2004. doi: 10.1891/0889-8391.14.1.106.

[80] Dongwon Lee, Junghoon Moon, Yong Jin Kim, and Mun Y. Yi. Antecedents and consequences of mobile phone usability. *Inf. Manage.*, 52(3):295–304, 4 2015. ISSN 0378-7206. doi: 10.1016/j.im.2014.12.001.

[81] John D. Lee and Neville Moray. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40(1):153–184, 1994. ISSN 1071-5819. doi: https://doi.org/10.1006/ijhc.1994.1007. URL `https://www.sciencedirect.com/science/article/pii/S107158198471007X`.

[82] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392.

[83] Yi Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. ”i hear you, i feel you”: Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–12. Association for Computing Machinery, 4 2020. ISBN 9781450367080. doi: 10.1145/3313831.3376175.

[84] Yin-Yang Lee and Julia L. Lin. The effects of trust in physician on self-efficacy, adherence and diabetes outcomes. *Social Science & Medicine*, 68(6): 1060–1068, 2009. ISSN 0277-9536. doi: https://doi.org/10.1016/j.socscimed.2008.12.033. URL `https://www.sciencedirect.com/science/article/pii/S0277953608006734`.

[85] Michael P Leiter and Christina Maslach. Interventions to prevent and alleviate burnout. In *Current Issues in Work and Organizational Psychology*, pages 32–50. Routledge, 2018.

[86] Michael P. Leiter and Wilmar B. Schaufeli. Consistency of the burnout construct across occupations. *Anxiety, Stress & Coping*, 9:229–243, 1 1996. ISSN 1061-5806. doi: 10.1080/10615809608249404.

[87] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. Umux-lite: When there’s no time for the sus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’13, page 2099–2102, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2481287. URL `https://doi.org/10.1145/2470654.2481287`.

[88] C. Liebrecht, Lena Sander, and C.M.J. van Hooijdonk. Too informal? how a chatbot’s communication style affects brand attitude and quality of interaction. In *Conversations 2020 : 4th international workshop on chatbot research*, 11 2020. URL `https://conversations2020.wordpress.com/`.

[89] G. Lietaer, J. Rombauts, R. Van Balen, and G.T. Barrett-Lennard. *The therapy pathway reformulated*, page 123–153. Leuven University Press, 1990.

[90] Joon Soo Lim and Ghee-Young Noh. Effects of gain-versus loss-framed performance feedback on the use of fitness apps: Mediating role of exercise self-efficacy and outcome expectations of exercise. *Computers in Human Behavior*, 77:249–257, 2017. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2017.09.006. URL `https://www.sciencedirect.com/science/article/pii/S0747563217305320`.

[91] Shu Hui Lin and Yun Chen Huang. Life stress and academic burnout. *Active Learning in Higher Education*, 15:77–90, 3 2014. ISSN 14697874. doi: 10.1177/1469787413514651.

[92] Sarah Linder, Geri Miller, and Paula Johnson. Counseling and spirituality: The use of emptiness and the importance of timing. In *Annual Conference of the American Counseling Association*, 3 2000.

[93] Sarah Ketchen Lipson, Emily G. Lattie, and Daniel Eisenberg. Increased rates of mental health service utilization by u.s. college students: 10-year population-level trends (2007-2017). *Psychiatric Services*, 70:60–63, 1 2019. ISSN 15579700. doi: 10.1176/appi.ps.201800332.

[94] Shuang Liu and Jessica F. Willoughby. Do fitness apps need text reminders? an experiment testing goal-setting text message reminders to promote self-monitoring. *Journal of Health Communication*, 23(4):379–386, 2018. doi: 10.1080/10810730.2018.1455768.

[95] Edwin A. Locke and Gary P. Latham. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57:705–717, 9 2002. ISSN 1935-990x. doi: 10.1037/0003-066x.57.9.705. URL `http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.57.9.705`.

[96] V A Lofthouse and D Lilley. What they really, really want: user centered research methods for design. *DS 36: Proceedings DESIGN 2006, the 9th International Design Conference*, 2006.

[97] Kien Hoa Ly, Anna Trüschel, Linnea Jarl, Susanna Magnusson, Tove Windahl, Robert Johansson, Per Carlbring, and Gerhard Andersson. Behavioural activation versus mindfulness-based guided self-help treatment administered through a smartphone application: a randomised controlled trial. *BMJ Open*, 4:3440, 2014. doi: 10.1136/bmjopen-2013.

[98] Kien Hoa Ly, Naira Topooco, Hanna Cederlund, Anna Wallin, Jan Bergstrom, Olof Molander, Per Carlbring, and Gerhard Andersson. Smartphone-supported versus full behavioural activation for depression: A randomised controlled trial. *PLoS ONE*, 10, 5 2015. ISSN 19326203. doi: 10.1371/journal.pone.0126559.

[99] Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. A fully automated conversational agent for promoting mental well-being: A pilot rct using mixed methods. *Internet Interventions*, 10:39–46, 2017. ISSN 2214-7829. doi: https://doi.org/10.1016/j.invent.2017.10.002. URL `https://www.sciencedirect.com/science/article/pii/S221478291730091X`.

[100] Dutch MacDonald, H.R. Shiever, Nancy Rekhelman, Rehaab Raza, Phil Gerrard, and David Heacock. Human-centered design is more important than

ever, 7 2021. URL `https://www.bcg.com/publications/2020/the-importance-of-human-centered-design`.

[101] James E. Maddux. *Self-Efficacy Theory*, pages 3–33. Springer US, Boston, MA, 1995. ISBN 978-1-4419-6868-5. doi: 10.1007/978-1-4419-6868-5_1.

[102] Leilani Madrigal and Anastasia Blevins. I hate it, it's ruining my life: College students' early academic year experiences during the covid-19 pandemic. *Traumatology*, 8 2021. ISSN 1085-9373. doi: 10.1037/trm0000336.

[103] François Mairesse and Marilyn A. Walker. Towards personality-based user adaptation: Psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20:227–278, 8 2010. ISSN 09241868. doi: 10.1007/s11257-010-9076-2.

[104] Mady Mantha. Introducing diet: State-of-the-art architecture that outperforms fine-tuning bert and is 6x faster to train, Aug 2021. URL `https://rasa.com/blog/introducing-dual-intent-and-entity-transformer-diet-state-of-the-art-performance-on-a-lightweight-architecture/`.

[105] Sarah Marsh. Number of university dropouts due to mental health problems trebles, 5 2017. URL `https://www.theguardian.com/society/2017/may/23/number-university-dropouts-due-to-mental-health-problems-trebles`.

[106] Christina Maslach and Susan E Jackson. The measurement of experienced burnout, 1981.

[107] Christina Maslach, Susan Jackson, and Michael Leiter. *The Maslach Burnout Inventory Manual*. Consulting Psychologists Press, 1997.

[108] Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3): 709–734, 1995. doi: 10.2307/258792.

[109] Trevor Mazzucchelli, Robert Kane, and Clare Rees. Behavioral activation treatments for depression in adults: A meta-analysis and review. *Clinical Psychology: Science and Practice*, 16(4):383–411, 2009. doi: 10.1111/j.1468-2850.2009.01178.x.

[110] Mary Mccarthy, Grace Pretty, and Vic Catano. Psychological sense of community and burnout. *Journal of College Student Development*, 31:211–216, 05 1990.

[111] Fiona H. McKay, Christina Cheng, Annemarie Wright, Jane Shill, Hugh Stephens, and Mary Uccellini. Evaluating mobile phone applications for health behaviour change: A systematic review. *Journal of Telemedicine and Telecare*, 24:22–30, 1 2018. ISSN 17581109. doi: 10.1177/1357633x16673538.

[112] D. Harrison McKnight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manag. Inf. Syst.*, 2:12:1–12:25, 2011. doi: 10.1145/1985347.1985353.

[113] Margaret McLafferty, Coral R. Lapsley, Edel Ennis, Cherie Armour, Sam Murphy, Brendan P. Bunting, Anthony J. Bjourson, Elaine K. Murray, and Siobhan M. Oâ€™Neill. Mental health, behavioural problems and treatment seeking among students commencing university in northern ireland. *PLoS ONE*, 12, 12 2017. ISSN 19326203. doi: 10.1371/journal.pone.0188785.

[114] Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'Neill. How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians. In Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler, editors, *Human-Computer Interaction - INTERACT 2017*, pages 441–459, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67744-6.

[115] David C. Mohr, Pim Cuijpers, and Kenneth Lehman. Supportive accountability: A model for providing human support to enhance adherence to ehealth interventions. *Journal of Medical Internet Research*, 13, 2011. ISSN 14388871. doi: 10.2196/jmir.1602.

[116] Paula M.C. Mommersteeg, Cobi J. Heijnen, Marc J.P.M. Verbraak, and Lorenz J.P. van Doornen. A longitudinal study on cortisol and complaint reduction in burnout. *Psychoneuroendocrinology*, 31(7):793–804, 2006. ISSN 0306-4530. doi: https://doi.org/10.1016/j.psyneuen.2006.03.003.

[117] Lea Müller, Jens Mattke, Christian Maier, Tim Weitzel, and Heinrich Graser. Chatbot acceptance: A latent profile analysis on individuals' trust in conversational agents. In *Proceedings of the 2019 on Computers and People Research Conference*, SIGMIS-CPR '19, pages 35–42. Association for Computing Machinery, Inc, 6 2019. ISBN 9781450360883. doi: 10.1145/3322385.3322392.

[118] Nature. Being a phd student shouldn't be bad for your health. *Nature*, 569: 307–307, 5 2019. ISSN 0028-0836. doi: 10.1038/d41586-019-01492-0.

[119] Geoffrey C. Nguyen, Thomas A. LaVeist, Mary L. Harris, Lisa W. Datta, Theodore M. Bayless, and Steven R. Brant. Patient trust-in-physician and race are predictors of adherence to medical management in inflammatory bowel disease. *Inflammatory Bowel Diseases*, 15(8):1233–1239, 01 2009. ISSN 1078-0998. doi: 10.1002/ibd.20883.

[120] Jennifer Nicholas, Kit Huckvale, Mark Erik Larsen, Ashna Basu, Philip J Batterham, Frances Shaw, and Shahbaz Sendi. Issues for ehealth in psychiatry: Results of an expert survey. *J Med Internet Res*, 19(2):e55, 2 2017. ISSN 1438-8871. doi: 10.2196/jmir.6957.

