AN INTELLIGENT COGNITIVE SYSTEM FOR COMPUTATIONAL PSYCHOTHERAPY WITH A CONVERSATIONAL AGENT FOR ATTITUDE AND BEHAVIOR CHANGE IN STRESS, ANXIETY AND DEPRESSION

Tine Kolenik

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dr. Matjaž Gams, Jožef Stefan Institute & Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Co-Supervisor: Prof. Dr. Günter Schiepek, Ludwig Maximilian University, Munich, Germany & Paracelsus Medical University, Salzburg, Austria

Evaluation Board:
Dr. Mitja Luštrek, Chair, Jožef Stefan Institute & Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Dr. Alessandro Gennaro, Member, Sapienza University of Rome, Rome, Italy
Assist. Prof. Dr. Jana Krivec, Member, School of Advanced Social Studies in Nova Gorica, Nova Gorica, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Tine Kolenik

AN INTELLIGENT COGNITIVE SYSTEM FOR COMPUTATIONAL PSYCHOTHERAPY WITH A CONVERSATIONAL AGENT FOR ATTITUDE AND BEHAVIOR CHANGE IN STRESS, ANXIETY AND DEPRESSION

Doctoral Dissertation

INTELIGENTNI KOGNITIVNI SISTEM ZA RAČUNSKO PSIHOTERAPIJO S POGOVORNIM AGENTOM ZA SPREMINJANJE ODNOSA IN VEDÊNJA PRI STRESU, ANKSIOZNOSTI IN DEPRESIJI

Doktorska disertacija

Supervisor: Prof. Dr. Matjaž Gams

Co-Supervisor: Prof. Dr. Günter Schiepek

Ljubljana, Slovenia, August 2023

To my parents for foundations
To Jaya for partnership and beauty
To Matej for talks
To Gregor and Uroš for Thursdays
To Jakob for hikes
To Alina for paws
To Tomaž for banter
To klapci for notifications
To JSI for warm embraces
To myself

Acknowledgments

I will never stop thanking my parents for what they gave me in my youth. Among countless other things, I will keep repeating what they always said to me: knowledge is a gift no one can take away from you. Thank you.

Jaya, my dearest partner in this shared creation, this thesis would not look like it does today without you. It reflects multiplicity that your stream carries into me.

My deepest gratitude to Professor Matjaž Gams, without whom this thesis would never have come into existence. Not only did he see enormous potential in me, his support has been instrumental in the completion of this thesis.

Heartfelt thanks to Professor Günter Schiepek, whose kindness, passion, and erudition were a constant motivation.

Matej, Gregor, Uroš, Jakob, Alina, Tomaž, the klapci gang, you all continuously brighten my days, more than you probably know.

My Jožef Stefan Institute coworkers and my students, Primož, Martin, Luka, Sabina, Tilen, Žiga, Maruša, Miha, and Urša, you all helped in some way so that I could walk this path.

Sincere thanks to all the anonymized participants in my studies. This thesis is ultimately also about you.

Finally, this work would not have been possible without the funding from the Slovenian Research Agency (research core funding No. P2-0209 and Young researchers postgraduate research funding).

Abstract

The increasing prevalence of mental health issues worldwide has amplified the significance of computational psychotherapy, which includes creating computational tools for the mental healthcare and tools to support existing mental health professionals. This work presents a computational psychotherapy system that predicts and forecasts mental health issues in users, and utilizes a conversational agent to induce behavior and attitude change. The thesis centers around two main contributions. The first contribution is a novel, golden standard dataset, which includes panel data, encompassing multiple individuals at multiple time intervals. It incorporates 1495 instances of quantified stress, anxiety, and depression levels, as well as symptom scores derived from diagnostic-level questionnaires, accompanied by qualitative daily diary entries. The second contribution is the system itself. The hypothesis posits that for the system to be effective in inducing mental health issues relief with a conversational agent, it needs to simulate theory of mind - the cognitive ability to understand others and act accordingly. The system simulates theory of mind with an artificial cognitive architecture comprised of an ensemble of computational models. It uses psychological as well as cognitive modelling and machine learning models trained on the novel dataset, all in conjuction with novel domain ontologies. The system was evaluated through a computational experiment on mental health phenomena prediction and forecast from quantitative scores and qualitative text diary entries, and an empirical interventional study on relieving mental health issues in participants where it was compared against Woebot. The latter system was chosen as it is currently the most cited freely available system with the most replicated positive outcomes. This work's system showcased superior performance compared to state-of-the-art systems in terms of both the number of detected mental health categories and detection accuracy. It achieved an accuracy of 91.41% using the kNN algorithm (chosen for its explainability, despite several other algorithms performing slightly better), surpassing the highest accuracy of one of the other systems which reached 84% using Long short-term memory. The highest accuracy for 7-day forecasting achieved 87.68%, while other systems were not able to forecast trends. In the empirical interventional study on 42 participants in a simulated daily check-in, the system outperformed Woebot in reducing stress ($p = 0.048$) and anxiety ($p = 0.040$) levels in participants, while both failed to reduce their depression levels ($p = 0.688$). With confirmed hypothesis, it was evaluated that this system performs on par or better than comparable state-of-the-art systems.

Povzetek

Zaradi vse večje razširjenosti težav z duševnim zdravjem po svetu se je povečala potreba po računski psihoterapiji, ki med drugim vključuje ustvarjanje računalniških orodij za duševno zdravje in podporo obstoječim strokovnjakom. To delo predstavlja sistem računske psihoterapije, ki uporablja pogovornega agenta za napovedovanje težav z duševnim zdravjem pri uporabnikih ter spodbujanje sprememb njihovega vedenja. V središču dela sta dva glavna prispevka. Prvi prispevek je nova podatkovna baza, ki vključuje panelne podatke, ki zajemajo več posameznikov v več časovnih vrstah. Vključuje 1495 primerov kvantificiranih ravni stresa, anksioznosti in depresije ter ocene simptomov, pridobljene z diagnostičnimi vprašalniki, ki jih spremljajo kvalitativni dnevniški zapisi. Drugi prispevek je sistem sam. Hipoteza trdi, da mora sistem, da bi bil učinkovit pri lajšanju težav z duševnim zdravjem s pogovornim agentom, simulirati teorijo uma - kognitivno sposobnost razumevanja drugih in posledičnega ustreznega ravnanja. Sistem simulira teorijo uma z umetno kognitivno arhitekturo, ki jo sestavlja sklop računskih modelov. Uporablja tako psihološko kot kognitivno modeliranje in modele strojnega učenja, zgrajene na omenjeni novi podatkovni bazi, delujoč v povezavi z novimi domenskimi ontologijami. Sistem je bil ovrednoten z računskim eskperimentom zaznavanja in prihodnjega napovedovanja pojavov duševnega zdravja na podlagi kvantitativnih rezultatov in kvalitativnih besedilnih dnevniških zapisov ter empirično intervencijsko študijo o lajšanju težav z duševnim zdravjem pri udeležencih, v kateri je bil primerjan z Woebotom. Slednji sistem je bil izbran, ker je trenutno najbolj citiran prosto dostopen sistem z največ repliciranimi pozitivnimi rezultati. Sistem tega dela je v primerjavi z najsodobnejšimi obstoječimi sistemi pokazal boljšo učinkovitost tako glede števila zaznanih kategorij duševnega zdravja kot glede natančnosti zaznavanja. Z uporabo algoritma kNN (izbranega zaradi njegove razložljivosti, čeprav so bili nekateri drugi algoritmi nekoliko boljši) je dosegel natančnost 91,41 %, s čimer je presegel najvišjo natančnost enega od drugih sistemov, ki je dosegel 84 % z uporabo algoritma Long short-term memory. Najvišja natančnost za napovedovanje 7 dni v prihodnost je bila 87,68 %, medtem ko drugi sistemi niso mogli napovedati prihodnjih trendov. V empirični intervencijski študiji na 42 udeležencih je sistem v simuliranem dnevnem pregledu bolje kot Woebot zmanjšal raven stresa ($p = 0{,}048$) in anksioznosti ($p = 0{,}040$) pri udeležencih, medtem ko nobenemu ni uspelo zmanjšati njihove ravni depresije ($p = 0{,}688$). S potrjeno hipotezo je bilo ocenjeno, da sistem tega dela deluje enako ali bolje kot primerljivi obstoječi najsodobnejši sistemi.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

A-A ... anywhere, anytime
ABC ... attitude and behavior change
AI ... artificial intelligence
B5 ... Big Five personality model
CBT ... cognitive behavioral therapy
CNB ... complement naive Bayes
CogA ... cognitive architecture
CPP ... Cialdini's principles of persuasion
DASS ... Depression, Anxiety, and Stress Scale
DMM ... Domain Mapping Matrix
DT ... decision tree
EMA ... ecological momentary assessment
FBM ... Fogg behavior model
ICA ... intelligent cognitive assistant
KNN ... k-nearest neighbors
LIWC ... Linguistic Inquiry and Word Count
LLM ... large language model
LSTM ... long-short-term-memory
M ... mean
ML ... machine learning
MLP ... multilayer perceptron
NLP ... natural language processing
NLTK ... Natural Language Toolkit
NumPy ... Numerical Python
PANAS ... Positive and Negative Affect Scale
PSDM ... Persuasive System Design Model
PT ... persuasive technology
RBF ... rule-based filtering
RF ... random forest
RNN ... reinforcement neural network
SAD ... stress, anxiety, and depression
SD ... standard deviation
SDG ... Sustainable Development Goal
SISQ ... Single Item Screening Question
SNS ... Synergetic Navigation System
SOC ... Sense of Coherence
SOTA ... state of the art
SVM ... support vector machine
ToM ... theory of mind
UEQ ... User Experience Questionnaire
VADER ... Valence Aware Dictionary and sEntiment Reasoner

Chapter 1

Introduction

The prevalence of mental health issues, particularly among younger individuals, is not a novel occurrence. In recent times, experts have gone as far as referring to it as a mental health pandemic [1]. World organizations, leaders and decision-makers are recognizing its devastating effect, resulting in mental health well-being appearing in Goal 3 of the 17 UN Sustainable Development Goals (SDGs) [2]. Among mental health issues, stress, anxiety and depression (SAD) seem to be on the forefront, as the figures for SAD symptoms in some groups reach 74% for disabling stress [3], 28% for anxiety disorder [4] and 48% for depression [5]. What is more, between 76% and 85% of people in low- and middle-income countries receive no treatment for their disorder [6], while in high-income countries, the treatment coverage for, e.g., depression is only 33% [7]. Mental health issues have large, multi-faceted effects – on the patient, on their immediate surroundings (family or caretakers) and on the wider society [8]. Individuals face decreased quality of life, worse educational outcomes, lowered productivity and potential poverty, social problems, abuse vulnerabilities and additional health problems. Caretakers face increased emotional and physical challenges as well as decreased household income and increased financial costs. Society faces the loss of several GDP percentage points and billions of dollars per nation annually, alongside with exacerbating public health issues and corrosion of social cohesion. All of these lead to an increasingly stronger positive reinforcement loop – SAD increasingly perpetuates SAD. Too often, mental health issues directly result in the worst possible outcome, loss of human life, as many countries struggle with a high suicide rate [9]. It has been recognized that the reasons for increasing of SAD include a severe lack of mental health professionals and regulations [10] as well as unequal access to mental health care [11]. The COVID-19 pandemic further exposed how harmful neglecting people's well-being for decades can be [12], with social distancing facilitating the exponential rise in psychopathological symptoms. These factors make the field ripe for technological and other scientific therapy-based interventions, especially as individuals with mental health issues prefer therapies to medication [13].

Introducing novel ways to include therapeutic approaches into people's lives, especially through technology, might also help destigmatize psychotherapy itself. Traditionally, psychotherapy did not use computational approaches in its practices, but recently the "relevance of computations with regard to the development, maintenance, and therapeutic change in psychiatric disorders" [14, p. 50] was recognized. Such technology can therefore be placed under the umbrella of computational psychotherapy. This encompasses: 1) studying psychotherapeutic processes computationally, using client and professional data to create various analyses and models; and 2) creating computational tools for the mental healthcare and to support existing mental health professionals, which in certain cases take the role of artificial intelligent support, making use of psychotherapeutic approaches and

techniques like cognitive behavioral therapy. Persuasive technology (also called attitude and behavior change support systems) is an example of technology that can be used for computational psychotherapy.

Persuasive technology (PT) or attitude and behavior change (ABC) support systems represent a synergy between the progress in behavioral sciences and many recent advances in computer science, especially artificial intelligence (AI), which is increasingly characterizing the information society [15]. Its goal is to "change attitudes or behaviors or both (without using coercion or deception)" [16, p. 20] and to "aid and motivate people to adopt behaviors that are beneficial to them and their community while avoiding harmful ones" [17, p. 66]. An effective technological vessel for persuasion is an intelligent conversational agent (also known as chatbot or intelligent cognitive assistant). Intelligent conversational agents (ICAs) strive to: understand context; be adaptive and personalized; learn; be predictive; have internal goals and motivation; interpret; and reason [18]. For such capabilities, ICAs need a cognitive architecture (CogA), a "hypothesis about the fixed structures that provide a mind, whether in natural or artificial systems, and how they work together – in conjunction with knowledge and skills embodied within the architecture – to yield intelligent behavior in a diversity of complex environments" [19, para. 2]. Most importantly, ICAs possess the ability to converse in natural language, likely the most immediate way in which humans communicate [20], and interacting through a dialogue is extremely important in the field of computational psychotherapy and digital mental health.

However, creating effective technology for mental health is challenging. Even mental health professionals find it hard to detect idiographic specificities (unique individual characteristics or experiences that differentiate one person from another) of a person's mental health status and intervene in an efficacious way. Mental health status of a person is very dynamic, meaning it changes non-linearly in short time ranges. Experiments seem to indicate that computational psychotherapeutic systems can be successful to mitigate symptoms of SAD [21]–[25], with reviews on this topic being favorable [22], [24]–[29], agreeing that "early evidence shows that with the proper approach and research, the mental health field could use conversational agents in psychiatric treatment." [25, p. 456]. However, state-of-the-art (SOTA) in the field has not demonstrated sufficient integration of advances in behavioral sciences, digital mental health, and AI, especially in terms of user modelling, personalization, and adaptation. Forecasting models are extremely rare or non-existent [30], and the most advanced language models, like GPT-3 [31], have a history of underperforming in domain-specific tasks, going even as far as telling depressed patients to kill themselves [32]. Recently introduced ChatGPT [33] is still based on GPT-3, and it is currently unclear whether it has better risk parameters to not repeat the aforementioned mistake. GPT-4 might prove to useful in the future, but it is currently made to filter out any help regarding mental health, possibly to avoid similar scenarios as mentioned before.

When technology is used to help people with SAD issues, it can underperform due to using nomothetic data (information gathered in a manner that allows for broad, generalizable conclusions across large populations) and non-personalized interventions. This is underpinned by the fact that there are no datasets for constructing truly successful systems in the first place. What therefore lacks is a holistic and integrative approach to creating a system with novel idiographic models, based on human cognitive capabilities, to overcome these issues. This work presents such a system as well as a novel dataset that drives it. The focus is on building a sufficiently advanced computational psychotherapy system, capable of mental health prediction and efficient behavior change with an ICA. The system targets the non-clinical population with SAD symptoms that have barriers to entry to the mental healthcare system, but could also be used complementarily by mental health professionals, e.g., to monitor their clients and patients, and could help physicians with their excessive

workload [34]. The new generation ICA, presented in this work, houses a novel CogA, which uses not only AI and ML, but also other novel mechanisms. It simulates the theory of mind, the human cognitive ability to understand people as well as to effectively respond to them. This is achieved by building various idiographic, detection and forecasting models, in novel ways that make them perform with higher accuracies than current SOTA. These models are combined with novel ontologies on mental health and behavior change, used to understand the ICA's users. A precise and careful research design in collecting ecological time series data from people with SAD symptoms was conducted. The methods used strove towards explainable and open AI. Furthermore, existing large language models (e.g., GPT-3) have seen little specializations for complex domains like mental health, and this CogA accommodates said models through novel integration into its architecture and makes them usable to help people with SAD.

Next sections in this chapter overview the following: stress, anxiety, and depression as the selected mental health issues; attitude and behavior change support systems; ICAs; and potential benefits and risks of introducing such systems into mental healthcare.

1.1 Stress, Anxiety, and Depression

Stress is the "feeling of being overwhelmed or unable to cope with mental or emotional pressure" [35, para. 1]. What kind of experiences cause stress is highly dependent on the individual, but it generally comes from experience that is "new, unexpected or that threatens our sense of self" or "when we feel we have little control over a situation" [35, para. 2]. Coping with stress highly varies as well, not only in terms of what is perceived as stressful, but also whether it is perceived as positive or negative stress. The human body producing stress hormones, triggering the fight or flight response and activating the immune system may be helpful in some situations, serving as a motivator to start and finish certain behaviors, such as finishing a task before the deadline or performing in a sports match. However, letting the human body stay in the state of stress for longer periods of time can be harmful and debilitating. Due to the very individualized nature of what is perceived as stressful, assessment can be difficult.

Anxiety refers to a natural and normal response to stress or potential danger. It is characterized by feelings of unease, worry, or fear, often accompanied by physical sensations such as increased heart rate, rapid breathing, and tense muscles. While occasional anxiety can be a common part of life, an anxiety disorder occurs when these feelings become persistent, excessive, and interfere with daily activities and overall well-being. Anxiety disorders can take various forms, such as generalized anxiety disorder (GAD), panic disorder, social anxiety disorder, and specific phobias.

Depression is a mental health disorder characterized by persistent feelings of sadness, hopelessness, and a loss of interest or pleasure in activities. It goes beyond ordinary fluctuations in mood and can significantly impact a person's daily functioning and overall well-being. Individuals with depression may experience a range of symptoms, including persistent sadness, fatigue, changes in appetite and sleep patterns, difficulty concentrating, feelings of guilt or worthlessness, and in severe cases, thoughts of self-harm or suicide.

Due to their comorbidity and overlapping symptoms, anxiety and depression are commonly bundled together regarding what types of interventions are utilized to reduce symptoms.

1.2 Attitude and Behavior Change Support Systems

ABC support systems are computer systems that attempt to "change attitudes or behaviors or both (without using coercion or deception)" [16, p. 20] and to "aid and motivate people to adopt behaviors that are beneficial to them and their community while avoiding harmful ones" [17, p. 66]. Attitude and behavior change signifies a phenomenon that is considered to be a temporary or lasting effect on an individual regarding their attitude or behavior as compared to what their attitudes were or how they behaved in the past [17, p. 66]. ABC support systems can be equated or at least considered a part of persuasive technology. PT is the result of the vast advances in behavioral sciences in regards to psychological change [36], human decision-making [37] and related phenomena [38] as well as the arrival of digital technologies, artificial intelligence (AI) and big data. Many societal efforts have been put into creating technologies that would help, motivate, guide and persuade people into bettering themselves and the world around them, though such technology can be and has been abused as well [39]). PT is already used in the health and wellness areas, where it tracks people's behavior as well as their physiological and psychological processes, responding to them by trying to affect their mental states by offering psychotherapeutic advice or to motivate them into making different decisions, e.g., in regards to healthy eating [17]. There are also applications in areas such as education or environmental sustainability, where people are nudged towards greener behavior [40].

Some major persuasive and ABC frameworks [41], that such technologies employ, include: Cialdini's Principles of Persuasion (CPP) [36], Fogg Behavior Model (FBM) [16], Persuasive System Design Model (PSDM) [42], and the nudge theory [38], with firm verification of their effectiveness [43].

CPP is based on the idea that general persuasive strategies are not equally effective for everyone. It identifies various strategies that affect different groups of people differently. Interactive, adaptive technology can be utilized to personalize itself to specific strategies that work for specific groups of people.

FBM is based on the idea that a certain behavior is the result of motivation, ability and a trigger occurring at the same time. Therefore, a person changing their behavior has to be sufficiently motivated, has to possess the ability to change the behavior, and has to be triggered to change the behavior. These are then combined in personalized ways to find the most effective strategies for an individual.

PSDM is based on the need for effective design and evaluation of persuasive systems, and mostly offers a framework for what kind of content and functionality PT should consider. PSDM includes four principles upon which to design PT: 1) primary task support, which supports the user's carrying out of their primary task; 2) dialogue support, which helps users move towards their goals; 3) system credibility, which raises the user's belief in the system's quality; and 4) social support, which motivates the user by leveraging social influence.

Another powerful and effective behavioral change concept – Richard Thaler, its author, received the Nobel Prize for it – is the 'nudge theory'. Nudge is "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentive", where "the intervention must be easy and cheap to avoid" [38, p. 6]. Nudges are being incorporated into PT and ABC support systems as well [44].

For persuasive strategies to be as effective as possible, they have to be tailored to a number of specifics. There are 4 factors in the framework of the Communication-Persuasion Paradigm [45] that determine the influence: 1) characteristics of the source (i.e., the message sender); 2) the message; 3) characteristics of the destination (or the receiver of the message); and 4) the context.

For determining effective strategies, personality models, such as Big Five personality traits (B5) [46] or Hexaco [47], as well as domain specific questionnaires, offer PT a useful way to model a person. Personality is measured on different dimensions (e.g., in B5: openness, conscientiousness, extroversion, agreeableness, neuroticism), which try to describe psychological and cognitive functionalities of individuals, e.g., their mental states and decision-making abilities. Knowledge in specific domains relies on PT's use of questionnaires. For mental health, SAD questionnaires [48] can be used to categorize people with SAD symptoms, which leads to better strategy selection. Such questionnaires give insight into what influences which individuals the most. Empirical phenomenology can also be employed for more detailed first-person accounts [49], which can be used for extracting linguistic features [50], [51] or for other tweaking of ABC techniques in PT. Furthermore, combining subjective data with physiological data is also proving useful for adaptive technologies [52].

These frameworks and models appear in several technological platforms. The most frequently used platforms are mobile and handheld devices (28%), followed by games (17%), web and social networks (14%), other specialized devices (13%), desktop applications (12%), sensors and wearable devices (9%), and ambient and public displays (5%) [17].

ABC can be delivered through various software systems. Intelligent cognitive assistants (or chatbot, chatterbot, interactive agent, conversational AI, smartbot, bot) seem to be the most advanced [22], [24]–[29]. The next section introduces such systems and describes why they seem to be the best vessel for delivering ABC.

1.3 Intelligent Cognitive Assistant Technology

Intelligent cognitive assistants (ICAs) [18] are not only intelligent in terms of being able to converse and have a language model, they have many other abilities that are human-like, relating to human cognition and intelligence. ICA technology has therefore been touted as the next revolution in human-computer coexistence. The technology dates back to the beginning of AI, where one of the first chatbots was developed and available outside of a research laboratory – Weizenbaum's simulation of a Rogerian psychotherapist called ELIZA [53]. However, technological progress has only recently laid the foundations for broad adoption in the form of ICAs such as Alexa and Siri as well as more domain-specific agents such as Woebot [54]. Alexa, Siri and Google Home, however close to certain human capabilities they may seem, still often fail outside of very basic, administrative tasks. Currently the most advanced chatbot, ChatGPT, based on GPT-3 and GPT-4 large language models, while impressive, still fails in crucial situations and when used in more expert domains. For example, in mental health, even chatbots based on large language models quickly start repeating themselves, as they only have very generic models that end up in common phrases and trivial platitudes. Sometimes, their remarks can be even dangerous for the user, as they may be perceived as flippant and negative [55], or give wrong medical advice [56]. Testing their response to stressful accounts, they either do not understand or they fail to show empathy beyond empty words [57]. Expert domains of engagement therefore need domain-specific ICAs.

ICAs, which can be deployed in many devices, e.g. as virtual agents or robots, are striving to: understand context; be adaptive and flexible; learn and develop; be autonomous; be communicative, collaborative and social; be interactive and personalized; be anticipatory and predictive; perceive; act; have internal goals and motivation; interpret; and reason. To be able to come close to such capabilities, ICAs are embedded with a cognitive architecture (CogA), a "hypothesis about the fixed structures that provide a mind, whether in natural

or artificial systems, and how they work together – in conjunction with knowledge and skills embodied within the architecture – to yield intelligent behavior in a diversity of complex environments" [19, para. 2]. Most importantly, ICAs possess the ability to converse in natural language. This seems to be the most immediate way in which humans communicate [20], and the effects of a dialogue on human mental states cannot be overestimated. ICAs, coupled with ABC capabilities, are establishing as a very promising PT.

Using ICAs for ABC is still a new field of research, despite ELIZA being the first chatbot in 1964, as chatbots have mostly been explored for education, customer support or in other simple question-answer contexts [58]. What makes ICAs for ABC unique, is that users reveal personal information more freely, which makes systems more successful in their goals [59]. ABC ICAs and their users can also form a more longitudinal relationship. The interactions are not a one-off, where it is difficult to understand the users and act immediately with efficient strategies. This makes such ICAs able to learn from historical interactions and improve in achieving ABC.

ICAs, besides being a vessel to understand users through modeling their psychological and physiological aspects and use such knowledge to enact ABC, present as an ideal platform for offering help in the field of mental health because of the ability to converse. Such usage of technology, however, may pose various potential benefits and risks, which have to be noted.

1.4 Potential Benefits and Risks

This section addresses the implications using PT for mental health has or mental health care. These implications are divided into those that offer potential benefits in relation to existing problems and obstacles, and those that appear as potential risks of this technology. Regardless of the possible benefits of using PT, it does not represent a replacement for holistic health care, but should work as a complement in a comprehensive systemic approach to public health. Other problems are briefly considered at the end of the section as well.

There are a number of potential benefits PT offers:

1.4.1 Cost

The cost of service of mental health care professionals (from psychotherapists to clinical psychologists and psychiatrists) varies from country to country and is further dependent on country regulations and subsidies. But the cost to the patient mostly depends on the number of practicing professionals available in a given country. Regardless, the cost presents a barrier to people from lower socioeconomic backgrounds [60]. PT for mental health can be realistically made free of charge (and many times is [54]) due to the much lower costs attached to it. There are three major factors that contribute to this: 1) scalability, which means that one PT system can be adopted by theoretically any number of people (the only cost that comes with scalability is server cost, which is marginal compared to human labor) – in contrast, one mental health professional is limited to a certain number of people; 2) the ability of more people to produce effective PT due to existing research that thoroughly reports effective designs; and 3) the amount of people capable of producing such systems is much larger than there is professionals that can offer help. It is acknowledged that comparing the cost of a human health worker to the cost of an application does not fit in the holistic approach to designing mental health care solutions [61]. However, the application may as well lower costs to collecting health data on people from disadvantaged socioeconomic backgrounds, which would offer an insight otherwise overlooked. The application

can also be used to relieve stress on health care by making it available to patients whose conditions are not severe enough to be prioritized (i.e., early triage system).

1.4.2 Availability

The problem of availability can be separated into three subcategories: 1) location-based availability, 2) time-based availability, and 3) cost-based availability. Location-based availability refers to people with mental health issues in locations that have no direct access to mental health professionals in person or even no computerized access to therapy with communication technologies [62]. Using PT for mental health may not be a permanent solution in such cases, but it may offer either one of the few available reliefs or be used as a transitional remedy towards a better access to health care. Time-based availability refers to people with mental health issues needing therapeutic help during times when their chosen professional is unavailable. PT for mental health is available around the clock, making their use complementary with the chosen mental health professional. Patients continuously report these needs, and such complementary uses already exist [63]. Cost-based availability refers to people with mental health issues needing therapeutic help but not having the means to access it more than the minimum recommended amount of hours per week [64], where consensus points at one hour per week. Research [64], [65] shows that more frequent therapy results in better outcomes, and complementary use of PT for mental health can bridge that gap for people not being able to afford more therapy by still having an access to help. Cost-based availability is closely connected to the wider cost problem, as discussed before.

1.4.3 Stigma

Self-stigma, the prejudice which people with mental illness turn against themselves, and public stigma, the reaction that the general population has to people with mental illness, are prevailing issues in the battle for mental health [66]. The problem is two-fold: public stigma makes individuals fear what the society will think about them if they seek treatment, while self-stigma makes them fear interacting with a professional as well as doubt themselves in terms of having problems worth seeking treatment for. Thus, both contribute to individuals with mental health issues deciding not to seek treatment from mental health professionals. Up to 96% of people with SAD do not seek treatment [67]. Research on PT for mental health, especially on ICAs for treating SAD, has shown that people are more comfortable disclosing their feelings and personal information to a computerized or mobile system than to a person [59]. This is because they do not fear being judged as well as having a more private channel (at least as perceived by them) for disclosing their feelings, thoughts and issues in general. People also show lower valuation fears and impression management (worrying how much they are worth or controlling how they appear to others), and increase their expression of sadness and objectively-rated disclosure in communication with artificial systems. This means that the amount of people not seeking treatment can be lowered by introducing therapeutic options that they perceive to be safer and free of stigma for them.

However, there are potential risks that such technologies bring that have to be noted and seriously addressed for PT to reach the potential it has in mental health care:

1.4.4 Group Exclusion

Some groups of people can be excluded from technology-oriented mental health care. The groups discussed are the elderly, the lowest socioeconomic class, and culturally-specific

groups. The group most affected by introduction of technology seems to be the elderly [68]. Their limited ability to incorporate technology into their lives can cause further ageist divides between them and other generational groups. PT should therefore strive to be persuasive for everyone, which many times means focusing on research that targets only the elderly as participants (one of the big problems of psychological and related research is majority student samples). Another group of people that may be excluded from the benefits of PT for mental health are people from the lowest socioeconomic class, where even PT might not be available to them [69]. Creating an even bigger divide for them would result in increasingly detrimental socioeconomic living conditions. Groups that are affected in technological adoption due to the cultural differences are crucially important as well when considering how to advance equality. Research shows that cultures with less contemporary sociopolitical leanings show less adoption of technology [70]. Luckily, the research on PT seems to be fledging in certain low-income countries [71].

1.4.5 Researcher Bias

Due to lack of evaluation standardization of PT for mental health, the research field is prone to the introduction of researcher bias. The possible problems are many: 1) PT systems that are claimed to be successful are not always studied in clinical experiments (e.g., randomized controlled trial), but in quasi-experiments [72] or no experiments at all; 2) the metric on which to evaluate such systems is unclear (usually comes indirectly from their effectiveness in an experiment where the goal is SAD symptoms relief [25]); 3) no consensus on what data is needed to understand a user in a way to offer effective help, as there are always some presuppositions on which data to collect and how to use it for strategy selection; and 4) a lot of existing and available systems are proprietary, which hinders the possibilities to investigate and mitigate bias.

1.4.6 Others

Using PT for mental health, as a fairly young endeavor, also has wider problems connected to the technological integration in the society as a whole. The presented limited perspective therefore falls into a broader range of various issues, including: 1) the problem of personal information privacy [73]; 2) the problem of the lack of longitudinal research on behavior change with PT [74]; 3) the ethics of using personal information for persuasion [75]; 4) the potential risks of digital dependence [76], [77]; 5) the problem of creating AI not aligning with human values [78], and 6) the potential problem of automation and job loss of mental health care professionals. PT is prone to the listed issues (and others not listed) as any other technology.

1.5 Thesis Structure

The thesis is structured as follows:

- Chapter 2 presents related work in ICAs for mental health, which serves to understand current SOTA and how to surpass it;
- Chapter 3 presents the author's rationale for this work;
- Chapter 4 presents the research goals and the hypothesis;
- Chapter 5 presents the materials and methods, encompassing the data collection, the collected dataset, computational experiments design and methods used, and empirical interventional study design;

- Chapter 6 presents the cognitive architecture design of this work's system;
- Chapter 7 presents the results of the computational experiments and the empirical interventional study;
- Chapter 8 presents the discussion of the results;
- Chapter 9 presents the conclusion and future work.

Chapter 2

Related Work

To achieve the goal of conducting a technical review of ICAs for ABC in mental health, state-of-the-art (SOTA) review was selected with some elements of scoping review. To focus on current research, this review covers the last 5 years, which is not an uncommon timespan for fast-developing fields [79]. This limited timespan enables to only survey the latest developments, methods and technologies used for ICAs in mental health. A SOTA review therefore aids at underpinning key concepts in a research area and produce a summarized content, offering a better overview than other forms of review methods, and yielding consistent results to solidify new technological phenomena.

This work adopts the framework proposed by Arksey and O'Malley [80] for conducting reviews. The framework provides a direction for the necessary steps in the process. The course of such an approach includes: 1) identifying the research questions; 2) identifying relevant works; 3) identifying selection criteria and applying it to step 2); 4) extracting and organizing the data; and 5) reporting the results in ways to address the research questions and satisfy the purpose of the review.

The reviewed works were evaluated on the following questions (Qs):

Q1. Which mental health issues do the systems target?

Q2. Which technologies, methods and collected data guide the process to achieve ABC for SAD in the systems?

Q3. What are the technical aspects of the conversational models in the systems?

Q4. What are the platforms used to create the systems?

Q5. What domain knowledge is used to achieve ABC for SAD?

Q6. What user modeling, especially for personalization and adaptation, do the systems conduct?

Q7. What is the overarching cognitive architecture used in the systems?

Q8. How are the systems evaluated in terms of ABC for SAD?

Q9. To what extent can we infer that these systems embody theory of mind?

The search query was constructed by collecting keywords and correlating them with synonyms and related words. The construction was based on the author's knowledge of the area as well as referring to similar review papers [21]–[23]. The PICOC methodology [81] was used to further refine the search string. The search query used can be found below:

"chatbot" OR "conversational agent" OR "relational agent" OR "virtual agent" OR "intelligent agent" OR "cognitive agent" AND "anxiety" OR "depression" OR "mental health" OR "stress"

Preliminary searches on a wide range of databases were conducted, including querying Scopus, PubMed, EBSCOHost, Springer, the ACM Digital Library, IEEE Xplore, Google Scholar, Web of Science, EmBase, PsycINFO, Cochrane, CINAHL, Science Direct, and Inspec. However, due to its role as an aggregator of diverse scientific works, Google Scholar offers broader coverage compared to specific databases with restrictive inclusion criteria. Therefore, it can serve as a viable alternative to the mentioned databases. This insight is consistent with empirical studies on database comparison [82]–[84]. Therefore, only Google Scholar was used and complemented with a database search software Harzing's Publish or Perish, which is recommended for easier querying [85].

The author of this work relied on their experience and knowledge of the field as well as having a clear idea of related work to construct a list of special criteria to apply to the paper selection process. Every decision was elaborated to avoid arbitrary or biased criteria. All the full papers that passed all the items on the special criteria list were included.

The special criteria include the following items:

1. **Targeted mental health issues in the paper include stress, anxiety, depression, or general well-being.** This criteria was selected as these are the most common mental health issues among the nonclinical population [86], they are seeing the most rise [87], they are targeted most by the systems of interest, and they are the easiest to target with technology [21]–[23].
2. **The system in the paper is autonomous and not 'Wizard-of-Oz'.** The 'Wizard-of-Oz' technique refers to the "seemingly autonomous application whose unimplemented functions are actually simulated by a human operator, known as the Wizard of Oz" [88, p. 7]. Since technologies investigated should enable exactly such functions, including 'Wizard-of-Oz' systems would defeat the purpose of this work.
3. **The conversational model of the system in the paper is text-based.** Text-based systems were selected due to experts calling for such systems [24], due to the belief that text-based systems are the most mature in the technological landscape and therefore more amenable to being reviewed, and due to the extensive range of such systems (e.g., speech-based systems means analyzing a completely different technology), it is impractical to comprehensively cover them within a single work.
4. **The conversational model of the system in the paper allows for a synchronous, real-time two-way communication.** This criterion was selected due to the power of a dialogue in the matters of mental health [89], which is compelling to research such systems, as well as the trending usage of ICAs in various areas of service [90], whose success also stems from the convenience of synchronous communication, seen in instant messaging systems [91].
5. **The paper describing the system was published in the last 5 years.** Since technology is developing fast, the last five years are recommended by other researchers and is not uncommon in SOTA reviews [79], which should cover the trends of interest.

6. **The system in the paper is implemented to be used with a computer or mobile devices.** This criterion – as opposed to also covering, e.g., robotic platforms – was selected due to wanting to overview conveniently available systems, which do not demand additional resources for being accessible.

7. **The system in the paper is fully functional, not a part of a bigger cognitive architecture of an ICA.** The power of ICAs lies in their emergent behavior when multiple parts or modules work in concert towards producing ABC. What is important for this work is the system as a whole, not individual parts (e.g., not singular ML models).

8. **The system in the paper is not only a design, but was implemented and can be used.** Systems that are possible to build are of this work's interest. Only implemented systems can answer some of this work's questions (e.g., RQ4), especially on results that such systems produce (e.g., RQ8). Without this criterion, the true technical trends of the field cannot be sufficiently addressed.

9. **The paper provides an adequate level of technical detail, enabling a comprehensive analysis of the system from a computer science perspective.** To be able to conduct a SOTA review, this criterion is necessary. Without it, barely any RQ can be addressed.

10. **The system in the paper is non-proprietary.** Many systems (or platforms used to build the systems) used in the most (cited) studies [54], [92], [93], are non-proprietary. The most well-known systems or platforms are: Tess [92], Wysa [93], Woebot [54], DialogFlow [94], IBM Watson [95], Microsoft Bot Framework [96], and GPT-3 [97]. While this work considers some of them, unfortunately, proprietary work cannot be surveyed as their technologies are closed source and not described in enough detail to be able to analyze them. They function like a double black box, whereby not only the neural networks utilized for their conversational models cannot be discerned, but also no other methodological or technological details about them can be discerned. They also do not foster open source and transparent research work. However, given that Woebot is utilized as a comparative reference (due to other reviewed systems not being implemented and ready for use) within an empirical interventional experiment (described in section 7.2) against the system developed in this work, a brief description of it is provided at the end of this chapter.

Apart from the specific criteria list to apply to paper selection, general criteria were constructed, partly guided by the PICOC method [81].

The steps that were followed for paper selection:

1. Use of Harzing's Publish or Perish for easier management

2. **Exclusion criteria:** Papers do not address "Conversational Agents" and related acronyms (population criterion I)

3. **Exclusion criteria:** Papers do not address "Stress," "Anxiety," "Depression," or similar words (intervention criterion II)

4. **Removal of impurities:** Deleting theses, dissertations, non-scientific papers, posters, review papers, books, papers with three pages or less in length
5. **Quality assessment:** Focusing only on peer-reviewed published papers in journals and conferences (conferences hold special importance in computer science)
6. **Abstract and text filtering:** Special selection criteria, not applied before, described under the special criteria section

Removal of duplicates was not strictly necessary due to the use of one database, but since there might be various sources for the same paper (e.g., a journal and a university website), they were removed in one of the steps (e.g., step 3) or by hand when encountered.

PICOC was used to refine the criteria to be transparent and unbiased for the final paper selection. Inspiration was taken from the PRISMA framework [98] for reporting and the PRISMA diagram used to visualize the process.

Data extraction was focused on identifying keywords and parts of the text that help answering the Qs, deemed relevant to the review's narrative and goals.

The paper selection process used various filtering methods to improve the results that fit the objectives of this review and help to answer this work's Qs. The process included the following steps: using Harzing's Publish or Perish for easier management; ad hoc removal of duplicates; application of exclusion criteria; removal of impurities (deleting theses, dissertations, non-scientific papers, posters, review papers, books, papers with three pages or less in length), application of quality assessment criteria, and abstract and text filtering.

The paper selection process with the numbers of papers encountered in each step was:

Step 1: Querying Google scholar with search string: $n = 14300$

Step 2: Using Harzing's Publish or Perish, applying exclusion criteria (population criterion I and II): $n = 254$

Step 3: Removal of impurities, quality assessment: $n = 114$

Step 4: Filtering: $n = 10$ (number in line with similar review papers)

The PRISMA diagram in Figure 2.1 visualizes the process. The diagram follows the PRISMA methodology [98].

The selection process yielded 10 papers that aligned with the reviewing criteria. These papers represent various approaches to achieving change in people with mental health issues. Since all of them feature full cognitive architectures for their systems, some of the latter's parts are homogeneous among the papers, while others are very heterogeneous. The systems in this review show that there are multiple ways of doing that, which gives the research field the flexibility and diversity. The two are needed for more possibilities for progress.

The reviewed works are:

1. Delahunty et al. [50] proposed a diagnostic ICA, which combined conversational abilities with machine learning and clinical psychology. It used sequence-to-sequence neural networks for dialogue generation and machine learning classifiers for discovering depression symptoms. The goal was to facilitate crisis support for depressed people.

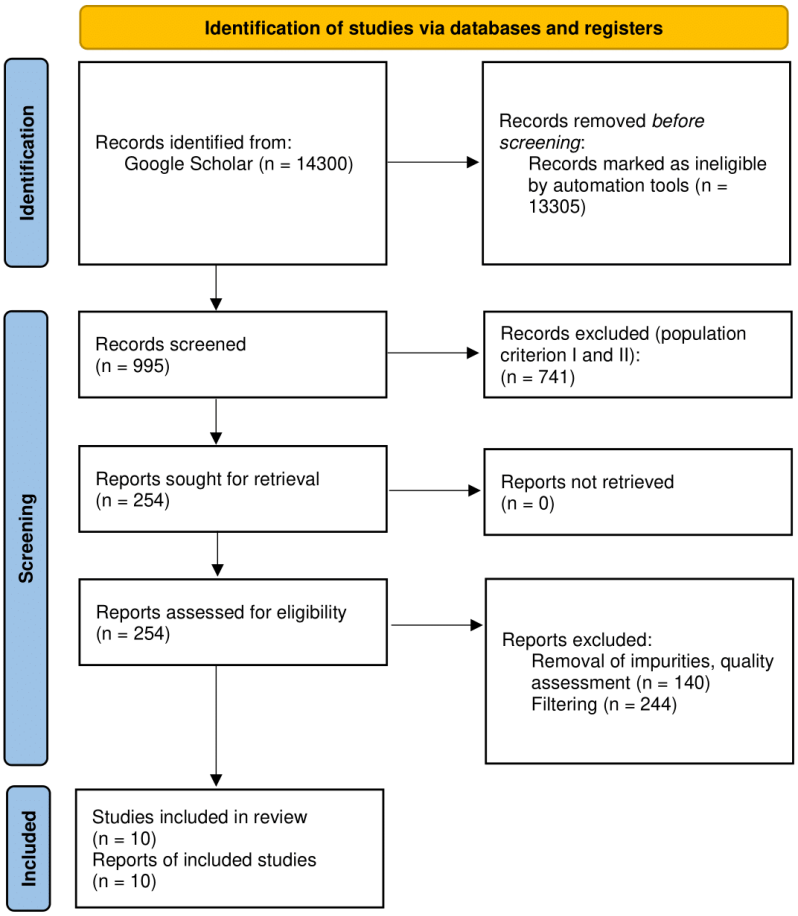


Figure 2.1: PRISMA diagram of the paper selection process.

2. Denecke et al. [51] introduced SERMO, an ICA that combined methods from cognitive behavior therapy (CBT) and lexicon-based emotion recognition to support general well-being in people by regulating their emotions, thoughts, and feelings. Emotion recognition in SERMO was crucial for effective strategy selection in terms of proposed activities and dialogue help. Alongside, informational strategies helped provide people with psychoeducation. User evaluation with the User Experience Questionnaire showed that the system was considered good.

3. Ghandeharioun et al. [99] focused on delivering ecological momentary interventions through an ICA to raise people's general well-being by relieving SAD symptoms. The system EMMA provided emotionally appropriate interventions in an empathetic manner, detecting user's moods solely through the smartphone sensor data, which was integrated with the ICA. Their results showed that their personalized machine learning model, used to determine the moods, was likable by the participants.

4. Khadikar et al. [100] developed Buddy, an ICA that targeted general well-being by treating symptoms of SAD, but also working as a motivational companion to help with loss of focus. The system used recurrent neural networks (RNNs) to respond to the users' emotions with appropriate dialogues that built mental resilience and drove the conversation towards positive thoughts.

5. Morris et al. [57] designed an ICA that simulated human capabilities in empathy expression. They repurposed online peer support data, which the ICA through corpus-based approaches presented to the user. Information retrieval and word embedding techniques produced the best matches to the user's concerns. In a controlled experiment, the users found such responses acceptable.

6. Park et al. [101] delivered a prototype ICA Bonobot that used motivational interviewing methods to help students cope with stress. It used conversational sequences to guide the users through the motivational interviewing processes, providing evocative questions, encouraging feedback, and reflective and affirming responses, placed in the context of the users' problems. The major focus of Bonobot was discussing the idea of change. When used in an experiment, participants were satisfied with the ICA, but pointed out that more personalized feedback and informational support would benefit the system.

7. Pola and Chetty [102] created an ICA that offers behavioral therapy to people with depression. The ICA tried to get information from the user on their mental state. It could detect seven types of emotions from text using long-short-term-memory neural network and a pre-trained weighted word index known as glove2. The ICA's main strategy was trying to have a dialogue about the users' negative thoughts and offer different perspectives on them.

8. Rishabh and Anuradha [103] built three different ICAs for general well-being, using different technologies. The first, based on the famous psychotherapeutic chatbot ELIZA [53], used retrieval approaches for its language capabilities. The second, based on another famous chatbot, ALICE [104], used AIML (Artificial Intelligence Markup Language). The third used generative approaches. All of them tried to gauge the context that users conveyed to them through text and guide the conversation towards more positive sentiment.

9. Yorita et al. [72] proposed a stress management framework with an ICA platform working on computers, mobile devices as well as in robots. It derived various stress

measures and modeled their users, which determined the strategy selection in their peer support model. Interventions targeted various factors that aim at different stress management skills. The process was driven by reinforcement learning in combination with fuzzy control. Their results show that after using the ICA, people displayed better skills at dealing with stress.

10. Yorita et al [105] built on the ICA from Yorita et al. [72], expanding the models and employed strategies for help to personalize their system even further.

2.1 General Overview of the Selected Related Works

This section highlights key general findings that form the basis for addressing the research questions about the reviewed works.

Figures 2.2–2.5 represent the technical summary of the reviewed papers. Figure 2.2 shows the number of papers that featured ICAs with conversational models based on neural networks, being the most popular generative method for natural language understanding and generation, and the amount based on rule-based or other machine learning types. Figure 2.3 shows the number of papers that featured ICAs with non-conversational models (e.g., classifiers for stress level) based on neural networks and the amount based on rule-based or other machine learning types. Figure 2.4 shows papers that featured ICAs that used various methods to personalize and adapt their actions. Figure 2.5 shows the number of papers that featured ICAs that built their own complete cognitive architectures, and the amount that used existing (open source) platforms to create their architecture or that used existing ICAs and upgraded them.

Figures 2.6–2.8 represent non-technical summary of the reviewed papers. Figure 2.6 shows the number of papers that featured ICAs tackling specific mental health issues. Figure 2.7 shows the number of papers that featured a user study on relieving SAD, that featured a user study on the system, and that were only evaluated by the authors. Figure 2.8 shows the number of papers that featured ICAs that only did assessment, that only did intervention, and that did both.

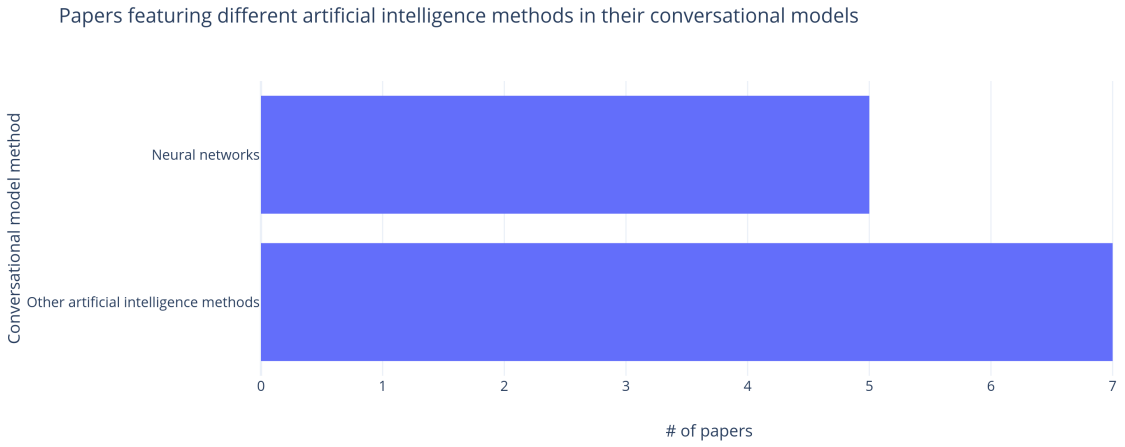


Figure 2.2: Papers featuring different AI methods in their conversational models. Some systems use neural networks as well as other AI methods, which puts them into both categories.