[121] Samantha K Norelli, Ashley Long, and Jeffrey M Krepps. Relaxation techniques. In *StatPearls [Internet]*. StatPearls Publishing, 2020.

[122] Michael I Norton, Daniel Mochon, and Dan Ariely. The ikea effect: When labor leads to love. *Journal of consumer psychology*, 22(3):453–460, 2012.

[123] Hanne Weie Oddli, John Mcleod, Sissel Reichelt, and Michael Helge Rønnestad. Strategies used by experienced therapists to explore client goals in early sessions of psychotherapy. *European Journal of Psychotherapy & Counselling*, 2014. ISSN 1469-5901. doi: 10.1080/13642537.2014.927380. URL `https://www.tandfonline.com/action/journalInformation?journalCode=rejp20`.

[124] Harri Oinas-Kukkonen. A foundation for the study of behavior change support systems. *Personal and Ubiquitous Computing*, 17:1223–1235, 8 2013. ISSN 16174909. doi: 10.1007/s00779-012-0591-5.

[125] Pawelczyk and Joanna. *Talk as Therapy*. Walter De Gruyter Inc, 2011. ISBN 1934078662.

[126] Olga Perski, David Crane, Emma Beard, and Jamie Brown. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? an experimental study. *Digital Health*, 5, 2019. ISSN 20552076. doi: 10.1177/2055207619880676.

[127] Aileen M. Pidgeon, Natasha F. Rowe, Peta Stapleton, Heidi B. Magyar, and Barbara C. Y. Lo. Examining characteristics of resilience among university students: An international study. *Open Journal of Social Sciences*, 02:14–22, 2014. ISSN 2327-5952. doi: 10.4236/jss.2014.211003.

[128] Jacqueline Pitchforth, Katie Fahy, Tamsin Ford, Miranda Wolpert, Russell M. Viner, and Dougal S. Hargreaves. Mental health and well-being trends among children and young people in the uk, 1995-2014: Analysis of repeated cross-sectional national health surveys. *Psychological Medicine*, 49:1275–1285, 6 2019. ISSN 14698978. doi: 10.1017/s0033291718001757.

[129] Pnn. Persbericht: Bijna helft promovendi heeft vergroot risico op mentale klachten, 40% overweegt te stoppen, 8 2020. URL `https://hetpnn.nl/2020/08/26/persbericht-bijna-helft-promovendi-heeft-vergroot-risico-op-mentale-klachten-40-overweegt-te-stoppen/`.

[130] Catherine Pricilla, Dessi Puji Lestari, and Dody Dharma. Designing interaction for chatbot-based conversational commerce with user-centered design. In *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 244–249, 2018. doi: 10.1109/ICAICTA.2018.8541320.

[131] Judith J Prochaska, Erin A Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study. *Journal of Medical Internet Research*, 23(3):e24850, 2021.

[132] Daniel S. Quintana and Donald R. Williams. Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using jasp. *BMC Psychiatry*, 18, 6 2018. ISSN 1471244X. doi: 10.1186/s12888-018-1761-4.

[133] S. Rachman. The evolution of behaviour therapy and cognitive behaviour therapy. *Behaviour Research and Therapy*, 64:1–8, 2015. ISSN 0005-7967. doi: 10.1016/j.brat.2014.10.006.

[134] Zeinab Rahmati. The study of academic burnout in students with high and low level of self-efficacy. *Procedia - Social and Behavioral Sciences*, 171:49–55, 1 2015. ISSN 18770428. doi: 10.1016/j.sbspro.2015.01.087.

[135] Nikita B Rajani, Nikolaos Mastellos, and Filippos T Filippidis. Self-efficacy and motivation to quit of smokers seeking to quit: Quantitative assessment of smoking cessation mobile apps. *JMIR Mhealth Uhealth*, 9(4):e25030, 4 2021. ISSN 2291-5222. doi: 10.2196/25030.

[136] Rasa. Why rasa?, Sep 2021. URL `https://rasa.com/product/why-rasa/`.

[137] Abhilasha Ravichander and Alan W. Black. An empirical study of self-disclosure in spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–263, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5030.

[138] Dorota Reis, Despoina Xanthopoulou, and Ioannis Tsaousis. Measuring job and academic burnout with the oldenburg burnout inventory (olbi): Factorial invariance across samples and countries. *Burnout Research*, 2:8–18, 2015. ISSN 22130586. doi: 10.1016/j.burn.2014.11.001.

[139] Carl R. Rogers. *On becoming a person: A therapist's view of psychotherapy.* Houghton Mifflin, 1961.

[140] Carl Ransom Rogers et al. *A theory of therapy, personality, and interpersonal relationships: As developed in the client-centered framework*, volume 3. McGraw-Hill New York, 1959.

[141] Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832 – 837, 1956. doi: 10.1214/aoms/1177728190. URL `https://doi.org/10.1214/aoms/1177728190`.

[142] Jeffrey N Rouder, Paul L Speckman, Dongchu Sun, Richard D Morey, and Geoffrey Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, April 2009.

[143] Fredrik Santoft, Sigrid Salomonsson, Hugo Hesser, Elin Lindsäter, Brjánn Ljótsson, Mats Lekander, Göran Kecklund, Lars Göran Öst, and Erik Hedman-Lagerlöf. Mediators of change in cognitive behavior therapy for clinical burnout. *Behavior Therapy*, 50:475–488, 5 2019. ISSN 0005-7894. doi: 10.1016/j.beth.2018.08.005.

[144] Wilmar B. Schaufeli and Toon W. Taris. The conceptualization and measurement of burnout: Common ground and worlds apart. *Work & Stress*, 19: 256–262, 7 2005. ISSN 0267-8373. doi: 10.1080/02678370500385913.

[145] Wilmar B. Schaufeli, Isabel M. Martínez, Alexandra Marques Pinto, Marisa Salanova, and Arnold B. Bakker. Burnout and engagement in university students. *Journal of Cross-Cultural Psychology*, 33:464–481, 9 2002. ISSN 0022-0221. doi: 10.1177/0022022102033005003.

[146] Wilmar B. Schaufeli, Michael P. Leiter, and Christina Maslach. Burnout: 35 years of research and practice. *Career Development International*, 14:204–220, 6 2009. ISSN 13620436. doi: 10.1108/13620430910966406.

[147] Astrid Schepman and Paul Rodway. Initial validation of the general attitudes towards artificial intelligence scale. *Computers in Human Behavior Reports*, 1:100014, 2020. ISSN 2451-9588. doi: https://doi.org/10.1016/j.chbr.2020.100014.

[148] Xenia Schmalz, José B Manresa, and Lei Zhang. What is a bayes factor?, Sep 2020. URL `osf.io/vgqbt`. Preprint.

[149] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans Georg Luigs, Anne Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2:476–486, 8 2020. ISSN 25225839. doi: 10.1038/s42256-020-0212-3.

[150] Thomas Schäfer and Marcus A. Schwarz. The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.00813. URL `https://www.frontiersin.org/article/10.3389/fpsyg.2019.00813`.

[151] David W Scott. *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons, 1992.

[152] Zindel V Segal, J Mark G Williams, and John D Teasdale. *Mindfulness-based cognitive therapy for depression: A new approach to preventing relapse.* Guilford Press, 2002. ISBN 1-57230-706-4 (Hardcover).

[153] Aparna Sharma. 7 ways to increase trust for your chatbot, 11 2018. URL `https://chatbotslife.com/7-ways-to-increase-trust-for-your-chatbot-19f7be70ead8`.

[154] Heung-Yeung Shum, Xiaodong He, and Di Li. From eliza to xiaoice: Challenges and opportunities with social chatbots. *CoRR*, abs/1801.01957, 1 2018. URL `http://arxiv.org/abs/1801.01957`.

[155] Tuva Lunde Smestad and Frode Volden. Chatbot personalities matters. In *International Conference on Internet Science*, pages 170–181. Springer, 2018. doi: 10.1007/978-3-030-17705-8_15.

[156] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. pages 293–304. Association for Computing Machinery, 3 2018. ISBN 9781450349451. doi: 10.1145/3172944.3172965.

[157] Matthias Söllner, Axel Hoffmann, Holger Hoffmann, and Jan Marco Leimeister. How to use behavioral research insights on trust for hci system design. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, page 1703–1708, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310161. doi: 10.1145/2212776.2223696.

[158] Changwon Son, Sudeep Hegde, Alec Smith, Xiaomei Wang, and Farzan Sasangohar. Effects of covid-19 on college students' mental health in the united states: Interview survey study. *Journal of Medical Internet Research*, 22, 9 2020. ISSN 14388871. doi: 10.2196/21279.

[159] Miguel A. Sorrel, Jose Angel Martinez-Huertas, and Maria Arconada. It must have been burnout: Prevalence and related factors among spanish phd students. *Spanish Journal of Psychology*, 23, 2020. ISSN 19882904. doi: 10.1017/sjp.2020.31.

[160] Helen M. Stallman. Psychological distress in university students: A comparison with general population data. *Australian Psychologist*, 45(4):249–257, 2010. doi: 10.1080/00050067.2010.482109.

[161] Zlatko Stapić, Ana Horvat, and Dijana Plantak Vukovac. Designing a faculty chatbot through user-centered design approach. In Constantine Stephanidis, Don Harris, Wen-Chin Li, Dylan D. Schmorrow, Cali M. Fidopiastis, Panayiotis Zaphiris, Andri Ioannou, Xiaowen Fang, Robert A. Sottilare, and Jessica Schwarz, editors, *HCI International 2020 – Late Breaking Papers: Cognition, Learning and Games*, pages 472–484, Cham, 2020. Springer International Publishing. ISBN 978-3-030-60128-7.

[162] Joachim Stoeber, Julian H. Childs, Jennifer A. Hayward, and Alexandra R. Feast. Passion and motivation for studying: Predicting academic engagement and burnout in university students. *Educational Psychology*, 31:513–528, 7 2011. ISSN 01443410. doi: 10.1080/01443410.2011.570251.

[163] Simone Stumpf, Erin Sullivan, Erin Fitzhenry, Ian Oberst, Weng-Keen Wong, and Margaret Burnett. Integrating rich user feedback into intelligent user interfaces. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, Iui '08, page 50–59, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781595939876. doi: 10.1145/1378773.1378781. URL `https://doi-org.tudelft.idm.oclc.org/10.1145/1378773.1378781`.