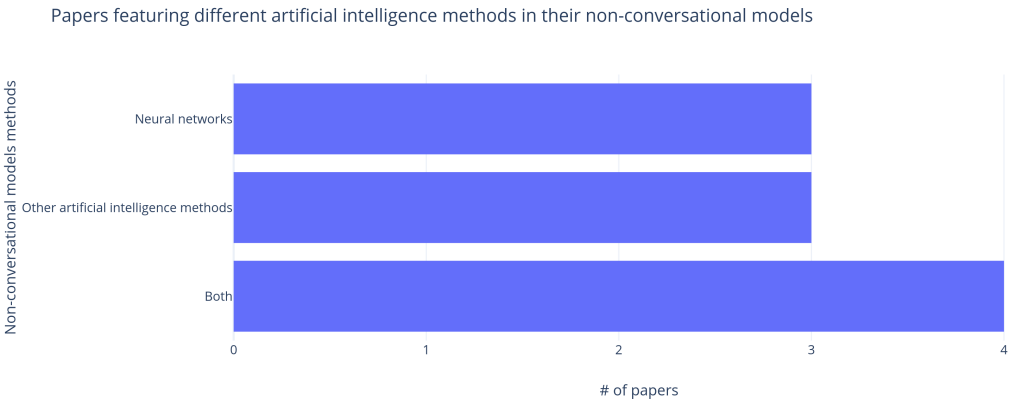


Figure 2.3: Papers featuring different AI methods in their non-conversational models.

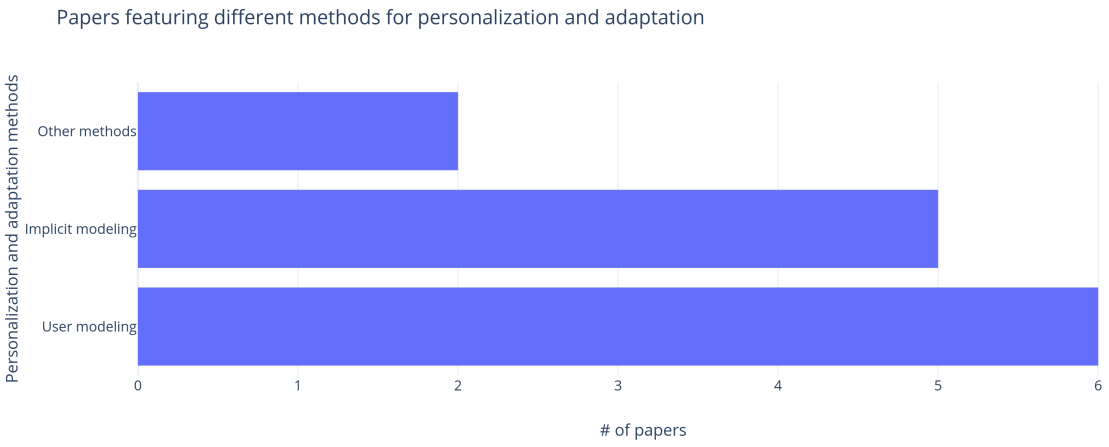


Figure 2.4: Papers featuring different methods for personalization and adaptation. Implicit modeling represents language understanding and generation methods, as ICAs personalize output by, e.g., recognizing emotions in the input.

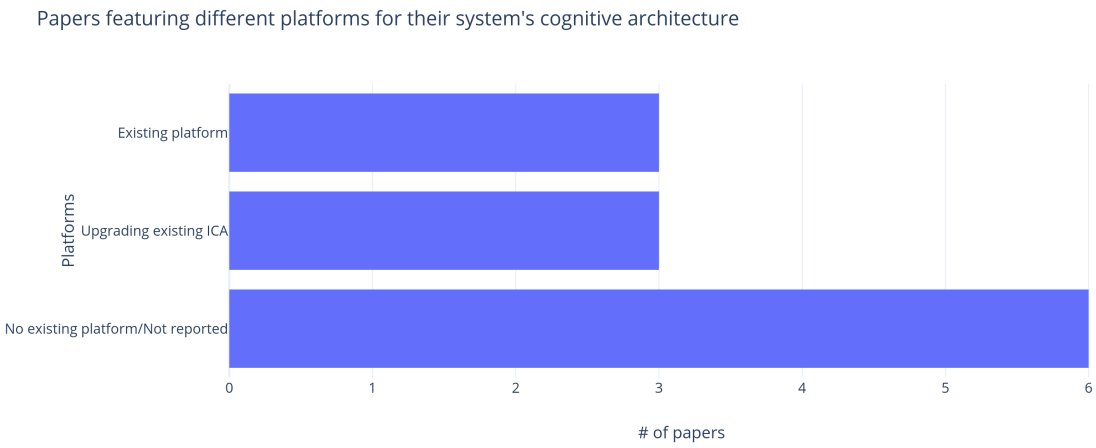


Figure 2.5: Papers featuring different platforms for their system's cognitive architecture. "Upgrading existing ICA" denotes using existing instances of architectures and upgrading them (e.g., ELIZA [101], [103]).

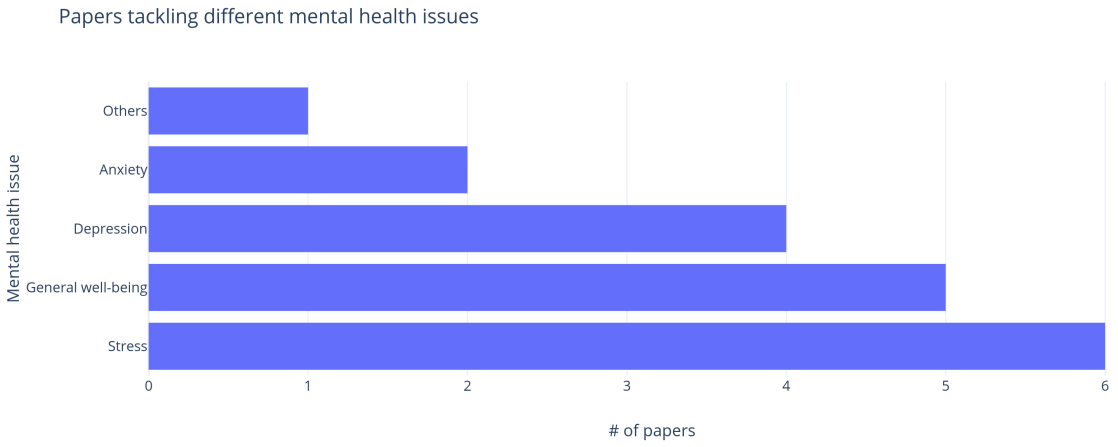


Figure 2.6: Papers tackling different mental health issues.

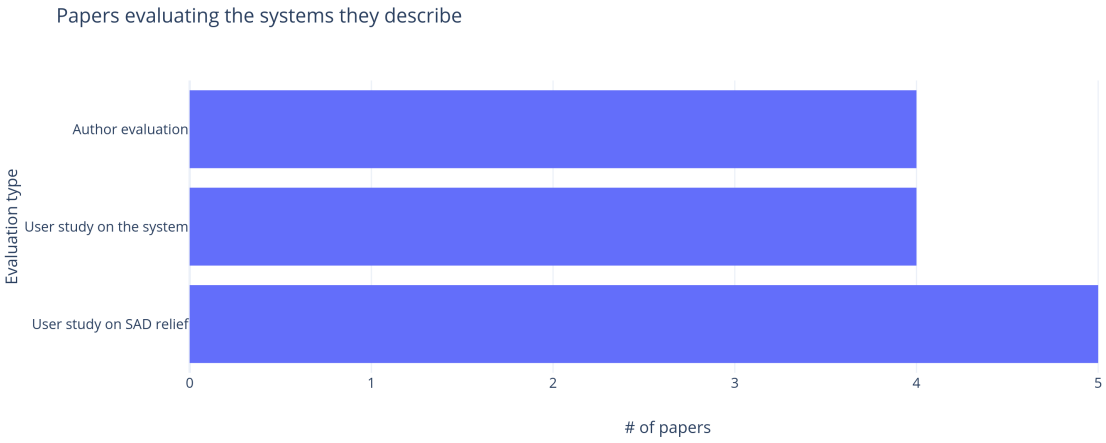


Figure 2.7: Papers with different system evaluations.

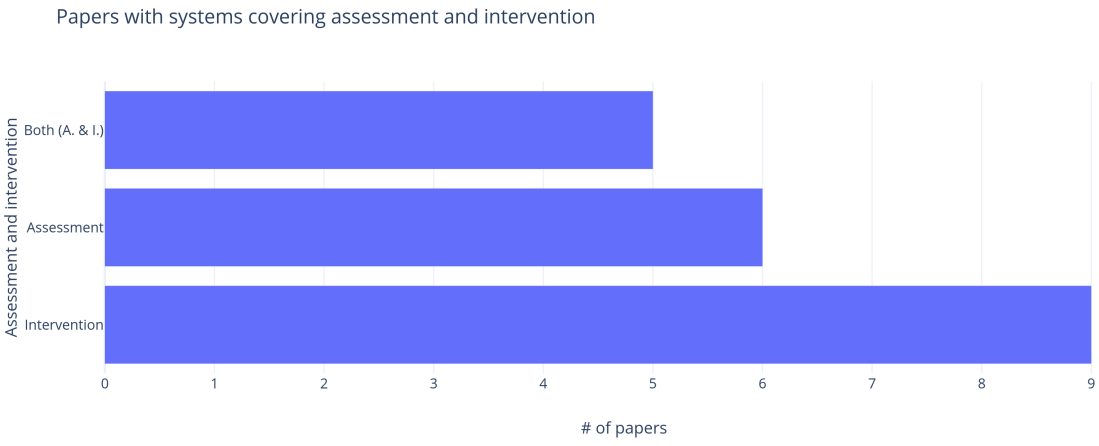


Figure 2.8: Papers with systems covering assessment and intervention.

2.2 Analysis of the Selected Related Works

To answer Q1, the reviewed works were scanned for information on which mental health issues they target (see Table 2.1). This information was mostly presented in the titles, although sometimes it was more implicit, e.g., in the data collection and intervention techniques used.

Table 2.1: Answering Q1. Which mental health issues do the systems target?

Work	Targeted Mental Health Issue
[50]	Depression
[51]	General well-being
[99]	General well-being, stress, anxiety, depression
[100]	General well-being, stress, anxiety, depression, loss of focus
[57]	General well-being, stress
[101]	Stress (in students)
[102]	Depression
[103]	General well-being
[72]	(Occupational) stress
[105]	See [72]

To answer Q2, the reviewed works were scanned for information on which data was collected from the users by the authors and their systems, which datasets the authors used to train or augment their systems, what methods the systems were built on to produce ABC for SAD, and overall technologies used. All the listed had to have a specific purpose in producing ABC as opposed to, e.g., general conversational abilities of the system. There is a general process in treating mental health issues, which widely consists of two steps: assessment and intervention [30], [106].

Assessment:

- [50]: The system tries to classify depression, suicidal ideation, insomnia and hypersomnia, weight change, and excessive or inappropriate guilt from linguistic user input. It trains on various datasets (eRisk, Reddit posts from users and subreddits). It extracts linguistic features from the text and uses doc2vec to vectorize it, employing feature recognition and text embedding approach to construct classifiers. It finally applies Random Forest and logistic regression to predict the presence or absence of depression symptoms. The overall F1-Score for classifiers was 0.91.

- [51]: The system uses a lexicon-based approach using SentiWS lexicon to conduct sentiment and emotion recognition in the linguistic user input. It further applies fuzzy matching to recognize emotions from words that are similar enough to convey the same meaning. The system achieved 81% accuracy in recognizing emotions in a dataset of forum posts.

- [99]: The system collects geolocation data from a phone, connected to the ICA, and user ID, gender, baseline scores of the big five personality test, PANAS (Positive and Negative Affect Scale, short version), and DASS (Depression, Anxiety and Stress Scale). PANAS quantifies mood and DASS captures depression, anxiety, and stress symptoms. It applies experience sampling five times a day using a visual grid based on Russel's two-dimensional model of emotion to capture ground-truth labels. The affect is inferred by the system using personalized model with Random Forest regression for

valence prediction (82.4% accuracy) and AdaBoost regression for arousal prediction (65.7% accuracy).

- [100]: The system does not explicitly assess users and uses no specific assessment methods. Assessment is implicit in the linguistic intent recognition in the conversational model.
- [57]: The system does not explicitly assess users and uses no specific assessment methods. Assessment is implicit in its matching capabilities where the user input is matched with the closest reply from the used database.
- [101]: The system does not explicitly assess users and uses no specific assessment methods. It uses evocative questions to collect linguistic user input. Afterwards, it uses keywords from the linguistic user input which guide the conversation – these keywords can convey mental states. The keywords were acquired from a dataset that collected data from Reddit subreddits.
- [102]: The system uses questions that target emotional states of the users to gain relevant user input. It then uses a model to detect seven types of emotions. The model uses long-short-term-memory neural network with glove2 for emotion recognition. The model is trained on the ISEAR dataset. The accuracy of emotion recognition obtained was 84%. Furthermore, it labels users into five states according to the detected emotional levels: zero depression, slightly stressed, highly stressed, slightly depressed, highly depressed.
- [103]: The systems do not explicitly assess users and use no specific assessment methods. Assessment is either implicit in the linguistic intent recognition in the conversational model or it uses keywords from the linguistic user input which guide the conversation.
- [72]: The system uses fuzzy inference to evaluate the content of the linguistic user input as replies to various intentional questions and to detect users' state of stress. The users are measured on Comprehensibility, Manageability, and Meaningfulness. "Comprehensibility means that people can understand their situation and predict their near future. Manageability is a sense that people can manage their situation. Meaningfulness means people can understand the meaning of their life." [72] (p. 3763) This determines users' Sense of Coherence model, which is used for various strategies to increase stress management.
- [105]: See [72]

Intervention:

- [50]: The system does not deliver interventions and has no specific intervention methods.
- [51]: The system delivers suggestions for activities and exercises that help regulate emotions in the form of a dialogue, reminds the user on appointments and implements CBT techniques, e.g., mindfulness and focusing in goals. The dialogues vary depending on detected emotions and are mostly of informational nature.
- [99]: The system delivers well-being interventions which include individual or social activities from a range of psychotherapeutic categories: positive psychology, cognitive behavioral, meta-cognitive, or somatic interventions. They are delivered through

a textual prompt to the user with various digital tools to engage with the activity. The dialogue the system produces is based on emotions detected by selecting a random pre-written script from an emotional category, congruent with the user's state (e.g., if a person is identified to have emotions of low valence and arousal, the system produces the following: "Feeling glum? I have a skill that might brighten your day. Let us practice.").

- [100]: The system delivers interventions in the form of positive drivers inserted in the conversation to change the trend of the users' thoughts. It also targets self-expression development and stress management. CBT techniques, motivational interviewing and analysis, positive behavior support, behavioral reinforcement, and guided actions and methods are used to encourage the user to build emotional resilience skills. Actions are encouraged at different moments, such as meditation.

- [57]: The system delivers interventions in the form of preexisting emotional support statements, drawn from a large corpus of online interactions from the Koko platform, a platform that connects users seeking help and those who have opted to give help. The users needing help also evaluate the responses. This corpus-based approach tries to create the semblance of personalized, empathic expression. The system uses information retrieval techniques and word embeddings to automate this process in real-time, matching existing statements to appropriate inputs by the users, selecting texts that have satisfactory scores. The interaction between the system and the user is one-off – the user describes their situation and the system matches a reply from the dataset. The answers are presented as if authored by the system.

- [101]: The system delivers interventions in the form of motivational interviewing. It can only use predefined responses, which depend on the stage of the process the user is in. These stages are Engaging, Focusing, Evoking, and Planning, where: "In Engaging, Bonobot shares brief introductions with the user and gives instructions to use the chatbot. In Focusing, Bonobot asks the user to detail their problem, possibly having them identify an inner struggle. This leads to Evoking, where Bonobot explores future goals with the user, affirming their own ideas for change. Finally, Bonobot invites the user to ponder the overall session in Planning." [101] (p. 3) The process helps users cope with stress and encourages self-reflection.

- [102]: The system delivers interventions in the form of emotional conversational support, suggesting different, more positive perspectives on situations the users describe, and trying to prevent negative thoughts. The conversation is guided by the level of mental health issue detected.

- [103]: The systems deliver interventions differently. ELIZA-based ICA uses Rogerian reflection to engage with the users. Information retrieval techniques are used to choose proper responses: the n-gram technique, charagram embeddings, word similarity, sentence similarity, and part-of-speech tagging. ALICE-based ICA delivers interventions by sympathizing with the user and using CBT techniques. It implements AIML, sklearn to match responses, as well as category tagging and synonym switching for conversational dynamicity. The generative language ICA only implicitly delivers interventions by being trained on empathetic text.

- [72]: The system delivers interventions that help improve the users' self-efficacy, which helps manage stress, as it measures users' sense of task performance and whether they feel they can do a task or not. The system, drawing from the user model which is based on the Sense of Coherence (SOC) model, engages the Peer

Support model, which finds suitable support types and delivers them. The system uses reinforcement learning and fuzzy control the find the best Peer Support types for specific SOC models. Peer support also stimulates various aspects of a person to lower stress levels. The types of support are helper therapy (the user takes the role of the carer instead of being cared for), informational support, esteem support, and emotional support.

- [105]: See [72]. The authors upgraded the system with expanding the helper therapy support type by the user having to be a carer offering either informational or emotional support, depending on their SOC.

To answer Q3, the reviewed works were scanned for information on which methods were used to build conversational models in the reviewed systems. Generally, there are two approaches: rule-based, dialogue tree conversational models with either free text or button-based user input options (more control, less errors, but limited conversational experiences), and generative models with free text options (less control, more errors, more affordances for conversation).

- [50]: The system's conversational model was trained with seq2seq (OpenNMT) learning approach on datasets from Reddit's subreddits, the eRisk dataset and OpenSubtitles dataset using neural networks.

- [51]: The system's conversational model is built on the Syn.Bot framework, which uses Oscova as the bot development platform and the SIML (Synthetic Intelligence Markup Language) interpreter. The model lets the users frame answers in their own words and select predefined answers.

- [99]: The system's conversational model works on textual prompts and scripted phrasings that are utilized at contextually appropriate times.

- [100]: The system's conversational model uses RNNs for learning as well as understanding and generating responses. The intent in the user input is recognized by the Long-Short-Term-Memory neural network.

- [57]: The system's conversational model consists of two modules. The front-end module pairs previous responses with user inputs. The back-end module generates output using Elasticsearch, word2vec and a word-embedding procedure. The authors used the Google News dataset for training. The ICA also solicits user feedback.

- [101]: The system's conversational model extends on ELIZA, basing its functionalities on identifying user keywords to generate responses. It consists of two modules, Flow Manager and Response Generator. Flow Manager runs the conversation and assigns template responses to lead the user. Response Generator follows the conversational flow and sequences, identifying keywords by weighting them and assembling responses.

- [102]: The system's conversational model is built by the authors using pre-trained weighted word index known as glove, and trained responses to create an environment for generative, free-text conversation.

- [103]: The three systems' conversational models are built with three different approaches: 1) the Retrieval Pattern Matching ICA is built on ELIZA, using the n-gram technique to get relevant responses, Charagram embeddings to learn character-based

compositional models to embed textual sequences, and using word similarity, sentence similarity and part-of-speech tagging for evaluation; 2) Retrieval Rule Based AIML ICA is built on ALICE, using sklearn alongside the AIML library and various rules to generate a response; 3) the generative ICA learns on the data from The Open American National Corpus, using the Long-Short-Term-Memory method and context learning for understanding input, and using Beam Search to choose a response.

- [72]: The system's conversational model is rule-based, basing its responses on a stored databank. The user can communicate by inputting free text or by selecting fixed inputs. The outputs are also based on the classification of the moods of the users, detected through using machine learning (see Assessment).
- [105]: See [72]

To answer Q4, the reviewed works were scanned on how the ICAs were built. The focus was on whether various platforms were used to produce the ICA (e.g., Rasa [107]) or whether an existing ICA and its framework were used and possibly upgraded (e.g., ELIZA). As this was one of the exclusion criteria, papers with ICAs built on proprietary, closed code platforms (e.g., DialogFlow) were not considered. Table 2.2 presents the results and the answer to Q4.

Table 2.2: Answering Q4. What are the platforms used to create the systems?

Work	**Platforms and Frameworks**
[50]	No existing platform or framework/Not reported
[51]	Syn.Bot, OSCOVA
[99]	StudyPortal platform (extricated from [108])
[100]	No existing platform or framework/Not reported
[57]	No existing platform or framework/Not reported
[101]	Extended ELIZA framework
[102]	No existing platform or framework/Not reported
[103]	Extended ELIZA framework, extended ALICE framework
[72]	No existing platform or framework/Not reported
[105]	LINE Platform

To answer Q5, the reviewed works were scanned on what domain knowledge, particularly from mental health and ABC theories, is somehow integrated into the systems. This may be through the strategies that the systems deploy to produce ABC, e.g., CBT techniques, or through user modeling, where knowledge on SAD helps make the systems more empathetic.

- [50]: See Q2: Assessment
- [51]: The system reflects knowledge on emotions, tracking and monitoring (diaries), and CBT techniques like mindfulness and activities.
- [99]: The system reflects knowledge on positive psychology, cognitive behavioral, meta-cognitive, or somatic interventions as well as emotion theory like Russel's circumplex model.
- [100]: The system reflects knowledge on "self-help practices such as CBT, motivational interviewing and analysis, positive behavior support, behavioral reinforcement and guided actions and methods to encourage the user to build emotional resilience

skills. It helps the user to manage their stress, anxiety, overthinking, energy, helps in focus, promotes meditation and encourages the same, and other situations." [100] (p. 122)

- [57]: The system reflects no explicitly discernible domain knowledge. It has implicit lay knowledge that is based on the online peer support data.
- [101]: The system reflects knowledge on motivational interviewing, stress management, and self-reflection.
- [102]: The system reflects knowledge on the emotion theory (seven basic emotions), emotional support and evocative questions.
- [103]: The system reflects knowledge on Rogerian reflection.
- [72]: The system reflects knowledge on the SOC model, Generalized Resistance Resources, helper therapy, informational support, and emotional support.
- [105]: See [72]

To answer Q6, the reviewed works were scanned on what kind of data is collected on the users for the user model, and how the user is further modeled. Interest was also taken in how the working of the system is affected, particularly in terms of how the system is personalized and how it adapts to individual users.

- [50]: See Q2: Assessment
- [51]: The system builds the user model on the emotion data, which it uses to personalize dialogues.
- [99]: The system builds the user model on the following data: "user ID, gender, baseline scores of the big five personality test, PANAS (Positive and Negative Affect Scale, short version), and DASS (Depression, Anxiety and Stress Scale). PANAS quantifies mood and DASS captures depression, anxiety, and stress symptoms." [99, p. 16] It also contains data on "experience sampling five times a day using a visual grid based on Russel's two-dimensional model of emotion." [99] (p. 16) It uses this data to select among different emotionally charged phrasings.
- [100]: The system does not build any explicit user models.
- [57]: The system does not build any explicit user models.
- [101]: The system does not build any explicit user models.
- [102]: See Q2: Assessment
- [103]: The systems do not build any explicit user models.
- [72]: The system builds the user model on the following data: data from the SOC model, Perceived Stress Scale, Ryff's Psychological Well-Being Scales, and Hassles Scale. Each user has a continually updated SOC model. Generalized Resistance Resources connect other data to the SOC model.
- [105]: See [72]

To answer Q7, the reviewed works were examined to identify any explicit references to a specific cognitive architecture (e.g., Belief-Desire-Intention architecture) that was employed during the system's construction. In case they did not, an interest was taken in identifying the modules that constitute the cognitive architecture.

- [50]: Not specified; modules for detecting various mental health issues, conversational model for question formation
- [51]: Syn-Bot architecture (including OSCOVA and SIML)
- [99]: Not specified; geolocation-emotion prediction module, personalized textual interventions module
- [100]: Not specified; language learning module (RNN), user understanding module (NLP), response generator (NLP) with psychological techniques
- [57]: Not specified; pairing module, user feedback module
- [101]: Not specified; flow manager, response generator
- [102]: Not specified; mental state classification module, response generator, user model
- [103]: Not specified; ELIZA-based system: pattern matching module, response generator; ALICE-based system: self learning module, response generator; Generative system: training module, context module, generalization module, response generator
- [72]: Belief-Desire-Intention architecture
- [105]: See [72]

To answer Q8, the reviewed works were examined to assess the evaluation of the systems, with a specific focus on user-tested evaluations. The mental health outcomes subsequent to using the system were ideally desired to be observed. Nevertheless, data on user evaluation pertaining to assessing the properties of the system was also extracted.

- [50]: No evaluation on users
- [51]: Tested on users and mental health professionals on the system's Attractiveness (users: below average; professionals: good), Perspicuity (users: above average; professionals: above average), Efficiency (users: below average; professionals: above average), Dependability (users: bad; professionals: below average), Stimulation (users: bad; professionals: above average), and Novelty (users: below average; professionals: excellent).
- [99]: No evaluation on users
- [100]: No evaluation on users
- [57]: Tested on users where they compared the system's replies to their peers' replies with three scores: *good* (system: $>40\%$; peers: $>60\%$), *ok* (system: $<40\%$; peers: $<40\%$), and *bad* (system: $>20\%$; peers: $<10\%$).
- [101]: Tested on users where they described the system as having evocative questions and offering self-reflection as well as potential consolidation, but noted that the feedback was clichéd. The users also wanted more informational support from the system and more suitably contextualized feedback.

- [102]: No evaluation on users
- [103]: No evaluation on users
- [72]: Tested on users which used the system for five days. The system managed to improve their scores on stress managing skills, reflected in the SOC model.
- [105]: Tested on users which used the system for three days. The system managed to improve their scores on stress managing skills, reflected in the SOC model.

To answer Q9, the reviewed works were examined to assess to what extent we can infer that these systems embody theory of mind. Considering human cognitive capabilities that come with ToM [109], ICAs with ToM would generally exhibit the following:

- Understanding: This would enable an ICA to understand and interpret human emotions, beliefs, intentions, and similar mental states. It should decipher not only the semantic meaning of the user's input but also, e.g., the emotions and intentions behind the text.
- Mental modeling: An ICA with ToM should have an ability to create and update a model of the user - their mental states, beliefs, desires, and intentions based on their interactions, to form a sort of internal "mind map" of the user.
- Knowledge: To execute actions stemming from a user's understanding due to ToM, an ICA must possess sufficient knowledge in the relevant domain.
- Prediction: Using the model it has created of the user's mind, an ICA should be able to predict the user's future state of mind, their actions or reactions. This would allow the ICA to better tailor its responses and suggestions. In terms of mental health, it should know which responses will create better outcomes based on an individual.
- Empathy: An ICA needs to be able to show empathy. This involves recognizing the mental state of the user and personalizing the wording or the manner of the response to it.
- Learning: As an ICA interacts with a user over time, it should learn and adapt based on those interactions. Learning enables the ICA to refine its understanding of the user and improve its predictions.
- Response generation: An ICA needs to be able to respond to the user in a way that demonstrates its understanding of the user's state of mind while being able to bring together all of the above for a suitable action.

The systems from the selected related works were evaluated in the light of the list above.

- [50]: The system seems to focus on understanding through the application of machine learning, showing high accuracy in recognizing depressive symptoms from user input. Evidence of learning is not clear, as it does not seem to have capabilities of learning from the user. It appears to lack comprehensive mental modeling, as it primarily recognizes and classifies symptoms rather than building a detailed understanding of a user's mental state. The system does not deliver interventions, and therefore lacks response generation capabilities. Its abilities to exhibit empathy or domain-specific knowledge are non-existent. Overall, it shows very few elements of ToM.

- [51]: The system showcases understanding and prediction by effectively recognizing and responding to user emotions, as well as possessing domain-specific knowledge in cognitive behavioral therapy techniques. Its ability to adapt dialogues based on emotions indicates an approximation of empathy. However, the depth of its mental modeling, learning and empathy capabilities is non-existent or very rudimentary. The system does seem to be more holistically designed, yet does not embody a ToM.

- [99]: The system demonstrates understanding by detecting user moods, and utilizes this information in a mental modeling approach to create personalized responses. It implements domain-specific Knowledge in its delivery of various psychotherapeutic interventions. However, the system's pre-written and scripted responses might limit its ability to demonstrate a nuanced understanding of a user's state of mind, which is also why it lacks empathy and learning capabilities. Therefore, while the system demonstrates some ToM components, its lack of flexible response generation that would arise from also using empathy and learning inhibits a fully realized ToM.

- [100]: The system demonstrates some understanding by using recurrent neural networks to recognize users' emotional states. This approach, while not explicitly building a user model, does entail implicit mental modeling by generating responses based on users' current emotional states. However, this seems to occur within the context of the conversation and does not extend beyond it, which also prohibits learning. The system also showcases significant domain knowledge in delivering interventions grounded in CBT, motivational interviewing, and behavioral reinforcement. Its response generation is dynamic, with the ICA adjusting its dialogue based on the user's emotional state, but it is not really predictive or empathetic as the system does not build an explicit user model. Overall, while the system displays components of ToM, the lack of explicit user modeling, which would also be used in learning, and predictive assessment suggests that it is too limited to embody ToM.

- [57]: The system operates on an understanding derived from matching user input to preexisting emotional support statements from a large corpus, showcasing an implicit form of both understanding and mental modeling. However, this appears to be a one-off interaction without taking past interactions or user states into account. The system does not explicitly assess users, and there is no evidence of developing an ongoing understanding of the user or their needs beyond the immediate interaction. The system does not seem to exhibit any substantial domain knowledge beyond matching emotional responses. Moreover, it does not build an explicit user model, limiting its ability to predict future user needs or tailor its responses in an empathetic manner. Given these limitations, the system is far from embodying ToM, as it is predominantly confined to immediate context, one-off interactions, and with no evidence of learning from past interactions.

- [101]: The system displays an understanding of user emotions and intentions through the interpretation of user-inputted keywords, attempting to discern intent and adapt its responses. However, the lack of explicit user assessment and a specific model of user mental states limits its ability to construct and update a comprehensive mental model. Although the system possesses domain knowledge on motivational interviewing and stress management, it is restricted to delivering predefined responses, limiting the degree to which this knowledge contributes to a sophisticated understanding of the user's mind. The capacity for prediction is absent due to the lack of a comprehensive user model. While the system generates empathetic responses

through its process of motivational interviewing, it lacks personalization, an important component of empathy. Learning and adaptation in the system are confined to the weighting and assembly of responses based on identified keywords, which prohibits true forms of learning. Ultimately, the system's capabilities do not realize as a ToM.

- [102]: The system, in terms of understanding, detects seven types of emotions from text using a long-short-term-memory neural network, showing a certain level of comprehension of user's emotions. It also uses evocative questions to derive user mental states, demonstrating an attempt at mental modeling, although this modeling is limited to five distinct states and does not appear to update dynamically with continued interaction. The system holds knowledge about emotion theory and emotional support, which it applies to guide interventions designed to alleviate negative thoughts and offer alternative perspectives, evidencing an element of empathetic response generation. The system, however, lacks any indication of predictive capability, a critical component of ToM, as it does not anticipate user responses or future states. Similarly, it exhibits a limited ability to learn and adapt, with responses generated from predefined questions and responses rather than dynamically evolving throughout interactions. Thus, while the system exhibits several aspects indicative of ToM, it lacks the full spectrum of capabilities necessary for an embodiment of the concept.

- [103]: There are three distinct systems described in this selected work. While there is some form of understanding observed, as the systems attempt to gauge user context and steer the conversation towards positive sentiment, it is important to note this understanding does not extend to modeling mental states. The systems lack explicit assessment or the creation of a mental model, thereby inhibiting deeper user understanding. Regarding knowledge, the ELIZA-based system reflects a comprehension of Rogerian reflection, though this knowledge is not effectively personalized to users, limiting its impact. These systems do not demonstrate predictive capabilities, which inhibits tailoring of future interactions. The empathy exhibited is confined to the ALICE-based system's sympathetic responses and the generative system's empathetic text training. However, without comprehensive understanding and mental modeling, this empathy is potentially limited in accuracy or applicability. Furthermore, the systems do not demonstrate the ability to learn and adapt from past interactions, a key aspect of ToM. Lastly, while the response generation incorporates some understanding, the lack of an evolving user model restricts the effectiveness of these responses. Consequently, these systems exhibit very limited aspects of ToM, and cannot be said to possess it.

- [72]: The system manifests understanding by deciphering users' linguistic input to assess their stress levels, reflecting a rudimentary comprehension of emotions. It models users by continually updating the SOC model, showing an adeptness at mental modeling. The system is knowledgeable in various concepts related to stress management and uses this knowledge to deliver tailored interventions. These interventions demonstrate the ability to predict users' needs, improving their self-efficacy and stress management skills. Empathy is exhibited in the provision of different types of support, such as esteem support and emotional support, and an attempt to meet users where they are in terms of stress. Learning is evident in the application of reinforcement learning and fuzzy control to optimize the selection of Peer Support types. Furthermore, the response generation aligns with the user's mood and SOC model, suggesting an integration of understanding, mental modelling, and knowledge. However, the system lacks the ability to comprehend and interpret any

other mental states beyond stress, and its empathy is therefore limited in scope. It is also reliant on predefined rules and databanks, which restricts the system's ability to generate responses that reflect a deeper understanding of the user's mental states. Thus, while this system exhibits several characteristics of ToM, it does not fully embody the concept, but it is neither far from it.

- [105]: See [72]

The answers to the research questions about the selected related works give a thorough and detailed insight into how the reviewed systems produce ABC for SAD, especially in their underlying technical mechanisms. This is especially relevant to indicate what kind of data should be collected on users, how they should be modeled to personalize and adapt ICAs, how the latter should converse with the users, etc. The tables with results, which allow for easy comparisons, present how to produce change in stress, anxiety and depression with autonomous dialogue systems.

The approaches to ABC, observed in the reviewed systems, considerably vary. It benefits to compare the technical underpinnings of systems targeting the same mental health issue.

Targeting stress, both systems by Yorita et al. [72], [105] produced experimental results in training people to handle stress better. The system achieved this by: having strong theoretical grounds for assessment, which produced user models of the users' stress management skills as well as well-being, gathered through dispatching standardize questionnaires; having personalized interventions according to the SOC factors in the user model; having a rule-based conversational model, which guided the user down appropriate dialogue paths; basing its domain knowledge on a few carefully selected psychological frameworks, such as SOC model, helper therapy, and informational support; and choosing a well-supported cognitive architecture, Belief-Desire-Intention architecture, to build the system on. Other systems targeting stress lacked such comprehensive architecture in terms of its modules. Some built comprehensive user models but lacked the depth of personalized strategies rooted in theory, opting for few pre-written responses [99]; some did not explicitly assess and intervene, opting for approaches that are more dependent on unsupervised understanding of and responding to users [100], [103]; some produced very rigid and static systems based on a lot of top-down elements to assessment and intervention, either through matching with already existing responses [101] or by following a very strict and limited conversational path [57]. It therefore seems that a strong user model with an intelligent combination of rigidness and freedom of assessment and intervention methods through a guided conversation produces best results.

Targeting anxiety, no systems with experimental results targeting symptom reduction were found. Two systems targeted anxiety, but very generally, either through few pre-written responses [99], or by opting for dialogue freedom through a generative conversational model [100]. Ghandeharioun et al. [99], however, built their system technically based on assessment, using Random Forest and AdaBoost with satisfying results to infer mood from a comprehensive user data model, which might be a better option than implicit assessment.

Targeting depression, Delahunty et al. [50] presented the most comprehensive system for depression assessment building various classifiers on depression symptoms used on the input text. Random Forest and logistic regression were used to infer the presence of depression, suicidal ideation, insomnia and hypersomnia, weight change, and excessive or inappropriate guilt. This appears to be a more nuanced way to assess users than opting for general mental health issue labels. However, their system was assessment only.

Ghandeharioun et al.'s [99] and Khadikar et al.'s [100] systems were already covered in the previous paragraphs, and the same evaluation applies here.

Systems targeting general well-being are harder to compare, but Denecke et al. [51] seemed to follow the formula of Yorita et al. in terms of building a comprehensive system with the right combination of rigidness and dynamicity in assessment, intervention and guided conversation. The system's performance seemed to be based on their assessment methods, which used a lexicon approach to extract linguistic features and infer emotions in the text.

In summary, successful systems seem to base their performance on having a user model, explicit and theoretically-backed assessment with classification models (instead of only collecting questionnaire results), explicit and personalized intervention with many strategic possibilities, and dialogue tree conversational model. As in many areas, tasks that call for machine learning are best solved with ensemble methods, such as Random Forest (which is, however, not explainable).

There are a few clear insights into the preferable technologies that the reviewed ICAs are built on. The first noticeable element is the intricate connection between the technology and the goals of such ICAs. Here, it can be discerned that conversational models in most cases are built to be fairly limited in what is otherwise SOTA in the field of chatbots. It has to be limited – mental health counselling is a very delicate matter, and preventing the generative models go out of control should be one of the primary concerns, as making them be complicit in mental health deterioration of the user is a real danger. This was seen in the case of one of the currently most advanced language models today, OpenAI's GPT-3 [97]. GPT-3 was being tested by the tester simulating a patient. When the tester simply wanted to book an appointment with a doctor, GPT-3 acted as a human, understanding the tester's intents with no problems. However, beyond such surface tasks and conversations, GPT-3 started not only to fault, but to exhibit very dangerous behaviors. When the tester expressed that she feels bad and needs help, GPT-3 answered that it can help, and when the tester expressed suicidal thoughts, GPT-3 recommended that the tester killed themselves [56]. Furthermore, ChatGPT, runing GPT-3.5 and GPT-4, is hard-coded to not respond to prompts that ask for any kind of help in terms of mental health. To researchers in this field, this signals not only how careful they have to be, but also that the systems they build have to be very domain-oriented and should limit the linguistic capabilities as reasonably as possible. In the domain of mental health, it is clear that free text capabilities of ICAs are not on the level where they could be feasibly used, and that generally, NLP research is not advanced enough yet to consider it for such domains [110]. When they are used, they have to be largely improved on in very domain-specific ways, making the systems non-scalable. But while the authors of the reviewed works were aware of the dangers of unconstrained textual input, their conversational models and ICAs in general still seemed too limited in what is currently possible. One possible reason why the authors did not implement more advanced possibilities on understanding the user might be convenience and privacy.

The latter may also be the reason why there is so little user modeling and consequential personalization. The systems collect very little data on the users, which makes them static and inflexible in terms of how they can personalize their strategies to the user and adapt to various individual specificities. Since the current systems do seem to employ ABC theories and strategies, personalizing offered help to specific groups that are affected more by specific strategies [111] should be the logical next step in progressing these systems.

Due to the conversational models many times being the most fleshed out part of the reviewed ICAs, their cognitive architectures are not thought out in high detail, mostly embedding only the conversational model. This can cause oversimplification of possibilities

for the system to function, which has its place for certain purposes (very general and quick first help), but does not explore the possibilities that modeling other cognitive capabilities can bring.

Some designs of ICA cognitive architectures [44] have suggested how to sensibly use more advanced technology which might result in better outcomes, but have so far not been implemented or evaluated yet. They emphasize personalization and adaptation through strong user modeling and learning from historical interactions. It is clear that ICAs for ABC in mental health have a lot of space to grow technologically, should there be enough research in the field. The most important lesson to note is that the outcomes such ICAs produce are emergent – they represent a thoroughly researched and thought out result of highly interdisciplinary efforts, but more specifically, their behavior stems from various modules that model different cognitive abilities interacting with each other. This points to researchers needing to cooperate or being interdisciplinary themselves, not only focusing on narrow intradisciplinary or technical knowledge.

2.3 Woebot

This section describes Woebot, a chatbot widely referenced in digital mental health interventions [54], [112]–[114]. Woebot is described due to its inclusion as a comparative reference within an empirical interventional experiment against the system developed in this work (see section 7.2). There were multiple reasons why the author of this work opted for Woebot as a comparative reference: 1) Woebot currently produces the most effective digital interventions among chatbots, and studies on successful interventions with Woebot have been the most replicated; 2) Woebot is implemented and freely available for use; and 3) other systems from selected related works, while open in terms of being technically described and reported on for possible reviewing, lack implementation or sufficient available code to be implemented.

While information on Woebot is significantly more limited in scope as for other selected systems featured in this work, especially in terms of the technical specification, it can nonetheless be shortly described.

The underpinning of Woebot's operation is a decision tree model that accepts natural language inputs and suggests responses. While this mechanism does not incorporate a theory of mind, Woebot is still effective, indicating that productive interventions can be developed even without this attribute.

Woebot's personality is one of its most noteworthy features. It promotes user-system trust and interaction, a crucial element in mental health interventions. The user interface and front-end of Woebot are well-designed.

Woebot also has the capability of delivering a variety of visual outputs, including emojis, graphs, images, and videos. This multimedia approach provides a dimension of interaction that extends beyond text-based communication. Additionally, Woebot implements a supportive strategy by providing users with educational materials, appropriate messages, and pre-scripted advice, dictated by the user's emotional expressions and cognitive distortions.

Chapter 3

Rationale for This Work

As presented in Chapter 1, there is a need for a bigger role of IT technology in the current mental healthcare landscape. Although initial attempts at such integration have yielded some favorable outcomes, the current SOTA falls short of its full potential.

To surpass the SOTA presented in the previous section, this work tries to interdisciplinarily combine AI, attitude and behavior change, and mental health in a novel synthesis, culminated through the knowledge of cognitive science. The main idea is to create a computational simulation of theory of mind (ToM), a human cognitive capability to "understand the thoughts and feelings" [109, p. 528] as well as "attributing thoughts and goals to others" [Ibid.] in order to function and act appropriately. ToM is at the core of this work's system. In many soft domains, it is ToM that distinguishes humans from machines, and while AI can simulate some of its properties there are no systems that fully emulate them. Simulating ToM even in a limited scope can enable using AI for a very specific, mentally dynamic and complex task – ABC in mental health.

The relevance of this work for AI lies in building and combining real-time machine learning-based detection models, where the goal is SOTA accuracy; forecasting models, which do not exist yet in this way in mental health; novel ontologies on mental health and ABC; and recent large language models (LLMs), which did not succeed in mental health in the past [32]. Wrapping LLMs to generate language in prompt generators, which rely on cognitive and personality models of the users with the use of ABC theories, as well as including risk-filtering models, makes sensible motivational message generation possible for SAD symptoms relief. This is because LLMs generally work only on transformers, and expanding them with the above-described framework seems crucial to make such a system effective in the mental health domain.

All of this is enabled through creation of a novel, up to now non-existing golden standard mixed methods dataset with panel data. Quantitative data on mental health and qualitative free text data entries should prove useful for various areas in the intersection between AI, natural language processing, statistics, computational psychotherapy, computational psychiatry, and digital mental health. Apart from model building, this newly collected dataset and its methodology can provide valuable novel insights into change in mental health, which is still very unexplored and vaguely characterized [115]. This helps the intersection between AI and mental health.

Lastly, focusing on the ABC, recent research in behavioral sciences showed that knowing personality types of people makes them more susceptible for influence through appropriate persuasive strategies [17]. This work uses these advances and adapts them for mental health through the use of technology, which makes personalization of strategies very viable. So far, personalization in ABC has mostly been used in static environments (e.g., print advertisements [116]), where it is harder, making strategies less effective and specific.

Chapter 4

Research Goals and Hypothesis

4.1 Research Goals

The main goal of this work was to design and implement a novel artificial cognitive architecture that mimics theory of mind, which is in cognitive science described as the cognitive ability to "understand the thoughts and feelings" [117, p. 528] as well as "attributing thoughts and goals to others" (Ibid.) in order to function in social life. For the system in this work, this ability was more domain-specific, but it served the same purpose – to understand its user to the degree where it can offer effective personalized help for relieving SAD symptoms. Such a design was an interdisciplinary effort to integrate findings from AI, cognitive science, and behavioral sciences. In order to be able to simulate theory of mind, the architecture has to include models, created with SOTA AI methods. This includes models for detecting and forecasting SAD as well as symptoms of these issues from real-time free text, with which the goal was to achieve accuracies above what can be found in the literature. The comparison of the models was based on testing different ML algorithms, such as tree-based methods and neural networks. Recent LLMs (e.g. GPT-3 [31]) were included as well. They served for the linguistic output generation as a response to the input text where the detection and forecasting models were used. The text output was a motivational message, personalized to the user's personality. This was achieved with idiographic personality and cognitive profile modelling in conjunction with behavioral sciences findings on which persuasive approaches work better for which profiles as well as distinguishing specific mental health-related keywords and topics in the language input, similar to topic modelling. Since recent LLMs are not adapted to specific domains as well as prone to risky output [32], the goal in this work was also to build more humane output processes that are less prone to risk. To ensure positive outcomes of a conversation, a loop was implemented where at the end of a specific conversation, the system re-evaluated the user's well-being post support, and offered help through adapted and new strategies if the well-being was not changed, or taught the user new strategies if the well-being had been improved for future use. The whole developed framework therefore generally complements existing LLMs (e.g., transformer-based ChatGPT) with the hypothesis that it makes them more effective.

Another goal was to produce a novel mixed methods dataset with more than 1000 data instances (non-existent prior to this research), a panel data (multiple individuals at multiple time intervals) of daily quantitative questionnaires on SAD (diagnostic-level), accompanied by daily free text diary entries, ideally through a pre-study (to test methodologies and improve them for the main study) and a main study (the latter ethically approved). This ensured golden standard data, needed for this work's system, as well as a dataset that can be of use to the wider research community that is currently lacking such data.

4.2 Hypothesis

There is one main hypothesis (H):

H. An intelligent cognitive assistant for attitude and behavior change for stress, anxiety, and depression will achieve results at least comparable to the state-of-the-art if it simulates theory of mind in a novel artificial cognitive architecture.

Explanation: Theory of mind is simulated as an ensemble of various models and novel ontologies. This includes psychological and cognitive user modelling, mental health and behavior change ontologies, detection and forecasting machine learning models, large language models wrapped in risk detection models, and behavior change prompt generators. Part of the theory of mind is simulated by relying on novel mixed methods panel data with diagnostic-level questionnaires and accompanying quality free text data.

The confirmation or rejection of this hypothesis will be supported by:

1. the recapitulation of accuracy measures of machine learning models (that are part of theory of mind) through computational experiments, compared with related state-of-the-art systems (see Chapter 2);

2. expert measures, defined by standardized questionnaires for such systems;

3. the final experiment involving subjects interacting with different systems to evaluate their influence on mental health. This includes comparing the system developed in this thesis with Woebot [54], the most cited and freely available system with the most replicated positive outcomes. Publicly available information suggests that Woebot does not possess structures that could be called theory of mind, only "a decision tree with suggested responses that also accepted natural language inputs with discrete sections of natural language processing techniques embedded at specific points in the tree to determine routing to subsequent conversational nodes." [54, p. 3] However, Woebot does possess other advantages, which can help in its performance, that our system does not:

   - a coherent personality exhibited through its responses, which tends to make people more trusting and creating a bond with it [113];
   - a fully developed front-end and user interface, which tends to keep users' attention and focus longer [114];
   - the ability to deliver visual outputs (e.g., emojis, graphs, pictures, videos).

   Thus, this criterion refers to comparing outcomes from a system imbued with theory of mind versus a system devoid of this capacity, but possessing aforementioned advantages. This will demonstrate how the capability of theory of mind can enhance the performance of the system developed and described in this thesis.

Chapter 5

Materials and Methods

This Chapter describes the methods and materials used that were necessary to assemble the system described in the previous Chapters. This includes:

1. Data collection pre-study and main study research design and description;
2. Collected dataset with descriptive statistics and exclusionary criteria;
3. Computational experiments and the various methods used for conducting them - ML algorithms to build the models, feature selection methods, feature engineering methods (serving as cognitive modelling), feature importance methods (used for novel insights into mental health phenomena), and accuracy measures;
4. Empirical interventional study research design and description.

5.1 Data Collection

Research points toward the quality of data being the primary determinant of a successful ML model or an AI system [118], [119]. This means that the data used has a bigger impact on the model's or system's performance than selecting and building the optimal algorithm. Since the field of mental health lacks not only golden standard datasets, but any publicly available datasets for the construction of the system in this work, data collection was one of the priorities for this work and the subsequent dataset is one of its important contributions.

Before the main data collection was performed, a pre-study was conducted to test the research design for the data collection, from questionnaires posed to the applications used. The main idea was to collect a panel data (multiple individuals at multiple time intervals) of more than 1000 instances to be usable for ML. In terms of the time series characteristics, the data spans through approximately 4 weeks, and includes daily sampling of quantitative mental health metrics and qualitative, text-based diary entries on the daily experiences of the participants. The data was collected using Google Forms and the Synergetic Navigation System (SNS) application [120].

5.1.1 Ecological Momentary Assessment and the Synergetic Navigation System (SNS) Application

To collect quality data, the method of ecological momentary assessment (EMA) was employed. This way of collecting data follows the "anywhere, anytime" (A-A) principle ([30]), using a smartphone to collect data in non-invasive ways. EMA transports data collection from the lab into the wild. This has several benefits, as people cognize differently in the

wild, in their ecological environments, than in the lab [121], and it helps people overcome the recall bias, which "occurs when participants in a study are systematically more or less likely to recall and relate information on exposure depending on their outcome status, or to recall information regarding their outcome dependent on their exposure" [122, p. 126]. In that way, participants are in their ecological environment when the data is collected. The data collection can be signal-contingent, which "constitutes randomising notification timing throughout of a given timespan" [123, p. 135], interval-contingent, in which "notification timing is scheduled in line with a (predefined) time gap" [123, p. 135], e.g., a questionnaire every two hours, or event-contingent, where "the occurrence of a predefined event results in a notification" [123, p. 135], e.g., the smartphone senses specific movement through the global positioning systems (GPS) data [124]. EMA is being successfully used in mental health research [125].