[164] Sara Tement, Saša Zorjan, Meta Lavrič, Vita Poštuvan, and Nejc Plohl. A randomized controlled trial to improve psychological detachment from work and well-being among employees: a study protocol comparing online cbt-based and mindfulness interventions. *BMC Public Health*, 20, 2020. doi: 10.1186/isrctn98347361.

[165] Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 239–245. Association for Computing Machinery, Inc, 1 2019. ISBN 9781450363242. doi: 10.1145/3306618.3314293.

[166] John Torous, Steven Richard Chan, Shih Yee Marie Tan, Jacob Behrens, Ian Mathew, Erich J. Conrad, Ladson Hinton, Peter Yellowlees, and Matcheri Keshavan. Patient smartphone ownership and interest in mobile apps to monitor symptoms of mental health conditions: A survey in four geographically distinct psychiatric clinics. *JMIR Mental Health*, 1, 7 2014. ISSN 23687959. doi: 10.2196/mental.4004.

[167] Don van den Bergh, Merlise A Clyde, Akash Raj, Tim de Jong, Quentin F Gronau, Maarten Marsman, Alexander Ly, and Eric-Jan Wagenmakers. A tutorial on bayesian multi-model linear regression with bas and jasp, Apr 2020. URL `psyarxiv.com/pqju6`.

[168] KEES VAN DEN BOS, RIËL VERMUNT, and HENK A. M. WILKE. The consistency rule and the voice effect: the influence of expectations on procedural fairness judgements and performance. *European Journal of Social Psychology*, 26(3):411–428, 1996. doi: https://doi.org/10.1002/(SICI)1099-0992(199605)26:3<411::AID-EJSP766>3.0.CO;2-2.

[169] Gabriela Villalobos-Zuniga and Mauro Cherubini. Apps that motivate: a taxonomy of app features based on self-determination theory. *International Journal of Human-Computer Studies*, 140:102449, 8 2020. ISSN 10715819. doi: 10.1016/j.ijhcs.2020.102449. URL `https://linkinghub.elsevier.com/retrieve/pii/S1071581920300513`.

[170] Kai Wang, Deepthi S. Varma, and Mattia Prosperi. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. *Journal of Psychiatric Research*, 107:73–78, 12 2018. ISSN 18791379. doi: 10.1016/j.jpsychires.2018.10.006.

[171] Weiquan Wang and Izak Benbasat. Trust in and adoption of online recommendation agents ∗. *Journal of the Association for Information Systems*, 6: 73, 2005.

[172] Roger Watson, Ian Deary, David Thompson, and Gloria Li. A study of stress and burnout in nursing students in hong kong: a questionnaire survey. *International journal of nursing studies*, 45(10):1534–1542, 2008. doi: 10.1016/j.ijnurstu.2007.11.003.

[173] Yanxia Wei, Pinpin Zheng, Hui Deng, Xihui Wang, Xiaomei Li, and Hua Fu. Design features for improving mobile health intervention user engagement: Systematic review and thematic analysis. *Journal of Medical Internet Research*, 22, 12 2020. ISSN 14388871. doi: 10.2196/21687.

[174] Joseph Weizenbaum. Eliza–a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9:36–45, 1 1966. ISSN 0001-0782. doi: 10.1145/365153.365168. URL `https://dl.acm.org/doi/10.1145/365153.365168`.

[175] Matthew Whalley. *Psychology Tools For Overcoming Panic*. Psychology Tools, 2017.

[176] Alyse Wheelock, Christine Bechtel, and Bruce Leff. Human-centered design and trust in medicine. *Jama*, 324(23):2369–2370, 12 2020. ISSN 0098-7484. doi: 10.1001/jama.2020.21080.

[177] Wysa. Everyday mental health. URL `https://www.wysa.io/client-case-studies`.

[178] Yi Xie and Siqing Peng. How to repair customer trust after negative publicity: The roles of competence, integrity, benevolence, and forgiveness. *Psychology & Marketing*, 26(7):572–589, 2009.

[179] Guo Yu. *Effects of timing on users' perceived control when interacting with intelligent systems*. Cambridge University, 5 2018. doi: 10.17863/CAM.37907.

[180] Christina Zaprudskaya. Millennials, trust and chatbots., 6 2017. URL `https://chatbotsmagazine.com/millennials-trust-and-chatbots-23f883bac9ff`.

[181] Renwen Zhang, Jennifer Nicholas, Ashley A Knapp, Andrea K Graham, Elizabeth Gray, Mary J Kwasny, Madhu Reddy, and David C Mohr. Clinically meaningful use of mental health apps and its effects on depression: Mixed

methods study. *J Med Internet Res*, 21(12):e15644, 12 2019. ISSN 1438-8871. doi: 10.2196/15644.

Appendix A

List of Burnout Exercises

A.1 CBT

Table A.1 (borrowed in part from Tement *et al.* (2020) [164]).

Exercise	Description
ABC belief monitoring	A type of functional analysis developed by Ellis [32]. Participants explore activating events (A), beliefs (B), and consequences (C). A key part of cognitive theory.
Alternative Action Formulation	In this exercise, participants describe an event and their thoughts, emotions, and behaviours during it. They are then encouraged to consider alternative responses and how those would affect their thoughts and behaviours, helping participants develop better coping strategies [164].
Thought Distortion Record	Participants learn to identify automatic negative thoughts and to link event and cognition. Specifically, they learn to identify distortions such as jumping to conclusions, should statement, and mind-reading, increasing their awareness of their internal biases. Another essential CBT practice [13, 79].
Decatastrophizing	A cognitive restructuring challenge to catastrophic thinking. Participants identify a worry then rate its likelihood, they rationally consider worst case scenarios, and devise coping strategies. They also come up with reassurances regarding their worry [175, 33].
Fact or Opinion	A key practice to teach participants that their thoughts are not facts about the world but assumptions. Challenges negative thoughts [13].

Table A.1: Common CBT Exercises

A.2 MAV

Exercise	Description
Body Scan	The participant brings non-judgemental detailed awareness to each part of their body. Participants improve their ability to sustain focus over time, developing calmness [152].
Non-Judgmental Journaling	Improves self-compassion [24]. Participants are encouraged to use 'stream of consciousness' writing to express themselves freely [92].
Breathing Regulation	The active manipulation of breathing, in this case following the 4-7-8 pattern which is based on the pranayama. Helps engender a physical feeling of relexation [68].
The Observer	A key eyes-closed ACT exercise. The participant first becomes aware of present sensations, then is asked to recall and re-experience a past event. The exercise highlights the continuity of consciousness and divorces lived experience from sense of self [56].
Value Inventory	The participant is asked to write out their fundamental values, for example, attributes they would like to see in their own obituary. These values then drive participant goals and actions [56].

Table A.2: Common Mindfulness and MAV exercises

Appendix B

Requirement Gathering

This research investigates the existence of a gap between in-need academic staff and students and psychological assistance and the possibility of bridging this gap with a trustworthy intelligent mobile based personal assistant/coach.

B.1 Psychological Background

- In your experience do you find cases of burnout in your practice at TU Delft correlate with the expected numbers for this environment (20%), have those numbers changed significantly in the past year?
- What treatment or treatments do you find most effective for alleviating those symptoms? When do you consider a treatment effective?

B.2 Therapy Barriers

- Do you find often staff or student hesitancy to contact counsellors to be a barrier for seeking therapy? What causes have you found most often driving this hesitancy (finance, travel time, long waitlists, stigma, cultural background, medical history, etc.)
- What other barriers do you most encounter when treating patients with regards to administrative issues (under-staffing, treatment duration), or your personal capacity (burnout, emotional disconnect, difficulty establishing rapport)?

B.3 eHealth

- Have you used ehealth interventions before, what is your experience of them?

Follow up Would you be interested in evaluating the system developed through this research?

Appendix C

First Expert Review

C.1 Introduction

Thank you for taking part of this review as part of a larger study to develop a burnout identification and support system for postgraduate students. There are three points of interest in this study. In Section C.2 we discuss the burnout intervention tools collected throughout the literature review process and their suitability. In Section C.3 we discuss how goals are arrived at in therapist-client relationships. Lastly, in Section C.4 we review the language choices in the initial prototype.

C.2 Burnout Interventions

In Appendix A we list the intervention tools collected in our research. Based on this information, and in your experience, do you consider the listed tools to be appropriate for their purpose? Do you find that you know of more appropriate tools which are not listed?

Our intervention structure is borrowed from the from the clear design created by Addis *et al* (2004) [3], wherein the individual is tasked with:

1. Understanding how ’burnout’ works and how it is related to their behaviour and experiences.
2. Identifying areas of their life where their current coping mechanisms are maladaptive and learning how to change them.
3. Learning how to approach difficult situations rather than avoid them.
4. Addressing longer-term issues that may affect or be affecting their degree of burnout.

Do you find this to be an appropriate approach? How do you generally structure your intervention plans in your own practice?

C.3 Goal Solicitation & Alignment

The locus of user control in this study is the setting and alignment of user goals in regards to their intervention plan. Consider your so-called Session 0 with your clients, how do you engage new clients in discussing their goals, their desired outcomes, and what they want to change in their life? How do you decide together on goal priorities? How implicit/explicit do you find this process to be?

C.4 Prototype Design

In Appendix D you can see the screens used in the initial prototype. Based on this information we pose the following questions:

- In your experience, what features would you expect or want from a burnout identification and support systems.
- Which goal editing flow did you find preferable D.9, or D.11 and D.12.

In Appendix E you can see conversation snippets with chatbot M. Based on this information we pose the following questions:

- Personally, do you find student clients to prefer more or less formal language from their therapists?
- Do you find student preferred extroverted or introverted personality styles in their therapists?
- How do you find therapist self-disclosure plays a role in encouraging client self-disclosure and engagement?
- How do you find your conversation style differs in therapeutic discussions as opposed to day-to-day discourse.