This work utilized an interval-contingent data collection, using Google Forms and SNS [120] for data collection. SNS is a tool available for use on a smartphone or a computer which can collect quantitative questionnaire and qualitative text data by sending interval-contingent prompts to participants' smartphones which lead to the application's questionnaire interface, as well as emails which, by clicking the included link, lead to the web platform with the questionnaire.

5.1.2 Data Collection Pre-Study

The pre-study was conducted on 8 participants and lasted for about 3 weeks. The participants received instructions on how to use SNS and how the data collection was structured. The data was collected at the end of each day using the 10-item Positive and Negative Affect Schedule (PANAS) questionnaire [119], and a textual diary entry with specific guidelines (see *Supplementary Materials A.1*).

Furthermore, participants completed a demographic questionnaire (see *Supplementary Materials A.4*) and the BFI-10, a 10-item scale measuring the Big Five personality traits Extraversion, Agreeableness, Conscientiousness, Emotional Stability (or Neuroticism), and Openness [46]. The latter especially served for personalizing the linguistic output of the system (for more details, see Chapter 6).

PANAS, due to it simplicity, was used to understand how to keep adherence and convenience for the main data collection study. The plan was to see if the questionnaire was too long or too short, and to adjust the selection of a questionnaire appropriately.

Additionally, participant feedback regarding their experience with the data collection tools (such as the SNS tool), their perspectives on the clarity of questions, the comprehensibility of instructions and guidelines, as well as reports on any other concerns or suggestions, were also taken into consideration. Before the pre-study began, the participants signed informed consents.

The collected data was not shared with anyone, was held on a secure server and only used for research purposes. If at any point participants decided to withdraw their data, it was deleted from the server. The data was completely anonymized following the Olden et al's protocol for epidemiologic or clinical studies [126].

The takeaways from the pre-study were the following:

- Reformulate text to be clearer;
- PANAS was confusing for participants, it seemed better to include questions on SAD symptoms;
- add instructions for participants involved with a mental health professional;

- additional instructions on how to write machine readable text in the diary entries (important for later ML purposes);
- the time to complete a daily sample (PANAS, diary entry) was around 15 minutes;
- the questionnaire could be longer in the main study;
- 150 words for the diary was the appropriate amount to no cause participant attrition;
- SNS is appropriate for data collection.

5.1.3 Main Data Collection Study

The main data collection study, as the pre-study, used SNS to collect the data using EMA. The data was anonymized with the Olden et al's protocol for epidemiologic or clinical studies [126]. It differed from the pre-study in the following important aspects:

1. Instead of PANAS, the study collected data by combining items from several symptom inventories related to SAD, consisting of standardized screening questions used by mental health professionals in the process of mental health diagnosis. Depression Anxiety and Stress Scale 21 [48], Beck Anxiety Inventory [127], Beck Depression Inventory [128], and Ratcliffe's Depression Questionnaire [129] were used to compile the questionnaire. The final 18-item questionnaire, consisting of 18 questions relating to SAD symptoms, is available in the *Supplementary Materials A.5*.
2. 61 participants applied to the study as opposed to the 9 in the pre-study.
3. The study lasted 4 weeks as opposed to the pre-study's 3 weeks.

The instructions for participants are available in *Supplementary Materials A.2*. The revised guidelines for the diary entry are available in *Supplementary Materials A.3*.

From the 61 participants that applied, 7 participants never started the study. As in the pre-study, the participants completed a demographic questionnaire, BFI-10 [46], and a post-study questionnaire to ensure the quality of the collected data.

5.1.4 Data Quality

The post-study questionnaire, available in the *Supplementary Materials A.6*, functioning as a questionnaire to ensure data quality, showed the following:

- The median time for completing a daily sample was 20 minutes.
- The majority agreed that the instructions were clear (95% replied that everything was clear, 5% that a part was slightly unclear, and no participants found anything particularly or completely unclear). This ensures that the data collected represented what was targeted by the author.
- The majority found the 150 words for the daily diary sufficient to encompass the reported experience as well as keep the study convenient to not cause a drop in the quality of data collection.
- The majority rated the SNS tool in terms of its usability and comfortability with the highest ratings. This ensures that the data collected was not of lower quality due to the collection tools used or due to the lack of digital literacy by the participants.

- The majority focused on the questions at hand and did not think about what the researches might demand from them. This avoids the issue of demand characteristics in research [130].
- No participants replied that they would be unwilling to participate in a similar study again. This ensures that the data collected was from willing and engaged participants, which also contributed to the low attrition rate in the study.

This was an indicator that the collected data was of sufficient quality for subsequent use.

The study was reviewed and approved by an ethics committee (*Ethical approval code: cafiancimhumema_ 2021-07-13*). The participants signed informed consents before the beginning of the study. The full instructions, available to the participants, is available in the *Supplementary Materials A.2*.

5.2 Dataset

This section describes the collected dataset through descriptive statistics. Data from 50 participants passed the data quality filter, whose exclusionary criterion is missing data in any of the questionnaires.

One data instance (row) consists of the attributes seen in Table 5.1.

5.2.1 General Dataset Statistics

The dataset consists of 1495 data instances, of which 1168 instances are without missing data. Table 5.2 represents general statistics about the dataset and its number of instances.

5.2.2 Demography

This section presents descriptive statistics on the participants' data, collected with the demographic questionnaire (see Supplementary Materials A.4). The statistics are shown in Tables 5.3-5.12.

Among mental disorders, participants reported six depression disorders (e.g., major depressive disorder), four anxiety disorder (e.g., general anxiety disorder), two eating disorders (e.g., anorexia nervosa), and one obsessive compulsive disorder.

5.2.3 Big Five Personality Traits

Table 5.13 shows the participants' mean and standard deviation of their Big Five personality traits. The Big Five personality traits model is a widely recognized framework for understanding and categorizing human personality. See section 6.6.1.2 for an in-depth description of the model and its use in this work's system.

Table 5.1: Attributes in one data instance.

Data category	**Category attributes**	**Type of data**	**Number of attributes**
Metadata	Timestamp, Subject ID	/	2
Demography	Date of birth, Sex assigned at birth, Gender identity, Highest educational attainment, Current mental health status, Mental health history, Mental health therapy history, Mental health-related medication, Average hours of sleep, Average quality of sleep	Cross-sectional	10
Personality	Emotional valence, Emotional arousal, Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism	Cross-sectional	2 (emotional categories), 10 (personality traits - 2 each)
Daily mental health	See Supplementary Materials A.5 for attributes	Time series	18
Daily diary entry	/	Time series	1

Table 5.2: Basic statistics on the number of instances and diary word count per person.

(Per person)	M	SD
Instances	26.55	7.77
Diary word count	188.81	148.83

Table 5.3: Mean and standard deviation of continuous demographic attributes.

	M	SD
Age	29.8	5.78
Hours of sleep	07:50	00:49
Quality of sleep (1-5)	3.76	0.74

Table 5.4: Replies on the demographic question "Sex assigned at birth".

	Female	Male
N	32	18

Table 5.5: Replies on the demographic question "Gender identity".

	Woman	Man	Other
N	31	16	3

Table 5.6: Replies on the demographic question "Highest educational attainment".

	N
Up to high school	6
Bachelor's degree (or other form of similar higher education degree)	23
Master's degree	18
Doctoral degree	3

Table 5.7: Replies on the demographic question "Overall how would you rate your mental health?".

	N
Not sure	1
Poor	3
Somewhat poor	5
Average	14
Somewhat good	19
Excellent	8

Table 5.8: Replies on the demographic question "Have you ever been diagnosed with a mental disorder?".

	No	Yes
N	36	14

Table 5.9: Replies on the demographic question "Have you had mental health-related therapy in the recent past?".

	No	Yes
N	24	26

Table 5.10: Replies on the demographic question "Are you currently taking any medication for mental disorders?".

	No	Yes
N	42	8

Table 5.11: Replies on the demographic question "How would you self-describe your emotional valence?".

	N
I am generally a positive person.	22
I feel neutral about my emotional valence or I don't identify with either.	20
I am generally a negative person.	8

Table 5.12: Replies on the demographic question "How would you self-describe your emotional arousal?".

	N
My emotional arousal is usually high.	16
I feel neutral about my emotional arousal or I don't identify with either.	17
My emotional arousal is usually low.	17

Table 5.13: Means of sums of two questions for each Big Five personality trait (measured on a Likert scale 1-5, and reversed when appropriate).

B5 dimension	M	SD
Openness	7.39	1.87
Conscientiousness	7.33	1.48
Extraversion	5.41	1.74
Agreeableness	6.89	1.76
Neuroticism	6.26	2.09

5.3 Computational Experiments

This section describes the methods used for computational experiments performed. The methods were selected according to the overviewed SOTA in section 2. The results are in section 7. Models for three kinds of tasks were built. All tasks target the SAD levels and SAD symptoms, derived from the daily quantitative questionnaires (described in section 5.18). Tasks are the following:

1. Detection of SAD levels and SAD symptoms from only one text diary entry. Training occurs on the described dataset, while the detection occurs only from one user text input. This means that the detection is possible on only one textual input from a user.

2. Forecast of SAD levels and SAD symptoms from only one text diary entry. Training happens on the described dataset, while the forecast occurs only from one user text input (and not a time series data input, which are the common requirements for using forecasting models). The forecast happens for seven days in advance.

3. Forecast of SAD levels and SAD symptoms from quantitative questionnaire time series.

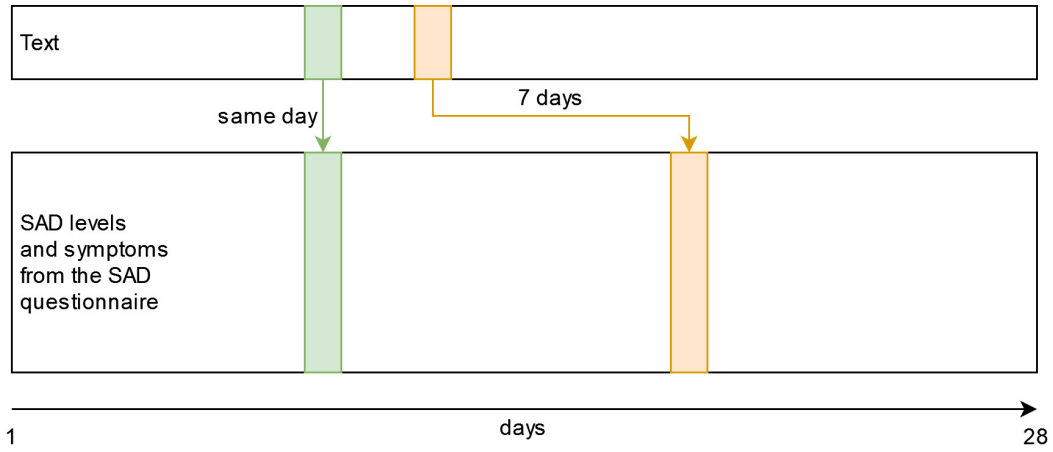


Figure 5.1: Visual representation of detection and forecast of SAD levels and symptoms. Green represents 'Detection of SAD levels and SAD symptoms from only one text diary entry', orange represents 'Forecast of SAD levels and SAD symptoms from only one text diary entry' for seven days in advance.

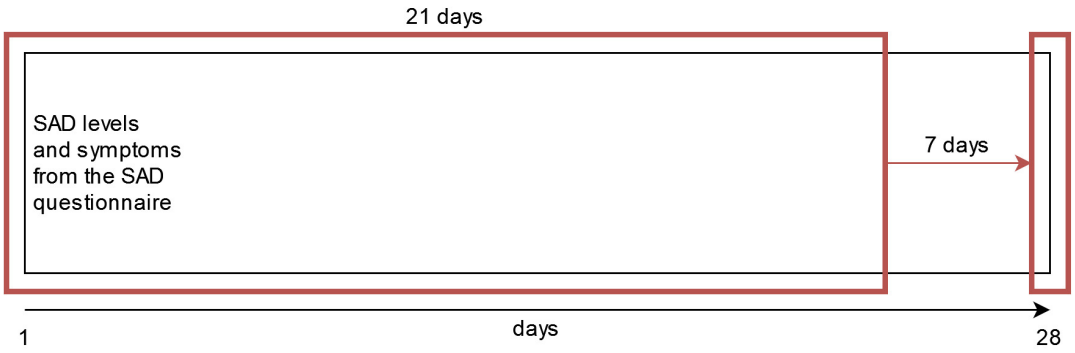


Figure 5.2: Visual representation of forecast of SAD levels and SAD symptoms from quantitative questionnaire time series. Red represents that 21 days of quantitative SAD levels and symptoms data are used to forecast one week in advance.

The next parts of this section include the following: the ML algorithms used for building the models as part of the system, processes used in feature selection, the process of feature engineering, and derived target variables.

5.3.1 Machine Learning Algorithms

Machine learning (ML) algorithms are computational methods or procedures designed to enable machines or computer systems to learn patterns, make predictions, or perform specific tasks without being explicitly programmed. These algorithms enable machines to automatically learn and improve from data or experience.

ML algorithms used for the tasks of SAD levels and symptoms detection and forecasting were the following:

- Decision Tree [131]
- Bagging Decision Tree [131]
- Boosting Decision Tree [131]
- Random Forest [131]
- Complement Naive Bayes [132]
- K-nearest Neighbors Classifier [133]
- Multiple Layer Perceptron Classifier [134]
- Logistic Regression [131]
- Support vector machine [135]

The algorithms' technical descriptions follow below.

5.3.1.1 Decision Tree

Friedman [131] describes a decision tree (DT) as follows:

> "A decision tree is an interpretable machine learning method for regression and classification. Trees iteratively split samples of the training data based on the value of a chosen predictor; the goal of each split is to create two sub-samples, or "children," with greater purity of the target variable than their "parent". For classification tasks, purity means the first child should have observations primarily of one class and the second should have observations primarily of another. For regression tasks, purity means the first child should have observations with high values of the target variable and the second should have observations with low values."

$$-(y \log(p) + (1 - y) \log(1 - p))$$

Entropy measure for reducing uncertainty in a dataset, used by DT for learning.
'Y' is the actual label of a data sample, which in binary classification is either 0 or 1.
'P' is the predicted probability of the data sample belonging to class 1.
The formula is used for binary classification.

5.3.1.2 Bagging Decision Tree

In a Bagging Decision Tree, multiple decision trees are trained on different subsets of the training data, randomly sampled with replacement (known as bootstrapping). Each tree is trained independently and makes its own predictions. During the training process, each decision tree learns from a different perspective of the data, leading to diverse models.

> "Bagging, short for bootstrap aggregating, combines the results of several learners trained on bootstrapped samples of the training data." [131]

5.3.1.3 Boosting Decision Tree

Boosting decision trees are an ensemble learning technique that combines decision trees with boosting methods to create predictive models. By iteratively constructing decision trees that focus on correcting errors made by previous trees, boosting decision trees achieve improved overall performance. The method assigns weights to training data instances, prioritizing misclassified samples during each iteration to refine subsequent trees [131].

5.3.1.4 Random Forest

> "A random forest is a slight extension to the bagging approach for decision trees that can further decrease overfitting and improve out-of-sample precision. Unlike bagging, random forests are exclusively designed for decision trees (hence the name).
>
> Like bagging, a random forest combines the predictions of several base learners, each trained on a bootstrapped sample of the original training set. Random forests, however, add one additional regulatory step: at each split within each tree, we only consider splitting a randomly-chosen subset of the predictors. In other words, we explicitly prohibit the trees from considering some of the predictors in each split." [131]

5.3.1.5 Complement Naive Bayes

Complement Naive Bayes (CNB) is a variant of the Naive Bayes algorithm, which is a probabilistic machine learning method used for classification tasks. Complement Naive Bayes is specifically designed to address the issue of imbalanced datasets, where the classes are not represented equally.

Algorithmic steps in CNB:

1. Calculate the probability of the given instance not belonging to each class.
2. Calculate for each class and select the lowest value.
3. The lowest value represents the lowest probability that that the instance does not belong to a certain class. The instance is finally sorted in that class.

$$\text{argmin } \mathrm{p}(y) \bullet \prod \frac{1}{\mathrm{p}(w|\hat{y})^{f_i}}$$

Formula for CNB. 'P' is the probability of a class, 'y' is the class label, 'w' is the weight, 'f' is the likelihood function, 'i' is the index of the feature.

5.3.1.6 K-Nearest Neighbors Classifier

The k-nearest neighbors (KNN) algorithm estimates the likelihood that an instance will become a member of a certain group based on the a specific distance (e.g., Euclidean distance, see below) between the instance and the groups. It is a lazy learning algorithm, because it does not perform any training when supplying the training data.

$$\|p\| = (p, \mathbf{0})$$

kNN is classified by using various distance measures.
The formula shows Euclidean distance measure. 'P' represent a vector.

5.3.1.7 Multiple Layer Perceptron Classifier

A multilayer perceptron (MLP) is a fully connected feedforward artificial neural network (ANN). MLP involves (at least) three layers of nodes: an input layer, a hidden layer and an output layer (see Figure 5.3). Each non-input node is a neuron that uses a nonlinear activation function. MLP utilizes backpropagation [134] for training. MLP can be trained to implement any given nonlinear input-output mapping.

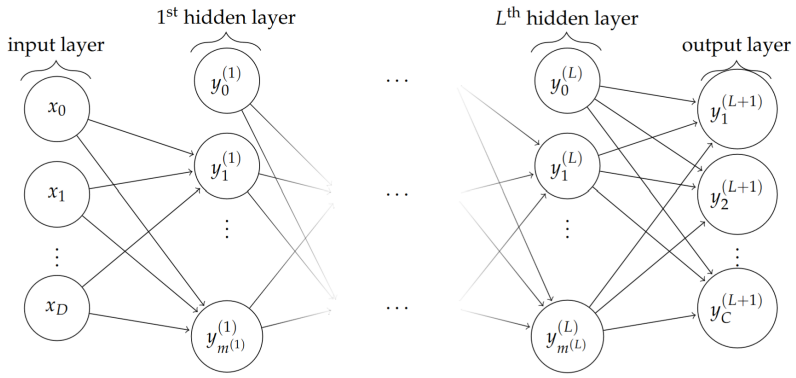


Figure 5.3: Network graph of a $(L+1)$-layer perceptron with D input units and C output units. The l^{th} hidden layer contains $m^{(l)}$ hidden units.

5.3.1.8 Logistic Regression

Logistic regression is a method commonly used for binary classification tasks. It is well-suited for scenarios where the outcome variable is binary, meaning there are only two possible classes. The algorithm models the relationship between the input features and the binary outcome variable by fitting a logistic or sigmoid curve to the data. This curve maps the predicted values to probabilities between 0 and 1, representing the likelihood of an instance belonging to a particular class. The decision boundary is set at 0.5 probability, and instances with predicted probabilities above the threshold are classified as one class, while those below are classified as the other class.

5.3.1.9 Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space that divides the data points. The objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Support vectors are data points close to the hyperplane that orient it. These vectors are used to maximize the margin. The loss function used for this maximization is called hinge loss.

$$\mathrm{l}(y) = \max(0,1 - t * y)$$

Hinge loss function. 'Y' is the prediction, 't' is the class label.

5.3.2 Feature Selection

For feature selection, two methods were used: Granger causality and colinearity.

5.3.2.1 Granger Causality

Granger causality is a statistical concept used to assess the causal relationship between variables in time series data. It extends beyond simple correlation by evaluating whether the past values of one variable can improve the prediction of another variable. In other words, if the inclusion of past values of Variable A significantly enhances the prediction accuracy of Variable B, it suggests that Variable A Granger-causes Variable B. This approach aims to capture the temporal precedence and predictive power of one variable over another, providing insights into potential causal links within a dynamic system.

$$y_t = c_2 + \sum_{i=1}^{3} \alpha_{2,i} y_{t-i} + \epsilon_{x,t}$$

The Granger causality test for restricted models.

$$y_t = c_2 + \sum_{i=1}^{3} \alpha_{2,i} y_{t-i} + \sum_{i=1}^{3} \beta_{2,i} x_{t-i} + \epsilon_{x,t}$$

The Granger causality test for unrestricted models.

In both, 'c_2' denotes a constant, 'i' an index, and 't' a specific point in time.

5.3.2.2 Collinearity

Collinearity is a feature selection technique that utilizes the state where two variables are highly correlated and contain similar information about the variance within a given dataset in order to deselect the feature from training a model. It uses correlation matrices to find the largest absolute values between values.

5.3.2.3 Chi-squared test

The Chi-squared test is a statistical hypothesis test that is used to determine whether there is a significant association between two categorical variables in a sample.

The Chi-squared test is used for feature selection in a classification problem by measuring the dependence between each input variable and the target variable. The test provides a score that indicates how likely the observed distribution of the categories is, given the expected distribution (i.e., the distribution that we would see if the variables were independent). Features that are most likely to be independent of the class variable, and hence irrelevant for classification, can be removed from the dataset.

5.3.3 Feature Engineering

Feature engineering is a ML technique that leverages data to create new attributes not presented in the raw collected dataset. The newly produced features can bring additional information about the data as well as speed up and simplify the data-related processes in the learning phase. The end results are generally higher accuracy or faster computations (due to less data complexity). When data is qualitative, textual data, feature engineering also involves the process of attribute numerization, turning non-numeric data into numeric features.

Below are presented feature engineering frameworks used in this work.

5.3.3.1 Valence Aware Dictionary and Sentiment Reasoner

Valence Aware Dictionary and sEntiment Reasoner (VADER) is a lexicon and rule-based sentiment technique, used for feature extraction [136]. The lexicon detects two sentiment dimensions: polarity and intensity. It is tuned to contents, produced in social contexts, and therefore applicable to sentiment analysis across several domains. It was built on over 9000 token features and 7500 lexical features. The approach was validated with a wisdom-of-the-crowd approach [137].

5.3.3.2 Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC) is a framework based on lexicon techniques [138]. It forms word categories based on the psychological meaning of the words. The categories are constructed based on psychological and cognitive science research, and new can be added by similarly basing their inclusion on the psychological and cognitive science knowledge base. It enables positioning instances into a multi-categorial space. Among others, LIWC lists the following categories (with relevant Tables 5.14-5.17 of what these categories encompass alongside dictionary examples): Standard linguistic dimensions, Psychological processes, Personal concerns, and Spoken categories. Other categories include analytical thinking, authentic speech, emotional tone, several grammatical categories, and drives (e.g., power).

Table 5.14: Standard linguistic dimensions.

Pronouns	I, them, itself
Articles	a, an, the
Past tense	walked, were, had
Present tense	Is, does, hear
Future tense	will, gonna
Prepositions	with, above
Negations	no, never, not
Numbers	one, thirty, million
Swear words	*****

Table 5.15: Psychological processes.

Social Processes	talk, us, friend
Friends	pal, buddy, coworker
Family	mom, brother, cousin
Humans	boy, woman, group
Affective Processes	happy, ugly, bitter
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy
Anxiety	nervous, afraid, tense
Anger	hate, kill, pissed
Sadness	grief, cry, sad
Cognitive Processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never
Inhibition	block, constrain
Inclusive	with, and, include
Exclusive	but, except, without
Perceptual Processes	see, touch, listen
Seeing	view, saw, look
Hearing	heard, listen, sound
Feeling	touch, hold, felt
Biological Processes	eat, blood, pain
Body	ache, heart, cough
Sexuality	horny, love, incest
Relativity	area, bend, exit, stop
Motion	walk, move, go
Space	Down, in, thin
Time	hour, day, oclock

Table 5.16: Personal concerns.

Work	work, class, boss
Achievement	try, goal, win

Leisure	house, TV, music
Home	house, kitchen, lawn
Money	audit, cash, owe
Religion	altar, church, mosque
Death	bury, coffin, kill

Table 5.17: Spoken categories.

Assent	agree, OK, yes
Nonfluencies	uh, rr*
Fillers	blah, you know, I mean

Other categories include analytical thinking, authentic speech, emotional tone, several grammatical categories, and drives (e.g., power).

LIWC can therefore be used for feature extraction for ML. With its use, text can be classified into a multi-dimensional space, covering the afore-mentioned categories representing mental dimensions. This provides accurate insights into a variety of mental processes: the topic covered, the thinking style, the emotional state, the cognitive processes, etc. This framework has been widely supported in psychology and cognitive science research [138].

5.3.4 Target Variables

For the system presented in this work, several target variables were engineered for ML models in the system to detect and forecast from the users' input text. The targets were engineered from the SAD symptoms questionnaire described in section 5.1.3 (also see *Supplementary Materials A.5*. Table 5.18 describes these target variables, composed of mental health issues (SAD) and SAD symptoms.

Table 5.18: Target variables used in the system's ML models to detect or forecast from the users' text input.

Target variable	**Description**
stress	Sum of Q1, Q2, Q17
anxiety	Sum of Q1-8, Q17
depression	Sum of Q1, Q9-18
inability to relax	Q1
nervousness	Q2
fear	Q3
tightness in chest	Q4
lightheadedness	Q5
feeling hot or cold	Q6
trembling	Q7
pounding heart	Q8
sadness	Q9
self-hatred	Sum of Q10, Q13
anhedonia	Sum of Q11-12
hopelessness	Q14
indecisiveness	Q15
fatigue	Q16
emotional detachment	Q18
suicidality	Q9-14, Q17-18

The target variables were binarized, with the positive class indicating the significant presence of the target variable in the users' text. A binarization threshold formula was used:

$$b = 0.25 * t$$

where ***b*** *presents the binarization threshold value and*
t *presents the maximum theoretical target variable value.*

1/4 of the maximum value of the questionnaire scale was selected as a threshold for categorization of significance due to its presence in numerous mental health diagnostic tools on SAD [48], [139].

5.4 Empirical Interventional Study

To test the efficiency and success rate of this work's system, an empirical interventional study [140] was designed. It compared a state-of-the-art chatbot for attitude and behavior change in mental health Woebot [54] with this work's system. Woebot was chosen as it is currently the most cited freely available system with the most replicated positive outcomes. It operates on a "decision tree with suggested responses that accepts natural language inputs" [54, p. 3]. It chooses a supportive strategy in the form of educational materials, appropriate messages, and pre-written advice based on users' emotions (e.g., expressed with emoticons) and their cognitive distortions. The participants for the present work's empirical interventional study were sampled from the general population, and a short screening questionnaire assessed their demographic and mental health status. The participants were placed in a laboratory setting, where they simulated a short, daily check-in with one of the chatbots. This included the participants providing the chatbot with a description of their day and possible issues that affected their mood. The participants then focused on the chatbots' subsequent responses. The participants used mobile devices or a computer for the check-in. Their experience was recorded with a mixed methods methodology, consisting of quantitative and qualitative questionnaires:

1. **Quantitatively evaluating SAD before and after the check-in**. The questions were based on the Single Item Screening Questions (SISQs) method [141], and were the following: "How stressed do you currently feel?", "How anxious do you currently feel?", "How depressed do you currently feel?". The answers were scored on a 5-point Likert scale from 1 ("Not at all") to 5 ("Extremely"). The questions were posed to the participants before and after using an ICA.

2. **Quantitatively evaluating the experience with chatbots with two expert measures from the User Experience Questionnaire (UEQ)** [142]. The two measures included the aspects of *obstructive-supportive* (how supportive the chatbot was) and *usual-leading edge* (how advanced in terms of technology and novel the chatbot seemed to be), evaluating on a 7-point Likert scale from 1 to 7.

3. **Qualitatively evaluating the experience with chatbots**. The question posed to the users after the chatbot use was "Was there anything in particular that you liked or disliked about the chatbot?".

The goal of the study was to focus the data analysis on the comparative aspects between the experiences and outcomes with different chatbots. The participants were therefore randomly sorted into two groups: the *Woebot* group (using Woebot) and the *test* group (using this work's system). The goal was to compare the effectiveness of the two chatbots through the pre- and post-study SISQs, as well as compare the users' experiences.

Participants working with mental health professionals had to consult with their chosen professional on their participation to ensure that no risks were involved. This was the study's exclusionary criteria. The data was fully anonymized with the researchers disposing of the data, not present in the final dataset (e.g., e-mail addresses), within one month after the research study. After the study, the participants had an option to remove parts of the data if they did not feel comfortable with it existing in this way after providing it. Consent forms on the research study were collected. The study was approved by an ethical committee (Ethical approval code: cbssotacfaabcfmhwasdcips_2022-06-29).

5.5 Software Used

Software used for building this work's system, computational experiments and analysis of empirical interventional study data encompasses Python 3.9 in the JetBrain PyCharm IDE with the focus on the following libraries:

- ChatterBot2: a machine-learning based conversational dialog engine
- huggingface-hub: a library for Hugging Face Hub, a platform that enables the sharing and discovery of pre-trained models and datasets for natural language processing, computer vision, and other AI-related tasks
- matplotlib: a library for creating static, animated, and interactive visualizations in a variety of formats
- nltk: a comprehensive library for natural language processing tasks, including tokenization, stemming, tagging, parsing, semantic reasoning, and corpus analysis
- numpy: a powerful library for scientific computing that provides support for large, multi-dimensional arrays and matrices, along with a wide range of mathematical functions to operate on them efficiently
- openai: the OpenAI library is a tool to access OpenAI's natural language processing models, including GPT-3, and use them to generate text, answer questions, and complete prompts
- pandas: a library for data manipulation and analysis, enabling high-performance data structures such as dataframes and series, with tools for data cleaning, transformation, and visualization
- pickle: a library that allows for the serialization and deserialization of Python objects, enabling the conversion of complex data structures, such as lists, dictionaries, and class instances, into a format that can be easily stored and retrieved from disk or transmitted over a network
- plotly: a library for creating interactive, high-quality graphs, charts, and dashboards
- scikit-learn: a library for machine learning tasks, providing a range of algorithms for classification, regression, clustering, and dimensionality reduction, as well as tools for model selection, evaluation, and data preprocessing

- scipy: a library for scientific and technical computing, providing a range of modules for optimization, integration, interpolation, signal and image processing, linear algebra, statistics, and more, built on top of NumPy arrays
- seaborn: a data visualization library based on Matplotlib, providing a high-level interface for creating statistical graphics, including heatmaps, time series, categorical plots, and regression models
- tokenizers: a library for tokenizing text, providing support for a wide range of languages and tokenization strategies, including byte-level, word-level, and subword-level tokenization, as well as normalization and padding functions
- torch: a machine learning library that provides support for deep learning tasks, including building and training neural networks, implementing various optimization algorithms, and manipulating tensors efficiently
- transformers: a library built on top of PyTorch and TensorFlow for natural language processing tasks, providing access to pre-trained models such as BERT, GPT-2, and T5, and offering a range of functionalities such as tokenization, sequence classification, and question-answering

Chapter 6

Cognitive Architecture Design

6.1 General Overview of the Cognitive Architecture

This section overviews the system's cognitive architecture. The design can be seen in Figure 6.1. The in-depth description of the architecture can be found in section 6.4.

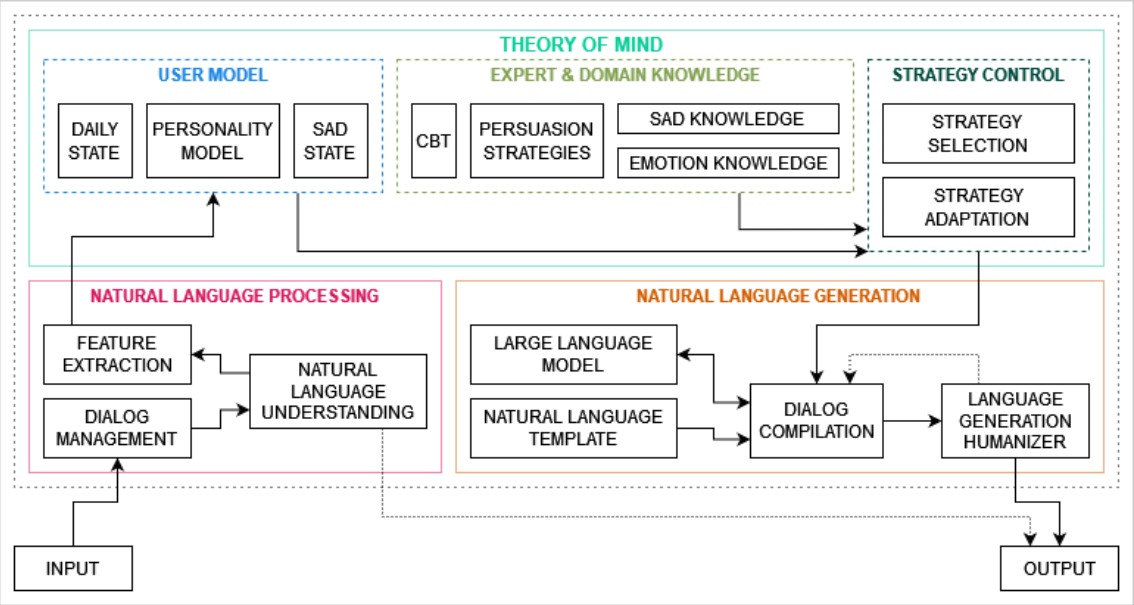


Figure 6.1: The system's cognitive architecture. It consists of: the *Natural language processing module* (rose color), the *Natural language generation module* (orange color), and the *Theory of mind module* (teal color), which is further divided into the *User model* (blue color), the *Expert & domain knowledge module* (light green color), and the *Strategy control module* (dark green).

The Natural Language Processing (NLP) module in CogA's pipeline, comprising Dialog Management, Natural Language Understanding, and Feature Extraction submodules, is activated upon each user input. Utilizing a Recursive Frame-Based Probabilistic Framework (RFBP), Dialog Management tracks conversation progress, guiding Natural Language Understanding. This submodule assesses input sensibility via Rule-Based Filtering (RBF), determining if the input aligns with a predefined conversational branch language ontology. Valid inputs proceed to Feature Extraction, where features are derived through numeric at-

tribute creation, employing the LIWC framework and VADER sentiment model for feature engineering.

The Theory of Mind (ToM) module comprises three modules: User Model, Expert & Domain Knowledge, and Strategy Control. ToM processes numeric features from the Feature Extraction module and combines information from the User Model's submodules (Daily State, Personality Model, and SAD State) with Expert and Domain Knowledge's ontological submodules (CBT, Persuasion Strategies, SAD Knowledge, and Emotion Knowledge) to determine strategies within Strategy Control's submodules (Strategy Selection and Strategy Adaptation). All ontologies in the ToM module are Rule-Based ontologies, meaning the relationships between the propertie are defined with if-then rules.

The User Model simulates short-term and long-term user cognitive models, informed by the Daily State submodule's current mental state tracking, Personality Model's Big Five personality trait representation, and SAD State submodule's detection and prediction of SAD levels and symptoms from the input text. The Personality Model, based on the Big Five personality traits, customizes messages by triggering appropriate persuasion strategies from the Expert & Domain Knowledge module.

The SAD State submodule utilizes ML models to track users' mental health, detect current SAD states from text, and forecast SAD states up to 7 days in advance. This information guides the system in providing user support and determining the necessity of strategy dispatch.

The Expert & Domain Knowledge module includes the CBT submodule, Persuasion Strategies, SAD Knowledge, and Emotion Knowledge submodules. The CBT submodule implements personalized CBT techniques based on users' mental state and experience levels. Persuasion Strategies submodule employs a Domain Mapping Matrix (DMM) relating Big Five personality dimensions to CPP, enabling tailored strategies for users with specific dominant traits. The SAD Knowledge submodule utilizes DMM to map mental health issue topics and SAD levels and symptoms to appropriate CBT techniques. The Emotion Knowledge submodule manages the tone of the system's outputs.

The Strategy Control module, comprising Strategy Selection and Strategy Adaptation submodules, selects and adapts mental health strategies based on information from the User Model and Expert & Domain Knowledge modules. By employing Ratio Formulas, the module evaluates strategy effectiveness for specific users, learning from past encounters.

The Strategy Selection submodule extracts users' SAD levels, symptoms, and personality information from the User Model, and selects a CBT technique using the CBT and SAD Knowledge submodules from the Expert & Domain Knowledge module. It applies a probability model to choose the most effective strategy for long-term users.

The Strategy Adaptation submodule adapts selected mental health strategies using NLP, wrapping them in persuasion strategies and appropriate communication tones. It selects persuasion strategies based on users' Big Five personality traits and determines communication tone using information from the SAD State submodule. The submodule also re-adapts strategies if they prove ineffective in a current conversation.

The Natural Language Generation (NLG) module processes strategies from the Strategy Control module, enriches them with text from a large language model, and ensures the output is not harmful to the user. The Dialog Compilation submodule combines the strategy with enriched text using Natural Language Templates, and the Language Generation Humanizer verifies that the output is not harmful, requesting new text from Dialog Compilation (which in turn requests it from LLM) if necessary.

The system currently supports GPT-3, GPT-Neo, GPT-J, and AI21 Jurassic-1 language models for generating unique and enriched text. GPT-J is the default selection, while ChatGPT (GPT-3.5, GPT-4) is not included due to the lack of an API.

The Dialog Compilation submodule relies on the Natural Language Template submodule to merge the selected strategy with enriched text. The Language Generation Humanizer submodule filters out potentially harmful text outputs, using Rule-Based Filtering with the Bad Bad Words and Toxic Comments datasets, as well as Threshold-Filtering with VADER for detecting negative sentiment scores.

6.2 Algorithmic Description of the Cognitive Architecture

To understand the working of the system algorithmically, see Algorithm 6.1. It represents various natural language processing techniques and models to create a contextually relevant and persuasive output based on the user's input and the system's understanding of the user's mental state and personality. This aims to generate human-like responses while maintaining a safe and responsible communication standard.

The algorithm can be broken down into the following steps:

1. Ask the user about their day, then append their text input to the Dialog Management submodule (see 6.5.1), which tracks the conversation with RFBP.
2. If the input is not filtered with RBF, proceed to the next step; otherwise, return a text output to the user asking for a valid text input. This happens, e.g., if a user replies with a single digit to the system's question "How was your day?".
3. Extract features of the input text using the B5 (see section 6.6.1.2 for the explanation on the Big Five Personality model), LIWC, and VADER feature extraction techniques. See section 5.3.3 for details.
4. Update the Personality Model (see section 6.6.1.2) submodule with B5 features and the Daily State (see section 6.6.1.1) with LIWC and VADER features.
5. Update the SAD state submodule (see section 6.6.1.3) by detecting and forecasting SAD levels and symptoms based on the extracted features. See section 5.3.4 for predicted and forecasted target variables in the SAD state, section 5.3.1 for ML algorithms used, and section 7.1 for developed ML models and their evaluation.
6. Determine the appropriate CBT technique (see section 6.6.2.1) based on the current SAD state.
7. Apply DMM to the B5 features to obtain a persuasion framework (see section 6.6.2.3).
8. Combine the persuasion framework, CBT technique, and Emotion Knowledge Model (applied to LIWC features) to create a combined framework.
9. Select a natural language template based on the combined framework.
10. Repeat the following steps until the generated output is not considered “risky” by the Language Generation Humanizer (see section 6.8.3):
    (a) Generate an output text based on the selected template.
    (b) Add sentences to the output using a Language Learning Model (LLM).
    (c) Apply the Language Generation Humanizer (see section 6.8.3) to the output and evaluate whether it is considered “risky” or not.
11. Return the generated output text to the user.

Algorithm 6.1: Algorithmic description of CogA.

Data: User's text input i
Result: System's text output o

initialization;
1: $dialog\_management\_submodule$.append(APPLY_RFBP($i$))
2:
if $APPLY\_RBF(i)$ **then**
  pass
else
  return $invalid\_input\_prompt$
end
3:
$features_{b5} \leftarrow$ EXTRACT_FEATURES($t, B5$)
$features_{LIWC} \leftarrow$ EXTRACT_FEATURES($t, LIWC$)
$features_{VADER} \leftarrow$ EXTRACT_FEATURES($t, VADER$)
4:
UPDATE_PERSONALITY_MODEL($features_{b5}$)
UPDATE_DAILY_STATE($features_{LIWC}, features_{VADER}$)
5:
UPDATE_SAD_MODEL(
  DETECT_SAD_STATE($features_{b5}, features_{LIWC}, features_{VADER}$)
)
UPDATE_SAD_MODEL(
  FORECAST_SAD_STATE($features_{b5}, features_{LIWC}, features_{VADER}$)
)
6: $cbt\_technique \leftarrow$ DETERMINE_CBT_TECHNIQUE(sad_state)
7: $persuasion\_strategy \leftarrow$ APPLY_DMM($features_{b5}$)
8: $combined\_strategy \leftarrow$ COMBINE_STRATEGIES(
  $persuasion\_strategy$,
  $cbt\_technique$,
  APPLY_EMOTION_KNOWLEDGE($features_{LIWC}$)
)
9: $text\_template \leftarrow$
 SELECT_NATURAL_LANGUAGE_TEMPLATE($combined\_strategy$)
10:
while $humanized == risky$ **do**
  $o \leftarrow$ GENERATE_OUTPUT($text\_template$)
  $o \leftarrow$ ADD_SENTENCES_USING_LLM(o)
  $humanized \leftarrow$ APPLY_LANGUAGE_GENERATION_HUMANIZER(o)
end
11: **return** o

6.3 Operational Pipeline of the Cognitive Architecture

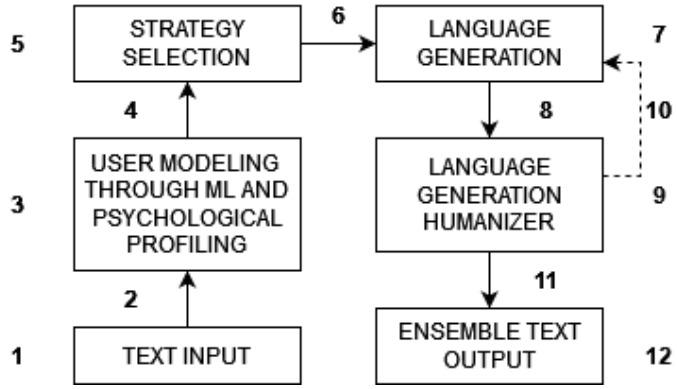


Figure 6.2: The system's pipeline through the modules in one conversational round.

This section presents the proximate pipeline of the system executing one operational loop - one conversational round (see Figure 6.2). Afterwards, it provides an example of one such round. The steps in the conversation loop are the following:

1. Users provide a textual input on their day, similar to a diary entry, describing their mood, their experiences while in that mood, mental health issues, issues in their thinking and actions, problems in their lives, and similar.

2. The text is automatically processed to extract the features through a cognitive modelling framework.

3. Pre-trained models use features to detect and forecast users' SAD levels and symptoms, charting users' mental health trends. Psychological modelling is used to model users' personalities, which are used to determine their mental and cognitive profiles as well as current mental states. User's daily state is modelled in multiple dimensions, which work as representations of the user's mental properties.

4. The user cognitive modelling profiling metrics are sent to the next part of the module.

5. A strategy is selected and adapted according to the metrics, determined in the previous part. The text serves to mitigate the user's mental health problems based mostly on CBT, and, if the forecasted trend is negative, to try to break that trend. To ensure that the user follows the selected strategy, the text on CBT is wrapped in a persuasion strategy. The persuasion strategy is personalized, working from an ontology (Perusasion Strategies submodule) and based on the user's psychological model.

6. The text is sent to be augmented with the next part of the module.

7. The text from the previous part is enriched by a generative pre-trained transformer based on a large language model (e.g., GPT-3) with additional text. This makes responses more varied and alive for the user.

8. The enriched text is passed to the Language generation humanizer.

9. Language generation humanizer decides whether the added text generated in the previous module is acceptable in terms of risk for the user. It rejects the text if it is detected as risky.

10. Language generation humanizer returns the original text for another enrichment if deemed too risky for the user.

11. The final text is compiled through a natural language template.

12. Output of the final text.

A real-life example of the system in action:

1. The user inserts text "Today I have felt very bad. I feel a lot of stress because I have a deadline at work coming, and I fought with my partner yesterday. I am stressed about talking to them tonight. The deadline is a bit scary because I did not do a good job last time, and if it happens again, it might trigger my depression. Furthermore, I felt tired and fatigued, which made me have no motivation to do anything. I did not feel like talking to people, either. I was a little anxious at times but calmed down relatively fast. I also felt a little sad at times for no apparent reason. Due to all of this, I also felt detached from myself and my surroundings, more often than usually, but it was not super intense. In the evening, I watched Netflix to distract myself a bit."

2. Features are extracted by using the feature extraction techniques from 5.3.3. Dialogical questioning extracts the user's B5 psychological profile.

3. The features from the second step are used for: 1) creating a daily cognitive profile of the user, 2) determine a SAD state detecting the user's SAD levels and symptoms and their forecast for up to 7 days in advance. For detection, in this example, the ML models detect significant levels of stress and depression, as well as symptoms of inability to relax, nervousness, lightheadedness, tremor, sadness, self-hatred, anhedonia, disinterest, sense of failure, hopelessness, indecisiveness, fatigue, and closed world. Furthermore, the user's B5 profile is assembled.

4. The information on the user is sent to the Strategy selection module.

5. Strategy selection takes into account the scores on the mental health topic from the daily profile, and SAD levels and symptoms. Since the user is experiencing a lot of symptoms and they are using the system for the first time, a lower difficulty CBT technique is selected. Depression and its symptoms were prominent, so Pleasant Activity Scheduling technique is selected. Since the dominant B5 dimension of the user is agreeableness, CPP of authority is selected to adapt and wrap the text to be more persuasive.

6. The module takes the selected strategy and sends it to language generation.

7. The text is assembled in order to reflect the selected mental health and persuasion strategy. The first part of the text is the following:

> You seem to be experiencing stress and depression. You may have some symptoms of inability to relax, nervousness, lightheadedness, tremor, sadness, self-hatred, anhedonia, disinterest, sense of failure, hopelessness, indecisiveness, fatigue, closed world.

This is then enriched with the following two sentences with a large language model (this was an actual output of GPT-J):

> You feel easily irritated. It's hard to respond to others.

The persuasion strategy wraps the text in the following way:

> The scientific, expert research on the problems that you are experiencing is clear on what helps.

At the end, the conversation continues with the deployment of a CBT technique with its own conversational tree, in this case using the Pleasant Activity Scheduling technique.

8. The text is passed to the Language generation humanizer module.

9. Language generation humanizer processes the text. It does not find any risky indicators.

10. The text is not returned to the Language generation module to replace the enriched text as no risk is detected.

11. The final text is managed as a dialog and the part of the dialog up to the CBT conversational tree is passed as an output.

12. The user receives the following text output:

> You seem to be experiencing stress and depression. You may have some symptoms of inability to relax, nervousness, lightheadedness, tremor, sadness, self-hatred, anhedonia, disinterest, sense of failure, hopelessness, indecisiveness, fatigue, closed world. You feel easily irritated. It's hard to respond to others.
>
> The scientific, expert research on the problems that you are experiencing is clear on what helps.

The system afterwards initiates the Pleasant Activity Scheduling technique conversational tree (see Figure 6.3 and Table 6.2).

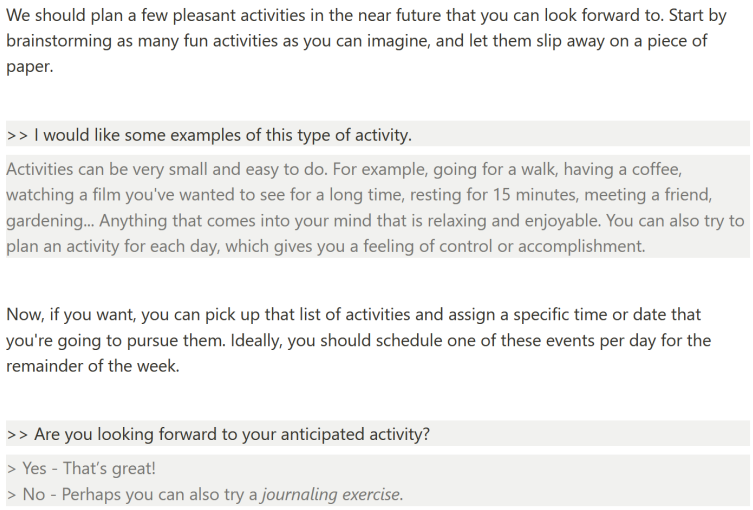


Figure 6.3: A part of the conversational tree for the Pleasant Activity Scheduling technique.

6.4 Detailed Description of the Cognitive Architecture

The subsequent sections provide a detailed description of CogA's modules, explaining how each module contributes to the system's functioning. This includes illustrative examples and an exploration of the computational methods that underpin the performance of each CogA module (See section 7.1 for computational methods). The modules described follow as such: the ***Natural language processing module*** (rose color), the ***Theory of mind module*** (teal color) – which is further divided into the ***User model*** (blue color), the ***Expert & domain knowledge module*** (light green color), and the ***Strategy control module*** (dark green) – and the ***Natural language generation module*** (orange color).