Appendix D

Application Screens (First Prototype)

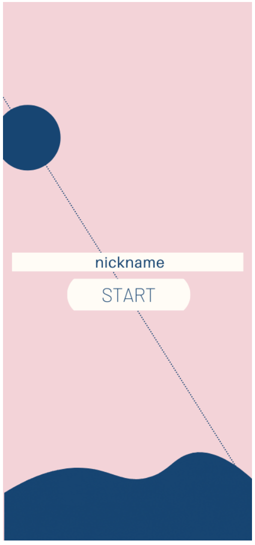
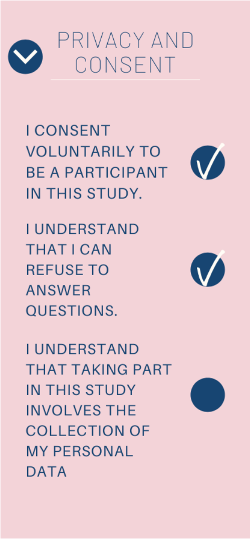


Figure D.1: Sign Up



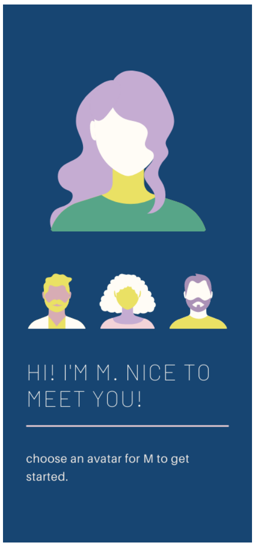
PRIVACY AND CONSENT

I CONSENT VOLUNTARILY TO BE A PARTICIPANT IN THIS STUDY. ✓

I UNDERSTAND THAT I CAN REFUSE TO ANSWER QUESTIONS. ✓

I UNDERSTAND THAT TAKING PART IN THIS STUDY INVOLVES THE COLLECTION OF MY PERSONAL DATA ●

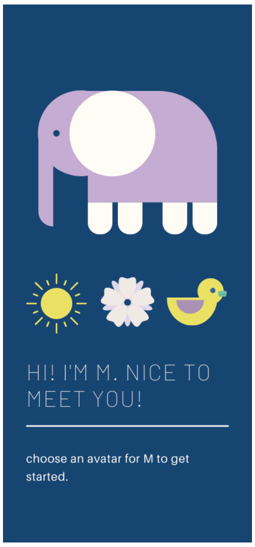
Figure D.2: Privacy and Data Use



HI! I'M M. NICE TO MEET YOU!

choose an avatar for M to get started.

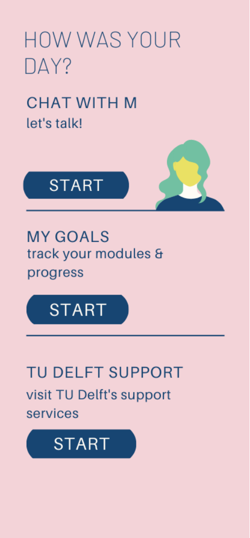
Figure D.3: Avatar Selection



HI! I'M M. NICE TO MEET YOU!

choose an avatar for M to get started.

Figure D.4: Avatar Selection Alt



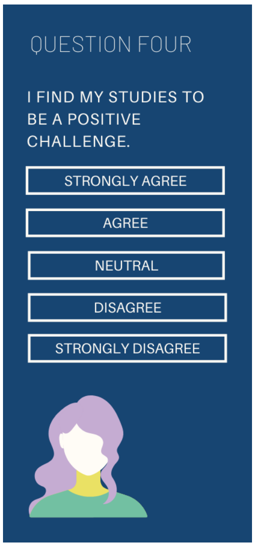
HOW WAS YOUR DAY?

CHAT WITH M
let's talk!

MY GOALS
track your modules & progress

TU DELFT SUPPORT
visit TU Delft's support services

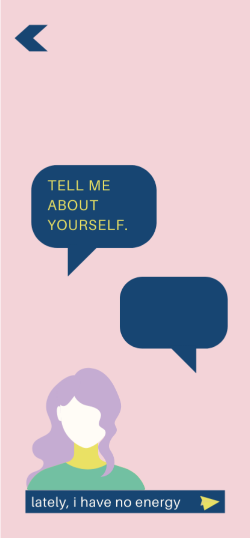
Figure D.5: Home Screen



QUESTION FOUR

I FIND MY STUDIES TO BE A POSITIVE CHALLENGE.

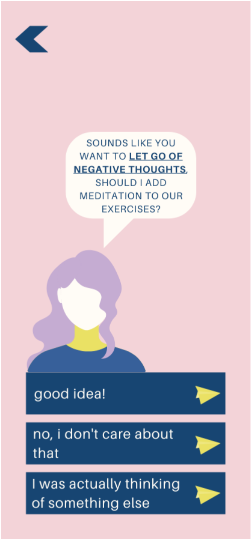
Figure D.6: OBLI Survey



TELL ME ABOUT YOURSELF.

lately, i have no energy

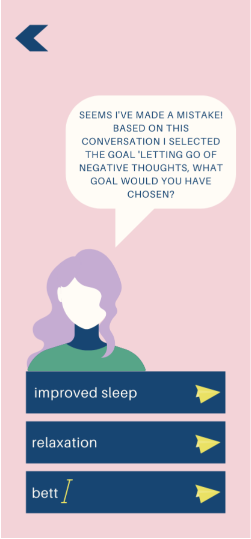
Figure D.7: Chatbot Interface



SOUNDS LIKE YOU WANT TO **LET GO OF NEGATIVE THOUGHTS**, SHOULD I ADD MEDITATION TO OUR EXERCISES?

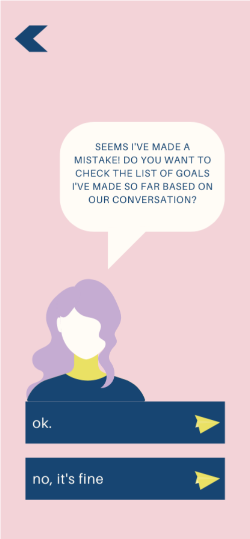
good idea!
no, i don't care about that
I was actually thinking of something else

Figure D.8: Plan Proposal



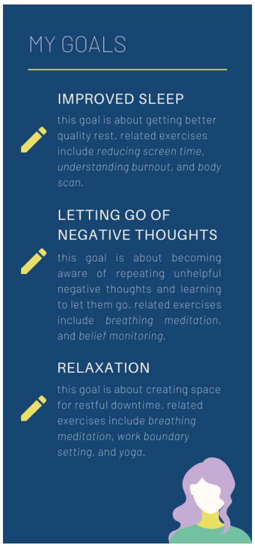
SEEMS I'VE MADE A MISTAKE! BASED ON THIS CONVERSATION I SELECTED THE GOAL 'LETTING GO OF NEGATIVE THOUGHTS, WHAT GOAL WOULD YOU HAVE CHOSEN?

Figure D.9: Goal Modification (A)



SEEMS I'VE MADE A MISTAKE! DO YOU WANT TO CHECK THE LIST OF GOALS I'VE MADE SO FAR BASED ON OUR CONVERSATION?

Figure D.10: Goal Modification (B)



MY GOALS

IMPROVED SLEEP

this goal is about getting better quality rest. related exercises include *reducing screen time*, *understanding burnout*, and *body scan*.

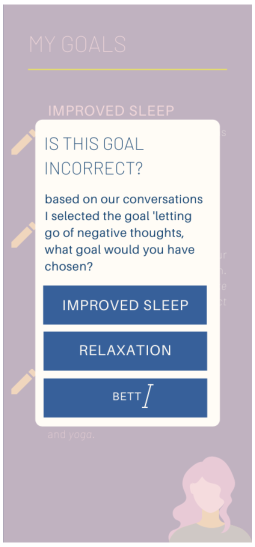
LETTING GO OF NEGATIVE THOUGHTS

this goal is about becoming aware of repeating unhelpful negative thoughts and learning to let them go. related exercises include *breathing meditation*, and *belief monitoring*.

RELAXATION

this goal is about creating space for restful downtime. related exercises include *breathing meditation*, *work boundary setting*, and *yoga*.

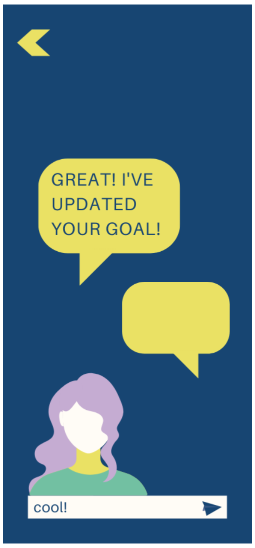
Figure D.11: Goal Modification (B2)



IS THIS GOAL INCORRECT?

based on our conversations I selected the goal 'letting go of negative thoughts, what goal would you have chosen?

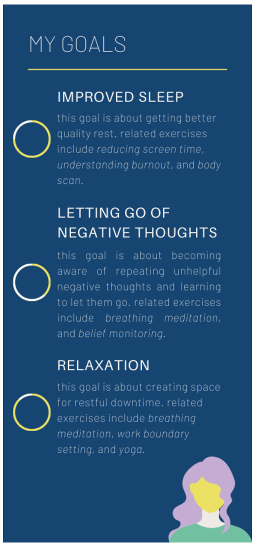
Figure D.12: Goal Modification (B3)



GREAT! I'VE UPDATED YOUR GOAL!

cool!

Figure D.13: Goal Modification Success



MY GOALS

IMPROVED SLEEP

this goal is about getting better quality rest. related exercises include *reducing screen time*, *understanding burnout*, and *body scan*.

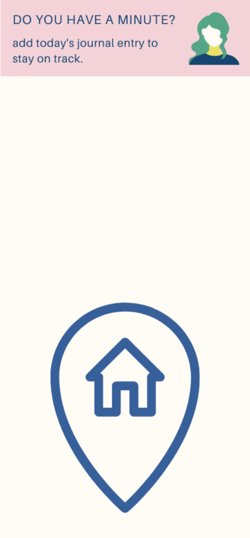
LETTING GO OF NEGATIVE THOUGHTS

this goal is about becoming aware of repeating unhelpful negative thoughts and learning to let them go. related exercises include *breathing meditation*, and *belief monitoring*.

RELAXATION

this goal is about creating space for restful downtime. related exercises include *breathing meditation*, *work boundary setting*, and *yoga*.

Figure D.14: Goal Overview



DO YOU HAVE A MINUTE?

add today's journal entry to stay on track.

Figure D.15: Push Notification

Appendix E

Mock Conversations

Conversations are excerpted from real-world Therapist-Patient conversations presented in [125] and heavily modified.

E.1 The Casual vs The Formal Agent

See Tables E.1, E.2, and E.3.

Context	Neutral
Language Features	Self Disclosure, Emotive Reaction, Question.
Alex	He has been so hard to get a hold of, prepping for the move and everything, I think he is ignoring me.
M	Well, I know even for me, a move can create a crisis. Packing up and moving and getting re-established. I wonder if he's been overwhelmed. What makes you say he's ignoring you?
Alex	He wasn't even willing to see me.
M	How painful.
Alex	And I really don't know what I should do, he wouldn't even tell me where his address would be.

Table E.1: Formality Neutral Example Conversation

E.2 The Introverted vs the Extroverted Agent

See Tables E.4, E.5, and E.6.

Context	Formal
Language Features	Self Disclosure, Emotive Reaction, Question. Would, No Contractions, Single Exclamation.
Alex	He has been so hard to get a hold of, prepping for the move and everything, I think he is ignoring me.
M	Well. I know even for me, a move would create a crisis. Packing up and moving and getting re-established. I wonder if he has been overwhelmed. What makes you say he is ignoring you?
Alex	He wasn't even willing to see me.
M	How painful!
Alex	And I really don't know what I should do, he wouldn't even tell me where his address would be.

Table E.2: Formal Example Conversation

Context	Informal
Language Features	Self Disclosure, Emotive Reaction, Question. Do, Contractions, Multi-Exclamation, Informal Words, Emoticon.
Alex	He has been so hard to get a hold of, prepping for the move and everything, I think he is ignoring me.
M	Yeah, I know even for me, a move creates a crisis!! Packing up and moving and getting re-established. I wonder if he's been overwhelmed. What makes you say he's ignoring you??
Alex	He wasn't even willing to see me.
M	Aww, how painful ❤!
Alex	And I really don't know what I should do, he wouldn't even tell me where his address would be.

Table E.3: Informal Example Conversation

Context	Neutral
Language Features	Self Disclosure, Validation, Question
Alex	I think, its like I need to control things which I know are uncontrollable.
M	I understand that. Letting go of control is difficult for me too. Why do you use the word 'need' here?
Alex	Well, even coming to university I pack almost everything, just in case, you know?
M	In case of what?
Alex	Emergencies, I lose something, stuff like that.

Table E.4: Personality Neutral Example Conversation

Context	Extravert
Language Features	Self Disclosure, Validation, Question, Restatements, High Verbosity, Informal Language, Exaggeration, But-Conjunction, Exclamation.
Alex	I think, its like I need to control things which I know are uncontrollable.
M	Yeah, I really understand that, I do. Letting go of control is truly very difficult for me too! But, I was thinking, why do you use the word 'need' here?
Alex	Well, even coming to university I pack almost everything, just in case, you know?
M	Yeah, in case of what, for example?
Alex	Emergencies, I lose something, stuff like that.

Table E.5: Extraverted Example Conversation

Context	Introvert
Language Features	Self Disclosure, Validation, Question. Formal Language. Negation. Longer words. Softener Hedges. Although-Conjunction. Complex Sentence Structure, Low Verbosity.
Alex	I think, its like I need to control things which I know are uncontrollable.
M	Right. Letting go of control is not easy for myself either, so that is quite understandable. Although, I am not sure why you choose 'need' here?
Alex	Well, even coming to university I pack almost everything, just in case, you know?
M	In what case?
Alex	Emergencies, I lose something, stuff like that.

Table E.6: Introverted Example Conversation

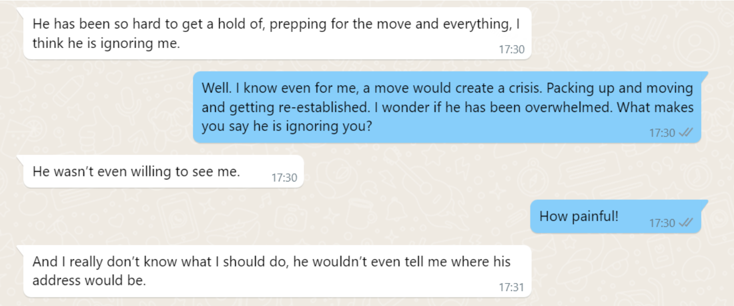
Appendix F

Design Study Survey

Thank you for taking part of this study for the development of a student burnout identification and support system. This questionnaire has 3 parts; design questions, language questions, and personal demographic questions. It should take no more than 15 minutes to complete. Please try to write as much as you can and as freely as you like!

F.1 Design

- Which feature was most important to you?
  - Avatar Selection (See D.3)
  - Chatbot messaging (See D.7)
  - Goal Editing (See D.8)
  - Interest Questionnaires (See D.6)
  - Push Notifications (See D.15)
- Why?
- Which feature was least important to you?
  - Avatar Selection (See D.3)
  - Chatbot messaging (See D.7)
  - Goal Editing (See D.8)
  - Interest Questionnaires (See D.6)
  - Push Notifications (See D.15)
- Why?
- Is there anything about this prototype you would change to make it easier to use or more enjoyable? For example navigation order, removing features, or avatar design.



He has been so hard to get a hold of, prepping for the move and everything, I think he is ignoring me. 17:30

Well. I know even for me, a move would create a crisis. Packing up and moving and getting re-established. I wonder if he has been overwhelmed. What makes you say he is ignoring you? 17:30

He wasn't even willing to see me. 17:30

How painful! 17:30

And I really don't know what I should do, he wouldn't even tell me where his address would be. 17:31

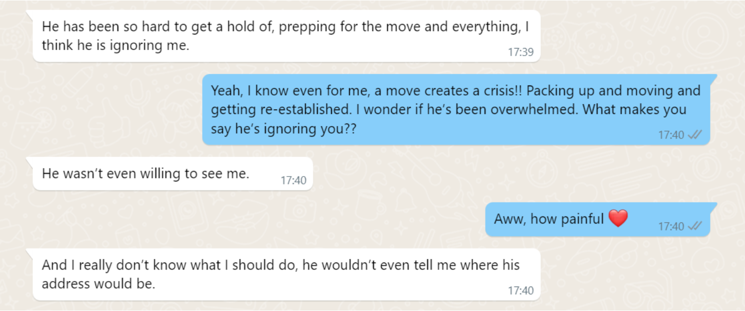
Figure F.1: Conversation 1 - Alex (White) and M (Blue)

- What three words would you use to describe this prototype?
- Which Avatar Selection Screen do you prefer?
  - Figure D.3
  - Figure D.4
- Why?
- Which Goal Editing Workflow do you prefer?
  - Figure D.9
  - Figure F.5
- Why?
- Were you able to complete the assigned task? Did you have any difficulties understanding how to do so?
- What features would you expect or want from a burnout identification and support system? For example mood tracking, sleep tracking, diary keeping, or so on.
- Do you have any other comments about the design? This can be anything; colour palette, font choices, avatar design, feature integration, and so on.

F.2 Language

Consider the following conversations (Formality).

- Which of these adjectives would you use to describe M in conversation F.1?
  - Introverted, Professional, Unprofessional, Boring, Personal, Disengaged, Engaged, Cold, Sociable, Likeable, Extroverted, Friendly, Impersonal, Caring, Formal, Casual.



He has been so hard to get a hold of, prepping for the move and everything, I think he is ignoring me. 17:39

Yeah, I know even for me, a move creates a crisis!! Packing up and moving and getting re-established. I wonder if he's been overwhelmed. What makes you say he's ignoring you?? 17:40

He wasn't even willing to see me. 17:40

Aww, how painful ❤️ 17:40

And I really don't know what I should do, he wouldn't even tell me where his address would be. 17:40

Figure F.2: Conversation 2 - Alex (White) and M (Blue)

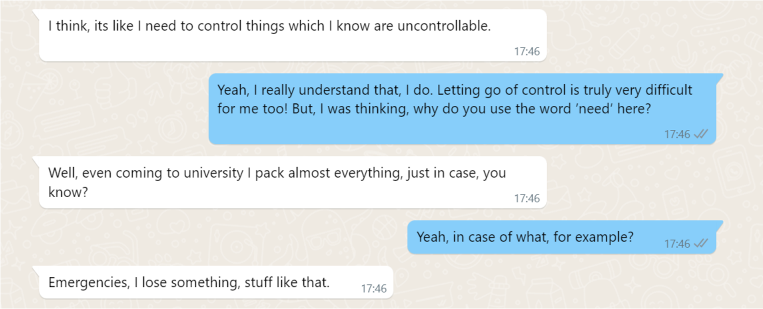
- In conversation F.1, M is interested in Alex's needs: Strongly disagree (1)—Strongly agree (5).
- In conversation F.1, M is effective at supporting emotional needs: Strongly disagree (1)—Strongly agree (5).
- Which circle (Figure F.6) do you think best represents M and Alex's relationship here?

- Which of these adjectives would you use to describe M in conversation F.2?
  - Introverted, Professional, Unprofessional, Boring, Personal, Disengaged, Engaged, Cold, Sociable, Likeable, Extroverted, Friendly, Impersonal, Caring, Formal, Casual.
- In conversation F.2, M is interested in Alex's needs: Strongly disagree (1)—Strongly agree (5).
- In conversation F.2, M is effective at supporting emotional needs: Strongly disagree (1)—Strongly agree (5).
- Which circle (Figure F.6) do you think best represents M and Alex's relationship here?

- Which conversation corresponds with the version M you would prefer to use?
- Why?

Consider the following conversations (Extraversion).

- Which of these adjectives would you use to describe M in conversation F.3?



I think, its like I need to control things which I know are uncontrollable. 17:46

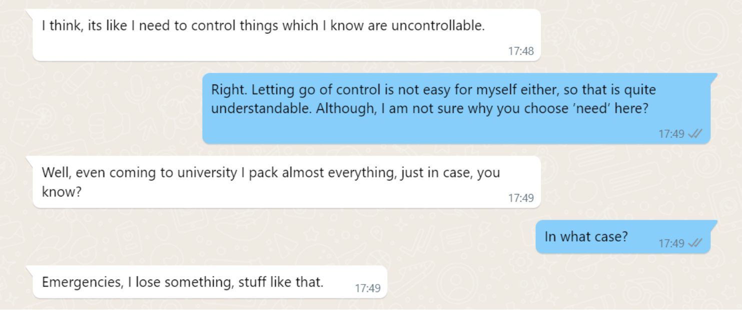
Yeah, I really understand that, I do. Letting go of control is truly very difficult for me too! But, I was thinking, why do you use the word 'need' here? 17:46

Well, even coming to university I pack almost everything, just in case, you know? 17:46

Yeah, in case of what, for example? 17:46

Emergencies, I lose something, stuff like that. 17:46

Figure F.3: Conversation 1 - Alex (White) and M (Blue)



I think, its like I need to control things which I know are uncontrollable. 17:48

Right. Letting go of control is not easy for myself either, so that is quite understandable. Although, I am not sure why you choose 'need' here? 17:49

Well, even coming to university I pack almost everything, just in case, you know? 17:49

In what case? 17:49

Emergencies, I lose something, stuff like that. 17:49

Figure F.4: Conversation 2 - Alex (White) and M (Blue)

  - Introverted, Professional, Unprofessional, Boring, Personal, Disengaged, Engaged, Cold, Sociable, Likeable, Extroverted, Friendly, Impersonal, Caring, Formal, Casual.