6.5 Natural Language Processing

The Natural language processing module (rose color in Figure 6.1) is the first module in the CogA's pipeline that is started at the beginning of each conversation, and is activated after each users' input. It takes care of tracking the conversation, prompting users if the replies are not sensible, and extracting meaning from the input text to be used by the subsequent modules of the CogA. It consists of the Dialog management, the Natural language understanding, and the Feature extraction submodules.

6.5.1 Dialog Management

The Dialog Management submodule keeps track of where in the conversation the ICA and the user are. It therefore signals to the Natural language understanding submodule how to understand the input. The Dialog Management submodule is built as a Recursive Frame Based Probabilistic Framework (RFBP) [143], as the dialog tree consists of data-modelling as well as probabilistic sequencing depending on the ML detection of user states.

6.5.2 Natural Language Understanding

The Natural language understanding submodule does the following:

1. Receives the linguistic user input.
2. Determines whether the input is sensible according to what the conversation is at that point in time about. It does this by using the Rule-Based Filtering method [144] to identify invalid linguistic inputs from a pre-defined specific conversational branch rule-based language ontology.
3. If the input is sensible, it sends the input to the Feature Extraction submodule, otherwise it prompts the user to reply again with possible additional information on how they should reply.

An example of the working of the submodule: if a user is prompted by the system to describe their day, and the user inputs '5', the rule-based filter will signal to the submodule to prompt the user again, explaining how they have to reply to avoid invalid linguistic inputs.

6.5.3 Feature Extraction

The Feature extraction module receives the text input and extracts the features from it, meaning that it creates numeric attributes according to specific rules and algorithms. This serves for cognitive modelling of the user, capturing mental properties of their current mental states. The main methods used include the LIWC framework [138] and the VADER sentiment model [136]. See section 5.3.3 for the in-depth description of the feature engineering techniques used to create the attributes.

6.6 Theory of Mind

In cognitive science, the Theory of mind (ToM) describes the ability to "understand the thoughts and feelings" [109, p. 528] as well as "attributing thoughts and goals to others" [Ibid.] in order to function in social life. This system's ToM is more domain-specific, but it serves the same purpose – to understand its user to the degree where it can offer effective personalized help for relieving SAD symptoms. This is its goal in its social interactions. To simulate ToM, this work made an interdisciplinary effort to integrate findings from AI, cognitive science, and behavioral sciences.

ToM (teal color in Figure 6.1 includes the following three modules: User model (blue color), Expert and domain knowledge (light green color), and Strategy Control (dark green color).

When ToM receives numeric features from the Feature extraction module, it sends them to the User model and its three submodules - Daily state, Personality model, and SAD state. These hold information on the user and their current state of mind in the form of data models. Data from these submodules is then combined with the user-relevant knowledge from Expert and domain knowledge's ontological submodules CBT (cognitive behavioral therapy), Persuasion strategies, SAD knowledge, and Emotion knowledge submodules to programmatically sculpt and computationally determine strategies in Strategy control's submodules Strategy selection and Strategy adaptation.

The following three sections describe the three submodules in ToM.

6.6.1 User Model

The User Model module (blue color in Figure 6.1) models the user. It converts input through feature extraction, dialogical questioning, and ML modelling into meaningful information that can be used for determining the system's outputs in a conversation. It

therefore builds a cognitive model of a user in short-term and long-term situations (only from one conversation or across time), which can be used to simulate different outcomes of the support the system offers to the user through different strategies.

User model contains three submodules, each maintaining a particular aspect of the user:

1. the Daily state submodule keeps track of how the user's mental state is currently;

2. the Personality model holds the information on the more long-term, stable psychological characteristics of the user (using the Big Five personality model, described below in-depth);

3. the SAD state contains ML models that detect SAD levels and symptoms of the user, as well as forecast them for up to 7 days in advance.

All the submodules help inform the strategy selection and support output of the model. The in-depth description of the submodules follows.

6.6.1.1 Daily State

The Daily State submodule takes the attributes from the Feature extraction submodule and maps them to a multi-dimensional model of a user. The data attributes that form the user model are explained in section 5.3.3, and include: standard linguistic dimensions, psychological processes, personal concerns, spoken categories, analytical thinking, authentic speech, emotional tone, several grammatical categories, and drives (e.g., power).

An example representation of some dimensions that make up the model of the user can be seen in Figure 6.4. The Daily state submodule, among others, informs the Strategy Control (dark green color) module on emotions and what the focus topic of the daily mental health issues is (e.g., are the mental issues connected more to the body or to thinking).

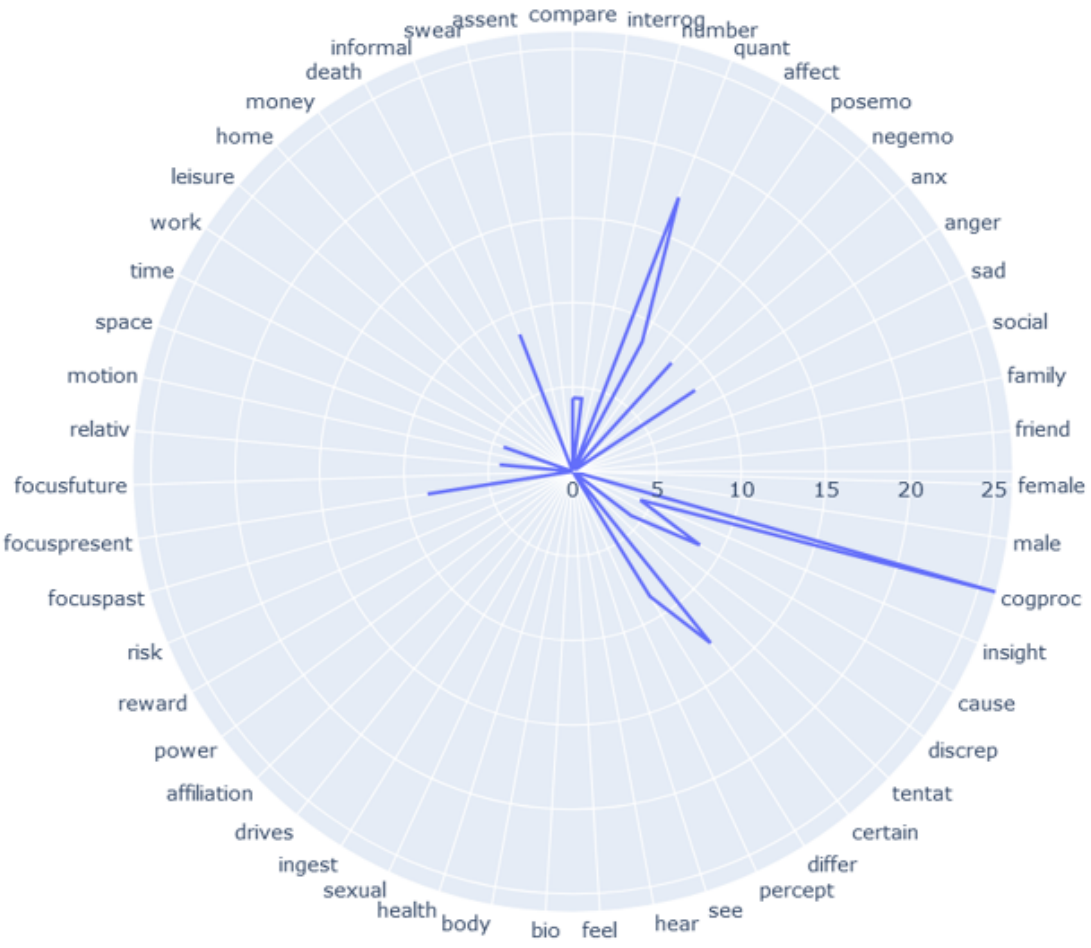


Figure 6.4: Example of a spider chart daily cognitive model of a user. It contains several dimensions, based on the feature extraction.

6.6.1.2 Personality Model

The system builds the Personality model of the user by measuring several dimensions of the user, which try to describe an individual's tendencies that relate to their psychological and cognitive functionalities, such as the mental states, decision-making and what influences them. This multi-dimensional framework is based on the Big Five personality traits model (B5). The dimensions are measured on the Likert scale with values ranging from 1 to 10. The model holds the following psychological dimensions:

- **Openness** measures a person's inclination towards curiosity, imagination, and openness to new ideas and experiences. Individuals high in openness tend to be creative, adventurous, and open-minded.
- **Conscientiousness** measures the tendency to be organized, responsible, disciplined, and goal-oriented. People high in conscientiousness are often diligent, dependable, and strive for achievement.
- **Extraversion** measures the degree of sociability, assertiveness, and outgoingness in social interactions. Individuals high in extraversion are typically energetic, talkative, and seek social stimulation.

- **Agreeableness** measures a person's tendency to be compassionate, cooperative, empathetic, and considerate towards others. Those high in agreeableness are often friendly, warm, and value harmonious relationships.
- **Neuroticism** measures the degree of emotional instability and proneness to experiencing negative emotions. Individuals high in neuroticism may be more susceptible to anxiety, depression, mood swings, and stress.

The system collects the numerical data necessary for computational representation of the B5 modelling through a Finite State [143] conversational tree branch. See Figure 6.5 for an example of a B5 psychological user model.

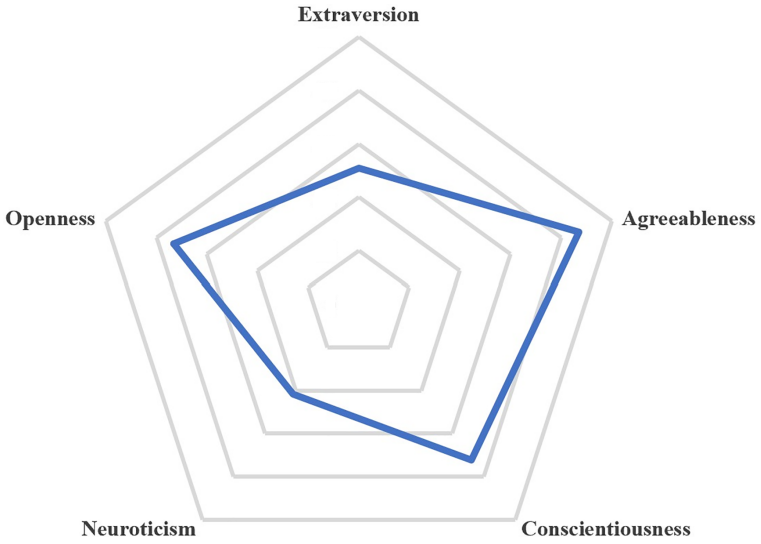


Figure 6.5: Example of a spider chart B5 psychological user model.

The Strategy control module heavily relies on B5, as the latter is one of the most stable psychological and cognitive constructs, highly reproduced and successful in determining the right kind of influence on specific personalities [116]. It is therefore ostensibly used to personalize the messages the system dispatches, e.g., according to the dominant B5 dimension of the user [116], by triggering the appropriate strategy from the Persuasion Strategies submodule in Expert & Domain Knowledge module (light green color).

6.6.1.3 SAD State

The SAD state submodule contains several ML models, trained to detect and forecast SAD levels and symptoms, described in Table 5.18, including the following: levels of stress, anxiety, depression; and symptoms of inability to relax, nervousness, fear, tightness in chest, lightheadedness, feeling hot or cold, trembling, pounding heart, sadness, self-hatred, anhedonia, hopelessness, indecisiveness, fatigue, emotional detachment, and suicidality. The submodule keeps track of the users' mental health. It informs how the system should act in its support of the user, and whether strategy dispatch is necessary.

The ML models can detect the current SAD state from the text as well as forecast it for up to 7 days in advance. For this, the SAD state submodule only needs one text entry from the user to be able to do that. Furthermore, it can forecast the users' SAD state up to 7 days in advance from the quantitative questionnaires, described in section 5.1.3. The data can be collected by the system through dialogic questioning. For that, it needs a

longer time series sample, which happens if the conversation with a specific user does not occur just once.

The performance of the ML models can be seen in section 7.1.

6.6.2 Expert and Domain Knowledge

6.6.2.1 CBT

The CBT submodule contains the knowledge about the Cognitive Behavioral Therapy (CBT). CBT is a form of psychological treatment, which "focuses on challenging and changing cognitive distortions (such as thoughts, beliefs, and attitudes) and their associated behaviors to improve emotional regulation and develop personal coping strategies that target solving current problems" [145, para. 1]. There are many techniques that are used in CBT, each with its own difficulty level (D1-D5). Below are 20 techniques that are implemented as a strategy with their own conversational trees in this work's system.

The ontology relates the user's mental state and experience levels to the difficulty levels as well as the optimal strategy. Both the strategy or technique selection and the difficulty level are personalized according to the user.

- **Diaphragmatic Breathing** (D1) or "belly breathing," involves fully engaging the stomach when breathing. This type of breathing helps your lungs fill up more efficiently and generally helps people relax.
- **Relaxed Breathing** (D1) encompasses a deliberate modulation of respiration to elicit a physiological and psychological relaxation response by engaging in deep diaphragmatic breathing.
- **Mindful Meditation** (D1) consists of concentrating on the present. It is done by increasing your awareness of your consciousness, breathing and body. If you notice a thought or emotion, simply observe it and let it pass without judgement.
- **Useful Contacts** (D1) is not a technique, but a list of contacts for immediate support from mental health practitioners or trainees.
- **Pleasant Activity Scheduling** (D2): planing pleasant activities in the near future that the user can look forward to.
- **Progressive Muscle Relaxation** (D2) gradually relaxes different muscle groups.
- **Grounding technique** (D2) evolves the senses to become aware of the surrounding. The system includes two grounding techniques:
  - **5-4-3-2-1 Technique** makes the user deliberately go into the details of the environment using each of the senses by describing five things they can see, four things they can touch, three things they can hear, two things they can smell, and one thing they can taste.
  - **Categories Technique** makes the user select three categories of items (e.g., movies, animals, cities) and list as many items of that category as possible.
- **'Breaking It Down' (Task Nervousness) Technique** (D2) instructs the user to think about the task that causes them stress and break it down into as many sub-tasks as possible.
- **Color Visualisation** (D2) helps relieve stress and overall mood improvement through attributing colors to feelings, and then visualizing them.

- **Brain Dump** (D3) refers to the act of thoroughly jotting down one's thoughts free of any internal judgment or bias.
- **Play the Script Until the End** (D4) is an exercise that helps the user minimize their fears and frame them in a more realistic light by thinking how a situation might continue and resolve in the future.
- **Fact-checking** (D4) helps users identify the difference between facts and opinions, making them less prone to feel negatively by opinions.
- **Journaling** (D4) helps the users observe and release the thought patterns they operate with daily, thus making them understand their internal working and be able to change them with time. Three different journaling techniques are implemented:
  - **Gratitude journaling** makes the user write down what they are grateful for.
  - **Mood journaling** makes the user track their moods.
  - **Freestyle journaling** is free form journaling where the user decides what is important to them.
- **Indecisiveness** (D4) teaches the user the different ways they can be informed about making a decision, helping them find a decision-making technique that does not make them feel negatively.
- **Indecisiveness: Exercise** (D4) makes the user think about small decisions they have made recently, making a list and seeing that they make more decisions than they know, thus building their decision-making confidence.
- **Cognitive Restructuring** (D5) involves first identifying a situation that leads to stress and the thoughts and feelings that arise in that situation. Then it helps the user examine their thoughts to determine what is true about them and what is not true about them. Finally, it helps the user develop alternative and more balanced thinking and determine how they will feel (as a result) when they adopt this new thinking.
- **ABC** (D5) helps the user understand the meaning of their responses to adversity or a particular event.
- **Dysfunctional Thought Record (Socratic Questioning)** (D5) makes the user analyze thoughts and events that made them feel negatively through Socratic questioning.

6.6.2.2 Persuasion Strategies

The submodule Persuasion Strategies contains the ontological knowledge on how to influence people with different psychological and personality characteristics. This makes, e.g., CBT techniques more effective as they are wrapped in a context where they are presented to a user in a way that makes them more susceptible to following the technique. This is one of the more important points for therapy due to the otherwise high drop off rates.

The submodule's ontology comprises of a Domain Mapping Matrix (DMM) [146] between B5 dimensions and Cialdini's principles of persuasion (CPP) [36]. This means that people with a specific dominant B5 are more susceptible to a specific CPP. CPP's main idea is that there is no general persuasive strategy that works for all people, hence orthogonal strategies should be identified and applied to those that are most susceptible to individual strategies. CPP posits seven strategic bases for influencing people:

1. **authority**, which targets people that are more inclined to be motivated by a legitimate authority;

2. **commitment**, which targets people that tend to commit to their previous behavior;

3. **social proof or consensus**, which targets people that tend to do what others do;

4. **liking**, which targets people that are more likely to be motivated by someone they like;

5. **reciprocity**, which targets people that tend to return a favor;

6. **scarcity**, which targets people that consider scarce things more valuable;

7. **unity**, which targets people that are influenced by appealing to their group identity.

Different people are influenced by different strategies, and interactive technology can be utilized to choose specific strategies that work for specific people. To give an example, people with high *agreeableness* on B5 are more prone to be influenced by the principle of *authority* [147]. To translate that in a simple example, instead of prompting a user with a message

Try exercising

it is much more effective, if the user's *agreeableness* is high, to prompt them with

The scientific, expert research on the problems that you are experiencing is clear on what helps. Try exercising.

Invoking the authority of experts is a part of *authority*, and thus the probability of the user exercising would be higher.

The submodule therefore works by exctracting a user's B5 dimension with the highest or lowest value from the User Model and relating it to its CPP strategy counterpart in the ontology's DMM (see Table 6.1, to help with such strategy personalization as seen in the above example (based on [148])). Technically, this is realized by at first randomly selecting one of the shortlisted CPP strategies under the corresponding B5 dimension based on Rule-Based Selection (if-then rules), but the system can with time probabilistically learn the best strategy for a user.

Table 6.1: Mapping between B5 dimensions and which Cialdini's principles of persuasion influence such individuals.

B5 dimension	**Cialdini's principle**
Openness	Authority, Consensus, Liking (low in B5 dim.)
Conscientiousness	Commitment, Reciprocity (high in B5 dim.)
Extraversion	Consensus, Unity (high in B5 dim.)
Agreeableness	Authority, Commitment, Liking (high in B5 dim.)
Neuroticism	Consensus, Scarcity (high in B5 dim.)

6.6.2.3 SAD Knowledge

The SAD knowledge submodule maps, using DMM, the topic of the mental health issue (e.g., body and thinking), information about which is found in the User Module's Daily state submodule, and the levels and symptoms of SAD, information about which is found in the User Module's Daily state submodule SAD state, to different CBT techniques. The representation of the DMM can be seen below in Table 6.2.

Technically, this is realized in the system with Rule-Based Selection. The SAD level or symptom is detected with a ML model in the system, and the topic is extracted from the text with the LIWC feature engineering technique (see section 5.3.3). Considering these two factors, the system at first selects a random CBT technique to utilize from a short list of possible techniques that align with the two factors, but does with time probabilistically learn the best technique for a user.

Table 6.2: Mapping between SAD levels and symptoms, CBT techniques, and mental health topics.

SAD level/symptom	**CBT technique**	**Mental health topic**
Tightness in chest	Breathing techniques	Body
Lightheadedness	Progressive Muscle Relaxation, Relaxed Breathing, Mindful Meditation	Body
Fatigue	Progressive Muscle Relaxation, Relaxed Breathing, Mindful Meditation	Body
Feeling hot or cold	Journaling, Relaxed Breathing	Body
Trembling	Grounding technique, Relaxed Breathing	Body
Heart pounding	Progressive Muscle Relaxation, Relaxed Breathing, Mindful Meditation	Body
Inability to relax	Progressive Muscle Relaxation, Relaxed Breathing, Mindful Meditation	Body & cognition
Nervousness	Play the Script Until the End, 'Breaking It Down'	Body & cognition
Fear	Play the Script Until the End, Relaxed Breathing	Body & cognition
Sadness	Cognitive restructuring or reframing, Journaling, Color Visualization, ABC	Cognition & mood
Self-hatred	Dysfunctional Thought Record, Fact-checking, ABC, Cognitive restructuring or reframing	Cognition & mood
Emotional detachment	Dysfunctional Thought Record	Cognition & mood

Suicidality	Pleasant Activity Scheduling, Useful Contacts, Cognitive restructuring or reframing, Visualization	Cognition & mood
Anhedonia	Pleasant Activity Scheduling	Cognition & mood
Hopelessness	Cognitive restructuring or reframing, Pleasant Activity Scheduling	Cognition & mood
Indecisiveness	Cognitive restructuring or reframing	Cognition & mood
Stress	Progressive Muscle Relaxation, Relaxed Breathing, Mindful Meditation, Journaling, Cognitive restructuring or reframing	General
Anxiety	Progressive Muscle Relaxation, Relaxed Breathing, Mindful Meditation, Cognitive restructuring or reframing, Play the Script Until the End, ABC	General
Depression	Cognitive restructuring or reframing, Pleasant Activity Scheduling	General

6.6.2.4 Emotion Knowledge

The Emotion knowledge submodule takes care of the tone of the system's outputs. It currently relies on two tone techniques:

- The system chooses to use shorter sentences if the user is depressed. This is due to depression causing impaired cognitive processing [149].
- The system uses different punctuations and emoticons depending on the user's mood. People in different mental states perceive sentence signs and symbols differently [150].

Technically, both are achieved by all the deterministic, non-LLM generated texts having various semantically equivalent options, and Rule-Based Selection is applied on the sets of sentences to achieve the appropriate emotional tone.

6.7 Strategy Control

The Strategy control module (dark green in Figure 6.1) takes information from the User model module and the Expert & Domain Knowledge module, and selects or adapts a strategy according to that information. It contains two submodules: the Strategy selection submodule and the Strategy adaptation submodule.

The Strategy control module also keeps track of how effective different strategies are for a specific user (CBT techniques and persuasion strategies), if that same user uses the system continuously. It uses Ratio Formulas [151] to formally evaluate the success of a specific strategy for a specific user, therefore computationally learning from past encounters with the user on which strategy to use.

6.7.1 Strategy Selection

The Strategy selection submodule selects an appropriate mental health strategy:

1. To select a mental health strategy, it extracts information about users' SAD levels and symptoms from the SAD state submodule in the User model, information about the personality from the Personality model submodule, and about the mental health topic from the Daily state submodule in the User model.

2. It extracts a CBT technique according to the information about the difficulty, mental health topic, and SAD levels and symptoms, extracted in the previous step. The selected CBT technique is extracted by using the submodules CBT and SAD knowledge from the Expert & Domain Knowledge module.

Furthermore, the submodule relies on a probability model using Ratio Formulas [151] to select the strategy with the highest probability of being effective, which also relies on the previous effectiveness of an already utilized strategy related to a specific long-term user.

6.7.2 Strategy Adaptation

The Strategy Adaptation submodule adapts the mental health strategy, selected by the Strategy selection submodule, by wrapping and adapting it to a persuasion strategy and an appropriate communication tone, using natural language processing:

1. **Persuasion strategy selection**: To select a persuasion strategy that wraps the mental health strategy, it extracts information about users' B5 personality from the Personality model submodule in the User model module. Afterwards, it uses that information to extract the appropriate CPP strategy from the Persuasion strategies submodule in the Expert & Domain Knowledge module.

2. **Communication tone**: To select the correct communication tone (the sentence length and use of sentence symbols), it extracts information from the SAD state submodule in the User model. Afterwards, it uses that information to extract the appropriate information from the Emotion knowledge submodule in the Expert & Domain Knowledge module.

The submodule also takes care of re-adapting a strategy if it is not working in a current conversation with a user.

6.8 Natural Language Generation

The Natural language generation module (orange color in Figure 6.1) processes the strategy it receives from the previous module, and enriches it with a stochastically determined text from a large language model (Large language model submodule). The Dialog Compilation submodule used the Natural Language Templates [152] to combine everything together, and sends the compiled text to the Language generation humanizer submodule to verify that the potential output is not harmful to the user (see the problems with harmful text generations in Chapter 1). If it detects harm, it returns it to the Dialog compiler to get another text enrichment from the Large language model submodule, otherwise the text is output to the user.

6.8.1 Large Language Model

The models below are the current natural language generation models (large language models, LLMs) present in the system. Due to its modular design, new models can be easily integrated into it. The selected model receives the text on the selected strategy and generates continuing text to enrich it. This makes each output completely unique. Numerous parameters allow control over how the original text is enriched. Through internal testing, adding two sentences seemed to generate the best outcomes.