- In conversation F.3, M is interested in Alex's needs: Strongly disagree (1)—Strongly agree (5).

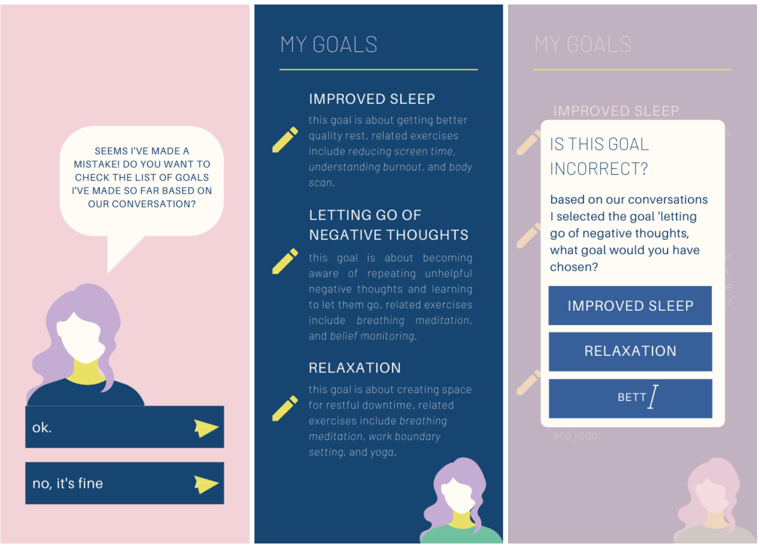
- In conversation F.3, M is effective at supporting emotional needs: Strongly disagree (1)—Strongly agree (5).

- Which circle (Figure F.6) do you think best represents M and Alex's relationship here?

- Which of these adjectives would you use to describe M in conversation F.4?

  - Introverted, Professional, Unprofessional, Boring, Personal, Disengaged, Engaged, Cold, Sociable, Likeable, Extroverted, Friendly, Impersonal, Caring, Formal, Casual.

- In conversation F.4, M is interested in Alex's needs: Strongly disagree (1)—Strongly agree (5).



SEEMS I'VE MADE A MISTAKE! DO YOU WANT TO CHECK THE LIST OF GOALS I'VE MADE SO FAR BASED ON OUR CONVERSATION?

IMPROVED SLEEP

this goal is about getting better quality rest. related exercises include *reducing screen time*, *understanding burnout*, and *body scan*.

LETTING GO OF NEGATIVE THOUGHTS

this goal is about becoming aware of repeating unhelpful negative thoughts and learning to let them go. related exercises include *breathing meditation*, and *belief monitoring*.

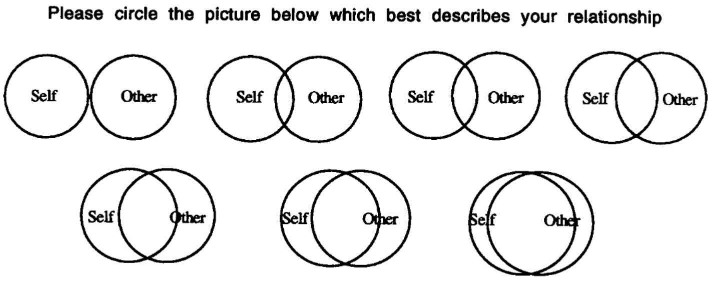
RELAXATION

this goal is about creating space for restful downtime. related exercises include *breathing meditation*, *work boundary setting*, and *yoga*.

IS THIS GOAL INCORRECT?

based on our conversations I selected the goal 'letting go of negative thoughts, what goal would you have chosen?

Figure F.5: Classical Navigation Goal Editing Flow



Please circle the picture below which best describes your relationship

Figure 1. The Inclusion of Other in the Self (IOS) Scale.

Figure F.6: Inclusion-of-the-Self-in-the-Other (Closeness)

- In conversation F.4, M is effective at supporting emotional needs: Strongly disagree (1)—Strongly agree (5).
- Which circle (Figure F.6) do you think best represents M and Alex's relationship here?

- Which conversation corresponds with the version M you would prefer to use?
- Why?

F.3 Demographic

About you: Let us know about yourself to help us understand our demographic.

- Gender
- Age Group
- Occupation Status: Undergraduate Student / Postgraduate Student / Recent Graduate / Prefer not to say
- Occupation: Architecture and Civil Engineering /.../ Social Sciences
- Geographic Area of Origin: North America /.../ Polynesia
- Have you used wellbeing support apps before?
- What was your experience with them? What did you like/dislike about them?
- How would you rate your experience with technology?

Appendix G

Design Study Codebook

Table G.1: Design Study Codebook with Descriptions and Examples

Code	Description	Examples	Category	Count
social presence	Includes references to the personhood of M. References to feelings of personal attention. References to friend-like behaviour. Can be discussed in a negative or positive light. Can be high (human-like) or low (robot-like).	"I felt like someone was there to help"	UX	24
				continues on next page

conversation	Explicit references to: talk, chat, conversation, or telling, speaking, and saying. Further, includes references to language and tone.	“the chatbox focused one feels more like a conversation. maybe the user can ask for additional info if necessary”	Function	13
supportiveness	Includes references to mental health support capacity such as sense of care. Many emotion words are seen in this code (warm, calm, hope, cold, abrupt). Can be high or low.	“Maybe [I would add] links to contacts that van help you? Mum, dad etc”	Function	23
engagement	References to user’s felt level of engagement or to their perception of M’s engagement with them. Can be high or low.	“It seems more involved, even if still it seems [like] an unlucky response”	UX	10

continues on next page

usefulness	References to the usefulness of the prototype and its tools to the user. Can be low (useless) or high (useful).	“It makes the app useful to me”	Function	9
disorientation	Technical or conceptual difficulties, creating unclear expectations for users. Can be in regard to the prototype itself or to the study design. Not dimensional. Low disorientation is in ease-of-use.	“The order is not too intuitive”	UX	23
ease-of-use	References to speed, directness, ease-of-use, or conciseness as perceived by the user. Not dimensional. Low ease-of-use is in disorientation.	“[Chatbot based editing] seems like less mess to consider [as opposed to classic navigation]”	UX	17

continues on next page

aesthetic	References to the look and feel of the prototype. Can be discussed in a negative or positive light.	‘As I said, I personally would prefer a more neutral design.’	UX	18
appropriateness	Includes references to M’s capacity to understand, and respect, relationship boundaries. Similarly, the suitability of the prototype to its purpose and context. Can be low (inappropriate) or high (well suited).	‘I think using emoji is not appropriate in this situation. Does not look that serious?’	UX	13
enjoyment	Includes references to joy, happiness, interest and fun aspects of the prototype or bot. Can be high or low.	‘It’s a nice delighter but it does not enhance the main function/goal.’	UX	8
inclusivity	References to inclusive representation of the target user group in the given selection of virtual chatbot avatars	“More abstract thus inclusivity is not a relevant matter”	Function	3

continues on next page

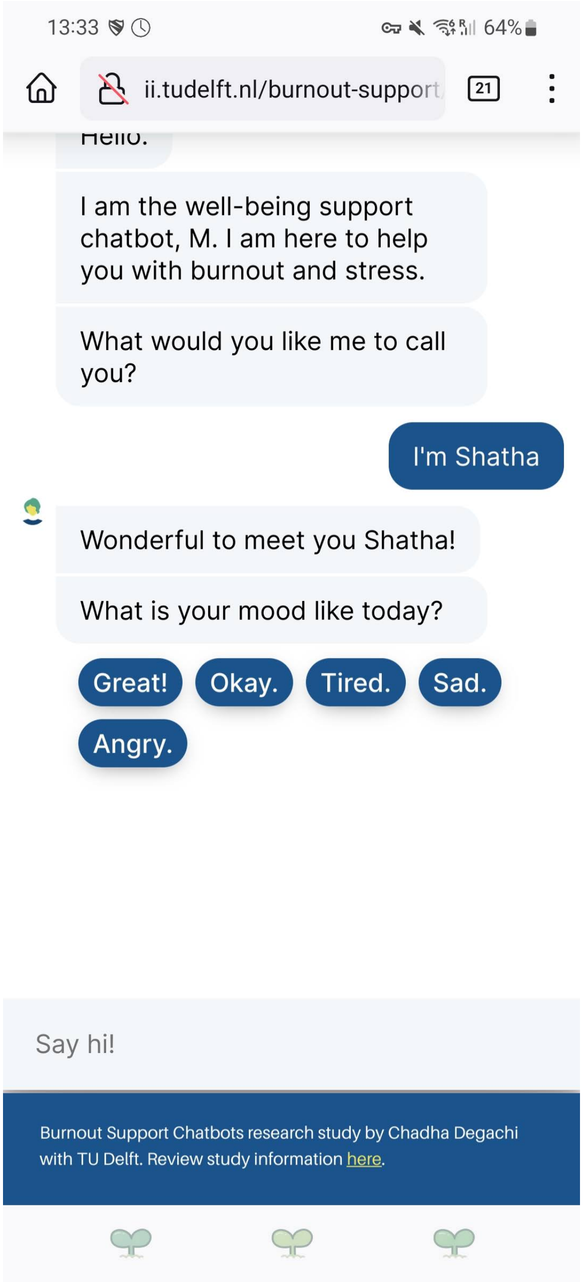
trust	Explicit discussion of trust as perceived by the user. Can be high or low.	‘Original, trustworthy, smooth’	UX	1
autonomy	References to editing, choosing, personalizing or controlling aspects of the systems whether desired, praised, or unhelpful. Can be high or low.	‘It gives you something you can modify’	UX	7
goals	Explicit references to goal setting or goal editing.	“When the goal list is reasonable, it gives me more hope.”	Function	3
expense	References to monetary elements and considerations. Can be discussed in negative or positive light.	“I like that it’s sort of a quick, guided meditation help. Although now most of the activities are paid :(”	Function	1
tools	References to specific psychological and technical tools. Can be discussed in a positive or negative light/	”’retreat’ feature, an emergency feature that guides me how to come back to a relaxed state”	Function	20

continues on next page

preference	Expression of importance or preference in relation to other items from the user point of view. References to comparison. Can be neutral (no preference).	‘The characters brings happy thoughts :) and the person one feels therapist-y’	25	
awareness	References to external or pre-existing knowledge of the subject matter (mental health apps) and its related topics. Can be high or low.	“I didn’t know [other wellbeing apps].”	1	
meaning unclear	Response meaning unclear to annotator.	‘Having a solution’		2

Appendix H

Pilot Screenshots



Hello.

I am the well-being support chatbot, M. I am here to help you with burnout and stress.

What would you like me to call you?

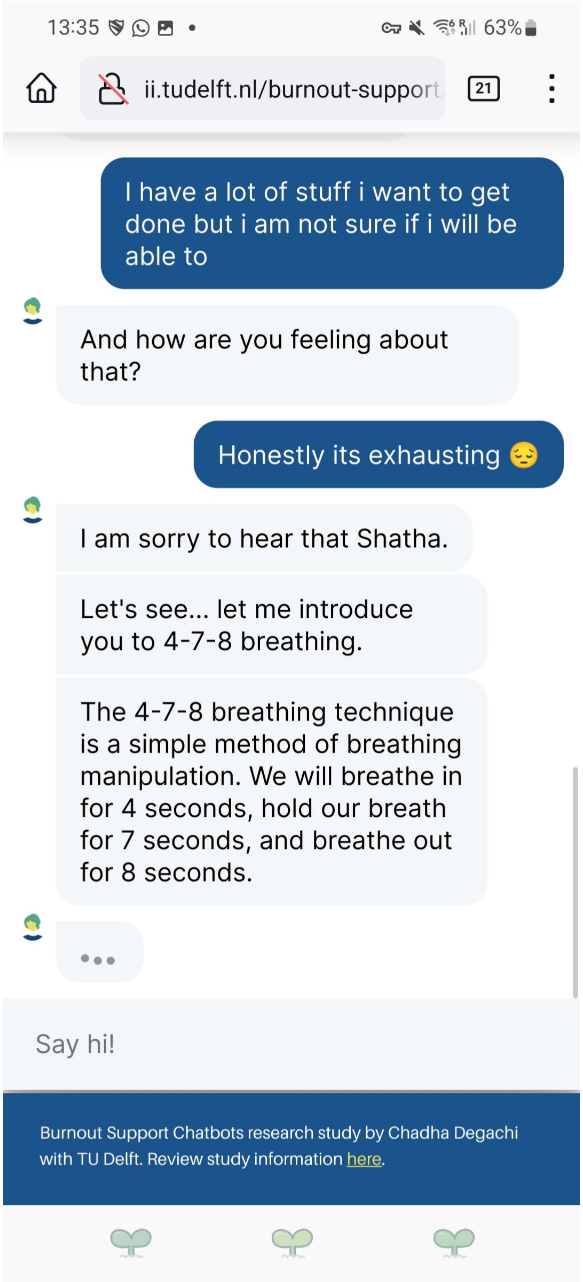
I'm Shatha

Wonderful to meet you Shatha!

What is your mood like today?

Burnout Support Chatbots research study by Chadha Degachi with TU Delft. Review study information here.

Figure H.1: Prototype Chat Screen: Mood recording (Screenshots by Author)



I have a lot of stuff i want to get done but i am not sure if i will be able to

And how are you feeling about that?

Honestly its exhausting 😔

I am sorry to hear that Shatha.

Let's see... let me introduce you to 4-7-8 breathing.

The 4-7-8 breathing technique is a simple method of breathing manipulation. We will breathe in for 4 seconds, hold our breath for 7 seconds, and breathe out for 8 seconds.

•••

Say hi!

Burnout Support Chatbots research study by Chadha Degachi with TU Delft. Review study information here.

Figure H.2: Prototype Chat Screen: Breathing Exercise (Screenshots by Author)

Appendix I

Pilot Survey

Thank you for taking part of this study for the development of a student burnout identification and support system. This study is being done by Chadha Degachi from the TU Delft as part of a Master's Thesis. We are very interested in your feedback to improve our design!

Your participation in this study is entirely voluntary, and you can withdraw at any time, up to survey submission. You are free to omit any question. After submission, your data can no longer be removed.

This questionnaire has 2 parts; design questions and personal demographic questions. It should take no more than 10 minutes to complete. Please try to write as much as you can and as freely as you like!

I.1 Prototype

- Which circle best represents your relationship with M after this conversation (Figure F.6)?
- On a scale of 0 (very poorly) to 10 (very well), how well do the following adjectives fit M?
  - Friendly
  - Personal
  - Sociable
  - Likeable
- I would describe M as: Introverted, Extroverted, Formal, Casual.
- Did you have any technical difficulties using the system? Why or Why not?
- Did you find that learning to use the system was easy? Why or Why not?
- Would you like to use this system frequently? Why or Why not

- M is interested in my needs: Strongly disagree (1)—Strongly agree (5).
- M is effective at supporting my emotional needs: Strongly disagree (1)—Strongly agree (5).
- Did you feel comfortable speaking with M? Why or Why not?
- Did M make any mistakes in the conversation? Can you give any examples?
- Do you have any other comments about the system?

I.2 Demographic

About you: Let us know about yourself to help us understand our demographic.

- Gender
- Age Group
- Occupation Status: Undergraduate Student / Postgraduate Student / Recent Graduate / Prefer not to say
- Occupation: Architecture and Civil Engineering /.../ Social Sciences
- Geographic Area of Origin: North America /.../ Polynesia
- Have you used wellbeing support apps before?
- What was your experience with them? What did you like/dislike about them?
- How would you rate your experience with technology?

Appendix J

Pilot Codebook

Table J.1: Pilot Study Codebook with Descriptions and Examples

Code	Description	Example	Category	Count
design	Includes references to visual elements of the prototype design.	No, the only thing there was a banner blocking my view	UX	1
correctness	Includes references to performance of M in relation to the outcomes expected by the users. Dimensional. High (exhibiting correct behaviour)—Low (exhibiting incorrect behaviour)	“Sometimes the chatbot did not understand my not that obvious answers”	Function	5
				continues on next page

ease-of-use	References to speed, directness, ease-of-use, or conciseness as perceived by the user. Dimensional.	"Yes it was easy, also the chatbot explained how ai works"	UX	5
informative	Related to usefulness. References to information/tips/advice given to the user by the system.	"I would like to use it, expecially if it keeps giving advices and information about techniques to reduce stress and anxiety"	Function	4
names	References to how the chatbot addressed the user. Largely covers instances of incorrect behaviour in the chatbot.	"Yes she said my name twice in the beginning of every sentence"	Function	3
				continues on next page

social presence	Includes references to the personhood of M. References to feelings of personal attention. References to friend-like behaviour. Can be discussed in a negative or positive light. Can be high (human-like) or low (robot-like)	“Somewhat comfortable. Only issue was the long paragraphs of facts being sent instantly makes it feel less human.”	UX	3
usefulness	References to the usefulness of the prototype and its tools to the user. Can be low (useless) or high (useful)	“My session lasted 5mins mainly because I was having a pretty good day. I think there should be a few more questions asked even if the person is feeling good at the moment.”	Function	5
praise	Laudatory talk in relation to the general prototype. Unconnected to a specific prompt question.	“Good job!”	UX	1
				continues on next page

feature request	Encodes requests for specific behaviours from the participants.	“Everything was pretty clear, maybe he had to let me finished the technique and then add the link”	Function	1

Appendix K

Functional Requirements

ID	Name	Description
FR1	Informed Consent	Prototype will inform the participant of 1) the purpose of the experiment, 2) what is expected of them, and 3) their rights as a participant.
FR2	Privacy	Prototype will inform the participant of 1) what data is collected, and 2) how it is used.
FR3	External Support	Prototype will direct users to TU Delft Counselling and Self Service pages.
FR4	Goals 1	Prototype will allow users to view their goals
FR5	Goals 2	Prototype will (in experimental groups) allow users to edit their goals.
FR6	Goals 3	Prototype will visualise participant progress along every goal.
FR7	Chatbot 1	Prototype will allow participants to initiate conversations with an intelligent chatbot.
FR8	Chatbot 2	Prototype Chatbot will conduct the OLBI only in Session 0 and other predetermined progress check-in points.
Fr9	Chatbot 3	Prototype Chatbot will engage participants in small talk
FR10	Chatbot 4	Prototype Chatbot will suggest goals to participants
FR11	Chatbot 5	Prototype will suggest psychologically valid therapeutic exercises to participants which align with their goals.
FR12	Chatbot 6	Prototype Chatbot will coach participants through psychologically valid therapeutic exercise.
FR13	Chatbot 7	Prototype Chatbot will check and log user mood once a day
FR14	System 1	Prototype will allow participants to set a nickname for themselves in the system.
FR15	System 2	Prototype will redirect participants to our experimental surveys at the end of their conversation.

Table K.1: Functional Requirements

Appendix L

OLBI-S Abbreviated

1. I find my studies to be a positive challenge.
2. This is the only field of study that I can imagine myself doing.
3. I feel more and more engaged in my studies.
4. I can tolerate the pressure of my studies very well.
5. While studying, I often feel emotionally drained.

Appendix M

Goal - OLBI Mapping

	Goal	Associated Questions
1	Being more aware and present	Q11
2	Taking more restful breaks	Q12
3	Setting and enforcing work-life boundaries	Q4
4	Finding small joys in day-to-day life	Q7
5	Refocusing on my core values	Q8
6	Learning a new language	Irrelevant
7	Finding a new hobby	Irrelevant
8	Learning to eat better	Irrelevant
9	Taking more time to read	Irrelevant
10	Being a tidier person	Irrelevant

Table M.1: Mapping from Goals to OLBI Questions

Appendix N

Pilot Dialogue Flow

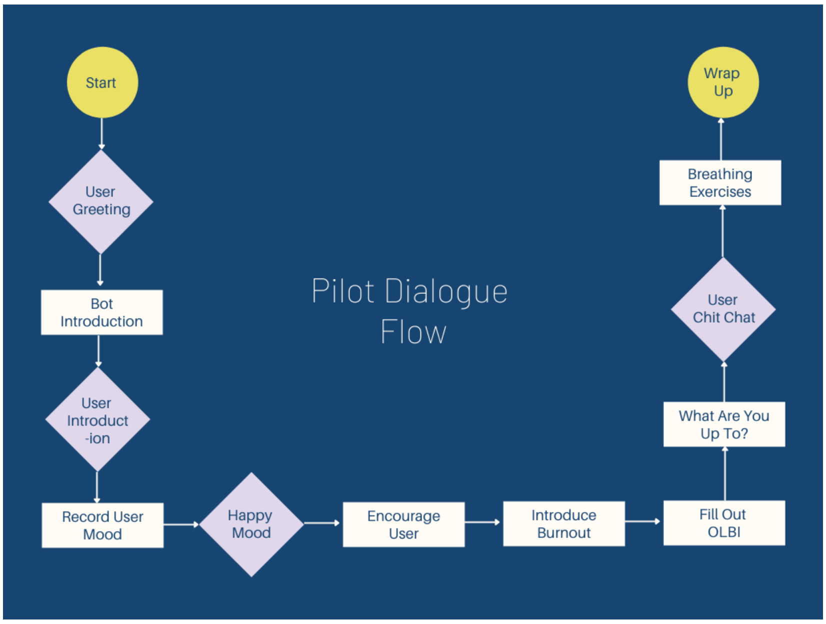


Figure N.1: Pilot Study Dialogue Flow

Appendix O

Pretest Survey

You are being invited to participate in a research study titled Burnout Support Chatbots, the study is being done by Chadha Degachi from the TU Delft as part of a Master's Thesis.

The purpose of this research study is to investigate the relationship between interactive learning systems and users. We are very interested in your experience! The study will take you approximately 20 minutes to complete.

We collected two type of data: multiple choice and free-text open answer questions. Your answers to open questions may be quoted in the published study, but will be stripped of all private information before doing so.

Your answers are used to address our research questions, additionally along with your feedback, and in order to have an idea of our users' backgrounds we do conduct a demographic survey at the end of this study and will ask you information regarding your age, gender, occupation, and geographic area of origin. All information collected in this survey will be anonymized and will only be published as an aggregate of all study participants. Raw data will only be accessible to the project researcher.

Your participation in this study is entirely voluntary, and you can withdraw at any time, up to survey submission. You are free to omit any question. After submission, your data can no longer be removed.

Our chatbot will discuss some emotional subjects with you. We do not expect this to be upsetting, however, if you do feel anxious or upset, we encourage you to get in touch with the TU Delft counselling team on their website or browse the TU-Delft e-Health Support Platform Gezondeboelfor further support.

As with any online related activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by anonymizing and aggregating any sensitive data, as well as cleaning text data of any private information. Raw data will only ever be accessible to the research team, it will be stored on TU Delft servers and storage solutions, and will be deleted at the end of the study (Jan 2023).

The research team conducting this study can be contacted at:

- Chadha Degachi (researcher): c.degachi@student.tudelft.nl
- Myrthe Tielman (supervisor): m.l.tielman@tudelft.nl
- Santosh Ilamparuthi (data steward): s.ilamparuthi@tudelft.nl

This questionnaire has 3 parts; a background section to be completed before using the system, an experience section to be completed after using the system, and a personal demographic section. Please try to write as much as you can and as freely as you like!

O.1 Demographic

About you: Let us know about yourself to help us understand our demographic.

- Gender
- Age Group
- Occupation Status: Undergraduate Student / Postgraduate Student / Recent Graduate / Prefer not to say
- Occupation: Architecture and Civil Engineering /.../ Social Sciences
- Geographic Area of Origin: North America /.../ Polynesia
- Have you used wellbeing support apps before?
- What was your experience with them? What did you like/dislike about them?
- How would you rate your experience with technology?

O.2 Attitude Towards AI

- How interested are you in using artificially intelligent systems in your daily life? Not interested at all (1) Extremely interested (5)
- How do you feel about the use of artificially intelligent systems becoming more common? Extremely negative (1) Extremely positive (5)

Thank you for completing this section. You can now interact with the system. The system will redirect you to this survey after your interaction is complete, you can keep this tab open or close it. This activity works best on a mobile phone! Click on this link to start.

Appendix P

Post Test Survey

Welcome back! Thank you for using our system. Click next to continue this survey.

P.1 Goals

- How many goals did M propose to you? 0 - 3
- How many goals did you accept? 0 - 3

P.2 Perceived Control

- How did you feel during the task?
  - M adapted to me (1) - I adapted to M (7).
  - I was controlling M's understanding of me (1) - M was controlling its understanding of me (7).
  - M intended to help me (1) - M intended to complete its own task (7).
  - I felt relaxed during this interaction (1) - I felt stressed during this interaction (7).
  - I felt confident using this system (1) - I felt unconfident using this system (7).

P.3 Trust

- I believe that M will act in my best interest. Strongly disagree (1)—Strongly agree (5).
- I believe that M will do its best to help me if I need help.Strongly disagree (1)—Strongly agree (5).

- I believe that M is interested in understanding my needs and preferences. Strongly disagree (1)—Strongly agree (5).
- I think that M is competent and effective in providing emotional support. Strongly disagree (1)—Strongly agree (5).
- I think that M performs its role as a burnout support chatbot very well. Strongly disagree (1)—Strongly agree (5).
- I believe that M has all the functionalities I would expect from a burnout support chatbot. Strongly disagree (1)—Strongly agree (5).
- If I use M, I think I would be able to depend on it completely. Strongly disagree (1)—Strongly agree (5).
- I can always rely on M for providing emotional support. Strongly disagree (1)—Strongly agree (5).
- I can trust the information presented to me by M. Strongly disagree (1)—Strongly agree (5).

P.4 Secondary Factors

- Which circle best represents your relationship with M after this conversation? (See Figure F.6).
- On a scale of 0 (very poorly) to 10 (very well), how well do the following adjectives fit M?
    - Friendly
    - Personal
    - Sociable
    - Likeable
- This system's capabilities meet my requirements. Strongly disagree (1)—Strongly agree (7).
- This system is easy to use. Strongly disagree (1)—Strongly agree (7).

P.5 User Experience

- Do you think M's goal recommendations improved over the course of your conversation? Why?
- Did you feel the ability to change the suggested goals was important to you? Why?
- What affected your sense of trust in M the most?

Appendix Q

Experiment Codebook

Table Q.1: Experiment Study Codebook with Descriptions and Examples

Code	Description	Examples	Category	Count
supportiveness	Includes references to mental health support capacity such as sense of care. Many emotion words are seen in this code (warm, calm, hope, cold, abrupt). Can be high or low.	As a bot it is unable to judge and I don't feel the pressure of talking to a real person	Function	15
				continues on next page

competence	Relates to correctness of behaviour and technical capability as a system and a chatbot. Dimensional, can be high or low. Different from ease-of-use-disorientation scale in that it does not focus on technical bugs.	Yes, because the bot does not necessarily understand what I'm saying, so it's very useful to be able to correct it	Function	27
benevolence	Relates to friendliness, good-intention, and care afforded the user by the system. Dimensional, can be high or low. Different from supportiveness as it is not intrinsically tied to system functions.	The responses were professional and friendly	UX	7

continues on next page

ease-of-use	References to speed, directness, ease-of-use, or conciseness as perceived by the user. Not dimensional. Low ease-of-use is in disorientation.	No because [goal editing]'s too much trouble	Function	1
disorientation	Technical or conceptual difficulties creating unclear expectations for users. Can be in regard to the prototype itself or to the study design. Not dimensional. Low disorientation is in ease-of-use.	She didn't understand my goal at first, so I had to go over the same conversation again	Function	14
autonomy	References to editing, choosing, personalizing or controlling aspects of the systems whether desired, praised, or unhelpful. Can be high or low.	Yes, [goal editing] gave a freedom of choice, the control is in my hands	UX	14

continues on next page

adaptation	References to changing behaviour on the part of M over the course of the interaction. Dimensional. Can be high (adaptive) or low (failure to adapt).	Yes, it went from completely generic to somewhat personalized	Function	9
attitude to AI	References to users' pre-existing conceptions and expectations of AI. Dimensional. Can be high (positive attitude) or low (negative attitude).	Personally, I do not trust a system like M very much. Not because I consider it to be hostile in some way (though I might if this was provided by an insurance company), but because it is rather transparently a chatbot. I don't generally trust systems like these because they can break easily.	UX	3
				continues on next page

appropriateness	Includes references to M's capacity to understand, and respect, relationship boundaries. Similarly, the suitability of the prototype to its purpose and context. Can be low (inappropriate) or high (well suited).	"[...]. Also, when burnt out, the last thing I want to do is add the additional stress of learning a new language."	UX	4
no improvement	The participant explicitly states they perceived no improvement in chatbot recommendation accuracy.	No, I did not understand why things where suggested	Function	11

continues on next page

neutral on improvement	The participant explicitly states they don't know whether the chatbot recommendation accuracy improved.	Hard to say. The goals were definitely taking the stock answers into account, but it weighted the desire to learn new things over feeling unable to cope with studies and suggested learning a new language on top of study pressures. I think it could have prioritised getting on top of studies instead of trying to add something else to the pile	Function	1
improvement	The participant explicitly states they perceived an improvement in chatbot recommendation accuracy.	Yes, M understood my needs better	Function	13
				continues on next page

explainability	Relates to the understanding of M's motivations, internal workings, and reasoning. Largely expressed as the desire to understand or the inability to understand.	I dont know what information is in M and what the recommendations are based on.	UX	3
usefulness	References to the usefulness of the prototype and its tools to the user. Can be low (useless) or high (useful)	Asking for another suggestion would have been useful, definitely. [...].	Function	5
neutral on editing	The participant explicitly states they are neutral on the incorporation of goal editing in their experience with M.	This may have been helpful just to get a better idea of the scope of goals M has to recommend, rather than going through them one by one, but i personally didn't mind [...].	Function	1
				continues on next page

no editing	The participant explicitly states goal editing was not or would not be important to their experience with M.	No because [goal editing]'s too much trouble	Function	5
want editing	The participant explicitly states goal editing was or would be important to their experience with M.	Yes, [the goals] would make it more customisable and relevant to me	Function	20
external context	References to the context in which the prototype is used, in this case a research study.	[What most affected my trust in M was that] It has academic support.	UX	2
privacy	Includes explicit references to data privacy concerns by participants.	[...] M itself didn't assure me that our conversation would be private. If we removed the context of this being a study that would definitely affect my sense of trust for it	UX	1