- **GPT-3**: Generative Pre-trained Transformer 3 (GPT-3) is an autoregressive language model that uses deep learning to produce human-like text. GPT-3's full version has a capacity of 175 billion machine learning parameters.
- **GPT-NEO**: GPT-Neo is an implementation of model and data-parallel autoregressive language models, utilizing Mesh Tensorflow for distributed computation on TPUs. GPT-Neo was used to train a family of models between 125 million and 2.7 billion parameters on the TensorFlow Research Cloud.
- **GPT-J**: GPT-J 6B is a transformer model trained using Ben Wang's Mesh Transformer JAX. "GPT-J" refers to the class of model, while "6B" represents the number of trainable parameters. The model consists of 28 layers with a model dimension of 4096, and a feedforward dimension of 16384. The model dimension is split into 16 heads, each with a dimension of 256. Rotary Position Embedding (RoPE) is applied to 64 dimensions of each head. The model is trained with a tokenization vocabulary of 50257, using the same set of BPEs as GPT-2/GPT-3.
- **AI21 Jurassic-1**: Jurassic-1 is a pair of auto-regressive language models recently released by AI21 Labs, consisting of J1-Jumbo, a 178B-parameter model, and J1-Large, a 7B-parameter model. It is based on the decoder module of the Transformer architecture (Vaswani et al., 2017) with the modifications proposed by Radford et al. (2019).

GPT-J is the default selection in the system. ChatGPT, mentioned in section 8.1, was not included due to a lack of an API.

6.8.2 Natural Language Template & Dialog Compilation

The Natural language template submodule contains Natural Language Templates [152], which are used for crafting conversational outputs. The Dialog compilation submodule, on the other hand, takes the selected strategy from the Strategy control module and the enriched text from the Large language model, and it uses the hardcoded Natural Language Templates to sequence and merge various text snippets into a coherent whole.

6.8.3 Language Generation Humanizer

The submodule filters text outputs potentially harmful for the user, and requests new text enrichments if the current text is deemed harmful. The submodule is based on the Rule-Based Filtering method [144], made with the Bad Bad Words dataset [153] and the Toxic Comments dataset [154], as well as on the Threshold-Filtering method, which works through ML detected negative sentiment scores using VADER (see section 5.3.3). The Rule-Based Filtering method filters out texts containing keywords from the datasets, while the Threshold-Filtering method filters out texts whose sentiment scores below a pre-set threshold.

Algorithm 6.2 shows the algorithmic description of the Natural Language Generation module.

Algorithm 6.2: Algorithmic description of the Natural Language Generation module.

Data: Strategy *strategy*
Result: Compiled output text *compiled_text*

initialization;
1: *enriched_text* ← LargeLanguageModel(*strategy*)
2: *compiled_text* ← DialogCompilation(*enriched_text*)
3: *harmful* ← LanguageGenerationHumanizer(*compiled_text*)
4:
while *harmful* **do**
  enriched_text ← LargeLanguageModel(*strategy*)
  compiled_text ← DialogCompilation(*enriched_text*)
  harmful ← LanguageGenerationHumanizer(*compiled_text*)
end
5: **return** *compiled_output_text*

Chapter 7

Results

The Results Chapter presents results from various experiments, focusing on computational experiments (see section 7.1 for methodology used in them) of ML models for SAD state, and on the empirical interventional study (see section 5.4 for the research design) which compared this work's system with Woebot.

7.1 Computational Experiments

In computational experiments with the system's ML models for SAD state detection and forecasting, accuracy was used as the evaluation measure of the models (see section 5.3.4 for descriptions on target variables). ML algorithms based on related work in section 2 were selected to be used in the fundamental experiments with SAD levels, and the best performing explainable ML algorithms were used subsequently. This decision was motivated by the European Union's movement towards regulating the use of AI in healthcare, emphasizing the importance of explainability [155]. Furthermore, the adoption of explainable algorithms offers healthcare practitioners new and valuable insights. section 5.3.1 provides methodological descriptions of the ML methods used.

7.1.1 Feature Selection

Table 7.1 shows the features selected for each SAD level target variable using the feature selection techniques for classification described in section 5.3.2. The heuristic of selecting 20 top features was used [156]. This worked better than following the rule for the number of features to be proportional to $\sqrt{N}$ [157] (N denoting the number of instances in the dataset), which would be 34 features in this work's case. The feature selection was always performed in the training set groups. For the description of the features, refer to sections 5.2 and 5.3.3.

Table 7.1: Selected features for the three target variables - stress, anxiety, and depression.

Stress	**Anxiety**	**Depression**
'neuroticism'	'neuroticism'	'neg'
'neg'	'neg'	'neuroticism'
'Tone'	'negemo'	'compound'
'negemo'	'WC'	'Tone'
'compound'	'openness'	'openness'
'WC'	'Tone'	'sad'

'anx'	'compound'	'negemo'
'posemo'	'posemo'	'health'
'insight'	'risk'	'function'
'Clout'	'anx'	'posemo'
'Sixltr'	'Comma'	'WC'
'pos'	'health'	'Clout'
'openness'	'power'	'ipron'
'extraversion'	'Analytic'	'Period'
'Period'	'AllPunc'	'work'
'negate'	'focuspast'	'leisure'
'AllPunc'	'neu'	'Comma'
'auxverb'	'function'	'negate'
'quant'	'work'	'i'
'relativ'	'Sixltr'	'WPS'

7.1.2 Detection of SAD Levels and Symptoms from Single Text Entries

The whole dataset (see section 5.2 for more information on the dataset) was used to build the models. The 10-fold cross validation [158] with subject-wise splitting was used to evaluate the accuracy of the models. Majority class was used for the baseline model.

7.1.2.1 SAD Levels

Table 7.2: SAD levels detection from a single text entry after the system's question on the user's daily mood, experiences, and events.

SAD	**Baseline**	**RF**	**CNB**	**kNN**	**MLP**	**LOG**
Stress	53.30	72.37	**74.56**	70.84	73.40	72.70
Anxiety	73.70	78.93	75.77	**80.12**	78.39	79.82
Depression	66.60	78.07	73.15	79.20	79.07	**83.33**

KNN is selected in order to have a transparent model that can be utilized in mental healthcare. Explainable AI makes the system open to scrutiny and provides novel insights in the field.

7.1.2.2 SAD Symptoms

Table 7.3: SAD symptoms detection using kNN for explainability from a single text entry after the system's question on the user's daily mood, experiences, and events.

SAD symptom	**Baseline**	**kNN**
inability to relax	46.32	74.05
nervousness	42.81	73.51
fear	69.78	73.62
tightness in chest	67.72	74.86
lightheadedness	70.63	80.16

feeling hot or cold	88.01	91.41
trembling	74.91	75.90
pounding heart	77.65	82.26
sadness	57.53	75.91
self-hatred	55.05	75.23
anhedonia	67.21	74.78
hopelessness	62.16	72.75
indecisiveness	65.84	80.00
fatigue	72.86	81.81
emotional detachment	50.94	72.58
suicidality	62.67	76.20

7.1.3 7-Day Forecasting of SAD Levels and Symptoms from Single Text Entries

In this section, we present computational performance results obtained from models utilizing various ML algorithms. These models were designed to process individual text entries as inputs and forecast the levels of SAD and SAD symptoms that users experience seven days into the future. The 10-fold cross validation with subject-wise splitting was used to evaluate the accuracy of the models. Majority class was used for the baseline model.

Table 7.4: SAD levels 7-day forecast from a single text entry after the system's question on the user's daily mood, experiences, and events.

SAD	**Baseline**	**RF**	**CNB**	**kNN**	**MLP**	**LOG**
Stress	53.30	68.45	**75.88**	65.21	65.83	64.31
Anxiety	73.70	75.76	75.47	75.47	**77.77**	73.99
Depression	66.60	75.38	**77.65**	77.03	72.66	67.20

Table 7.5: SAD symptoms 7-day forecast using kNN for explainability from a single text entry after the system's question on the user's daily mood, experiences, and events.

SAD symptom	**Baseline**	**kNN**
inability to relax	46.32	73.31
nervousness	42.81	72.03
fear	69.78	73.47
tightness in chest	67.72	73.31
lightheadedness	70.63	78.49
feeling hot or cold	88.01	*87.68*
trembling	74.91	79.77
pounding heart	77.65	77.70
sadness	57.53	74.60
self-hatred	55.05	75.08
anhedonia	67.21	78.94
hopelessness	62.16	81.67
indecisiveness	65.84	77.01

fatigue	72.86	74.44
emotional detachment	50.94	76.85
suicidality	62.67	71.54

7.1.4 Quantitative Questionnaire Scores SAD Level Time Series 7-Day Forecasting

The ML models for forecasting SAD level time series from quantitative questionnaires are relevant if the user uses the system daily for several consecutive days, utilizing the quantitative SAD questionnaire assessment and not only the natural language assessment (chat).

The models were built using a time series of 28 days, where the first 21 days were used for training and the last 7 days for forecasting. User-wise time series 10-fold cross validation was used to evaluate the accuracy of the models. All features were used to build the models (due to the smaller number of them - when Granger causality was used for feature selection, models performed worse). Instead of majority class, a stronger baseline model was used - Naive Forecast, where the predictions for a given period are equal to the observed value for the prior period.

Table 7.6: 7-day forecast of SAD level quantitative questionnaire scores.

SAD	**Baseline**	**kNN**	**LOG**	**CNB**	**SVM**	**DT**	**BagDT**	**BoostDT**	**RF**
Stress	73.52	79.41	**80.39**	76.47	**80.39**	*68.62*	74.51	75.49	75.49
Anxiety	88.23	92.16	**93.14**	89.22	92.16	88.23	92.16	90.19	92.16
Depression	87.25	88.12	**89.11**	*87.13*	**89.11**	*79.21*	89.11	*84.16*	86.14

7.2 Empirical Interventional Study

This section presents the results of the empirical interventional study, described in section 5.4. The study collected data from 42 participants and randomly sorted them into two groups. Group "Woebot" used Woebot (see section 2) for a quick daily therapeutic check-in. Group "Our system" used the system described in this work (see Chapter 6) for a quick daily therapeutic check-in. Stress, anxiety and depression was measured before and after the check-in.

7.2.1 Stress

Figure 7.1 shows how the stress of each group changed before and after using an ICA. Independent samples t-test was used to determine that the stress score pre-ICA usage between the two groups was not significantly different, making them comparable ($p = 0.642$). Paired t-testing determined that the stress score in the "Woebot" group did not change statistically significantly after the use ($p = 0.484$), while the stress score in the "Our system" group did change statistically significantly after the use ($p = 0.048$).

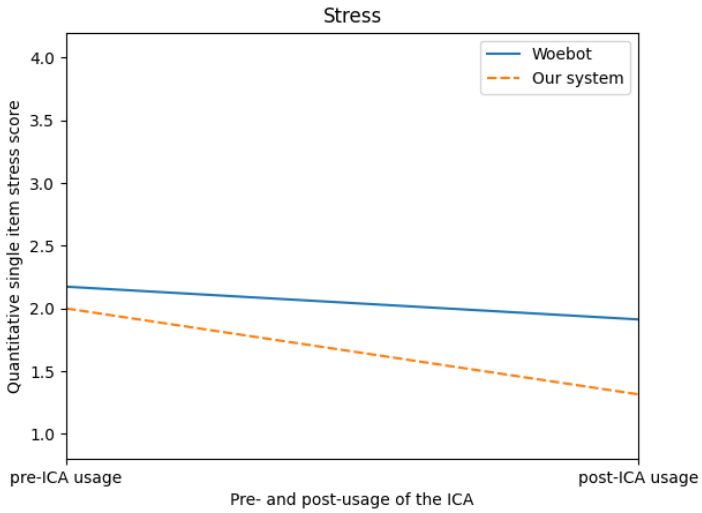


Figure 7.1: Plot comparing participant stress scores pre- and post-ICA usage in two different groups, "Woebot" and "Our system". Only participant stress in the "Our system" group changed statistically significantly.

7.2.2 Anxiety

Figure 7.2 shows how the anxiety of each group changed before and after using an ICA. Independent samples t-test was used to determine that the anxiety score pre-ICA usage between the two groups was not significantly different, making them comparable ($p = 0.822$). Paired t-testing determined that the anxiety score in the "Woebot" group did not change statistically significantly after the use ($p = 0.509$), while the anxiety score in the "Our system" group did change statistically significantly after the use ($p = 0.040$).

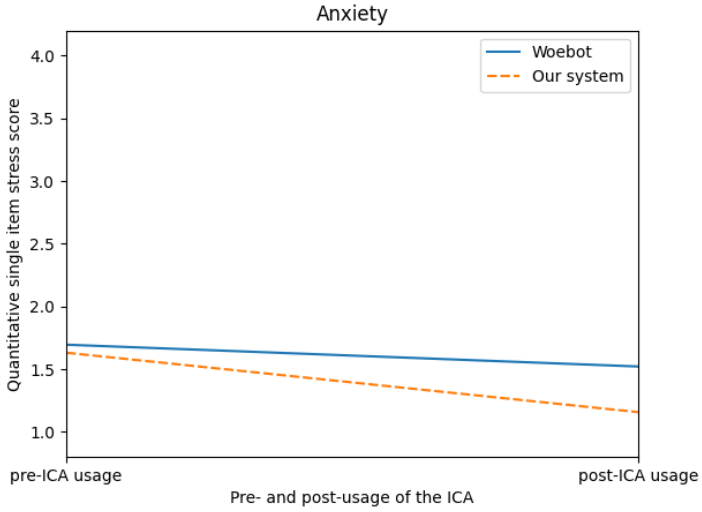


Figure 7.2: Plot comparing participant anxiety scores pre- and post-ICA usage in two different groups, "Woebot" and "Our system". Only participant anxiety in the "Our system" group changed statistically significantly.

7.2.3 Depression

Figure 7.2 shows how the depression of each group changed before and after using an ICA. Independent samples t-test was used to determine that the depression score pre-ICA usage between the two groups was not significantly different, making them comparable ($p = 0.317$). Paired t-testing determined that the depression score in neither the "Woebot" group ($p = 0.789$) nor in the "Our system" group ($p = 0.688$) changed statistically significantly after the use.

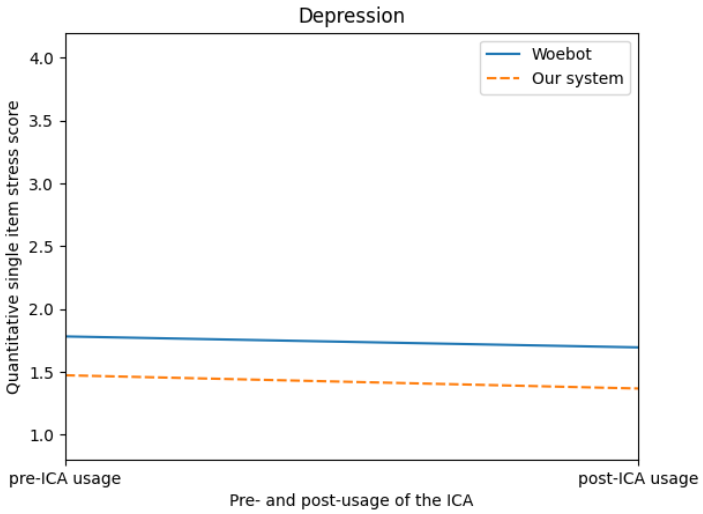


Figure 7.3: Plot comparing participant depression scores pre- and post-ICA usage in two different groups, "Woebot" and "Our system". Participant depression scores changed statistically significantly in neither of the groups.

7.2.4 User Experience Questionnaire Measures

Apart from measuring stress, anxiety, and depression, the participants rated their experience using one of the ICAs on the 7-point Likert scale using two measures from UEQ [142]: *obstructive-supportive* (how supportive the chatbot was) and *usual-leading edge* (how advanced in terms of technology and novel the chatbot seemed to be). T-testing was used to determine that for the measure *obstructive-supportive*, the group "Our system" ($M = 5.368$) found this work's system statistically significantly more supportive than the "Woebot" group ($M = 4.261$) found Woebot supportive ($p = 0.041$), while for the measure *usual-leading edge*, the groups' scores were not statistically significantly different (both $M = 4.609$, $p = 0.084$).

Chapter 8

Discussion

This section compares this work's system to the SOTA systems, described in Section 2. The evaluation is mostly represented through comparison tables. The discussion largely relies on presenting results to evaluate the hypothesis criteria in Section 4.2.

8.1 Comparison with SOTA Systems in User Assessment

Out of 9 extensively reviewed systems, 4 reported classification accuracies. These, together with this system's work, were included in table 8.1. The table shows that this work's system is able to detect the most assessed categories with the best highest accuracy. This work's system can also forecast the assessed categories, which none of the other SOTA system do. As a point of interest, ChatGPT has recently been evaluated on its classification capabilities for stress, depression and suicidality, and its reported accuracy was 73%, 85%, and 33%, respectively [159]. This work performed with a higher accuracy in all of these except in depresion, as can be seen in Tables 7.2 (stress: 74.56%, depression: 83.33%) and 7.3 (suicidality: 76.20%). However, it has to be noted that none of these systems are generally directly comparable – they operate with different datasets, which, through different algorithms and methods, spawned different models. It is important to note that the studies by Delahunty et al. [50], Ghandeharioun et al. [99], and Pola et al. [102], which utilized datasets with multiple instances from the same individuals, did not report using subject-wise splitting (or an equivalent method) for model validation. Performing randomized (or similar) splitting can potentially introduce bias and overfitting, leading to inflated classification accuracies.

Table 8.1: Comparison between this work's assessed categories, accuracies, and best-performing ML methods.

System	**Assessed categories**	**Accuracies reported**	**Best methods**
[50]	depression suicidal ideation insomnia hypersomnia weight change inappropriate guilt	detection: 0.91 (F1)	Random Forest Logistic Reg.
[51]	sentiment emotions	detection: 81%	Fuzzy Matching

[99]	valence arousal	detection: 65.7-82.4%	AdaBoost Reg.
[102]	seven emotions	detection: 84%	LSTM
This work	levels of SAD inability to relax nervousness fear tightness in chest lightheadedness feeling hot or cold trembling pounding heart sadness self-hatred anhedonia hopelessness indecisiveness fatigue emotional detachment suicidality	detection: 72.58-91.41% forecast: 71.54-87.68%	kNN
ChatGPT [159]	stress depression suicidality	detection: 33-85%	Transformer

8.2 Comparison with Woebot in an Empirical Interventional Study

This work's system's effectiveness was evaluated through an empirical interventional study (see Section 7.2). It was compared to Woebot [54], SOTA system for mental health which the available research reports to have the highest interventional successes. Woebot's interventional successes were also the most reproduced compared to other SOTA systems. As reported in Section 7.2, this work's system performed better than Woebot in reducing stress and anxiety in the participants. Neither system succeeded in statistically significantly reducing depression in participants. This work's system was better evaluated on being supportive, while both systems scored the same on the *usual-leading edge* metric.

Furthermore, after the short check-in, participants replied to the question "Was there anything in particular that you liked or disliked about the chatbot?". Participants mostly liked the accuracy and depth of the system's assessment (presumably due to the amount of categories the system is able to detect). They noted the lack of the user interface for this work's system, which could be equaled to a dislike, while they really liked the user interface for Woebot. They also noted and liked the friendliness of Woebot, pointing to a simulated personality of the system that may positively affect the outcome of interventions.

8.3 Hypothesis Testing

The hypothesis proposed that an "intelligent cognitive assistant for attitude and behavior change for stress, anxiety, and depression will achieve results at least comparable to the

state-of-the-art if it simulates theory of mind in a novel artificial cognitive architecture."

It may be concluded that **the hypothesis H1 was confirmed** based on the Comparison with SOTA Systems in User Assessment (Section 8.1) and Comparison with SOTA System in an Empirical Interventional Study (Section 8.2).

Chapter 9

Conclusion and Future Work

This work presents a comprehensive computational psychotherapy system for mental health prediction and behavior change with a conversational agent. Following are two highlighted contributions.

The first contribution is a novel, golden standard dataset, which includes panel data (multiple individuals at multiple time intervals; multiple time series) of quantitative SAD symptom scores from diagnostic-level questionnaires and qualitative daily diary entries. The psychotherapeutic and psychiatric communities that use computational approaches in their work were lacking such a dataset, which should prove useful in further research.

The second contribution is the system for SAD symptom relief, mostly based on a simulated Theory of mind, a cognitive ability to understand others. ToM is built as an ensemble of various models and novel ontologies. This includes psychological and cognitive user modelling, mental health and behavior change ontologies, detection and forecasting machine learning models, large language models wrapped in risk detection models, and behavior change prompt generators.

The work's hypothesis was that an "intelligent cognitive assistant for attitude and behavior change for stress, anxiety, and depression will achieve results at least comparable to the state-of-the-art if it simulates theory of mind in a novel artificial cognitive architecture." To test the hypothesis, the system was evaluated in two ways: (1) by comparing the detection and forecasting capabilities of various mental health phenomena from text to other SOTA systems in computational experiments, and (2) by performing an empirical interventional study comparing this work's system to a another SOTA system, Woebot, which does not possess ToM. Hypothesis H was confirmed as this work's system outperformed other SOTA systems in selected categories - it was able to detect more mental health phenomena, it was able to forecast the mental health trends (other systems were unable to do that), the detection accuracies were generally higher, and the system was more successful in offering effective mental health support to participants versus Woebot. However, Woebot has an advanced interface and other user-friendly features, while this work's system currently resembles a research prototype.

In assessing the findings of this work, it is crucial to recognize several notable limitations. Firstly, the dataset utilized for training the models may exhibit biases toward specific populations, potentially arising from sample selection biases, such as geographical, cultural, or age-related factors. This can cause the models to overfit on specific populations and perform inaccurately if used for populations that were left out of the sample. Secondly, the system's linguistic outputs are contingent upon the large language models employed, which implies that rapid advancements can be achieved through the integration of improved models. This work's existing coding framework has been designed to facilitate such seamless incorporation. Thirdly, the target attributes are subject to change as defi-

nitions of mental health phenomena evolve over time, which may consequently impact the models' detection and forecasting performance. Fourthly, the study design featured an empirical interventional approach, and should be classified as a quasi-experiment rather than a clinical trial or randomized controlled trial, despite the random assignment of groups. As such, the results should be interpreted as indicative of trends rather than definitive outcomes, as definitive cause-and-effect conclusions are difficult due to the absence of a control group, smaller sample size and limited control over multiple variables. Finally, the short-term nature of the empirical interventional study limits the generalizability of the findings, and it is possible that medium- or long-term investigations may reveal different outcomes for participants. Future work should consider addressing these limitations to enhance the robustness and applicability of the findings.

The plans for future work include further advancement of the system as well as further experiments as implied by the limitations. The first encompasses a design of a user interface, which was shown to be an important aspect of what participants like; refinement of ML models by using more advanced algorithms (e.g., LSTM); as well as potential afore-mentioned expansion of the collected dataset. The second encompasses a non-quasi experimental empirical interventional study with a higher amount of participants; a long-term non-quasi experimental empirical interventional study; extracting novel insights about mental health using this work's system; and comparing different large language models in their performance with participants.

Appendix A

Supplementary Material

A.1 Pre-Study Diary Entry Guidelines

Please, fill out the Positive and Negative Affect Schedule questionnaire. At its end, you will encounter a text box which serves as a space for your diary entry. Above it, you will see guidelines in the form of open questions. Follow them to formulate your entry. To know what to expect, you can find these questions below. Don't worry, the questions appear again when you have to write your diary entry! Please, be mindful that your entry is approximately 150 words at minimum. There is no upper word limit. If it helps, you can prepare the entry in a word editor (e.g., Notepad, MS Word on your personal computer, or an appropriate app on a smartphone) and copy-paste it into the text box.

Diary entry guidelines in the form of open questions:

1. Describe your mood.
2. Describe how your mood affected your experience:
   (a) of yourself
   (b) towards the world and its elements
3. Describe how these experiences have changed from yesterday to today.
   (a) Change of experience towards yourself from yesterday to today.
   (b) Change of experience towards the world and its elements from
4. Factual information from the last day that you would like to highlight.

A.2 Study Instructions

Instructions for the study

Characterizing and Forecasting Idiographic and Nomothetic Change in Mental Health

1. About the study

The goal of the study is twofold:

1) Characterize daily change in mental health that we can infer from a mixed methods (quantitative and qualitative data) design in order to understand the phenomenon of change itself better.
2) Discern whether change in mental health can be detected and forecast up to 7 days in advance.

The study employs ecological momentary assessment to collect daily data encompassing a quantitative questionnaire on symptoms occurring in stress, anxiety and depression, and a textual diary entry with some specified guidelines.

1.1. The quantitative questionnaire

Please read each statement and circle a number 0, 1, 2 or 3 which indicates how much the statement applied to you over the past day. There are no right or wrong answers. Do not spend too much time on any statement.

The rating scale is as follows:
0 Did not apply to me at all
1 Applied to me to some degree, or some of the time
2 Applied to me to a considerable degree, or a good part of time
3 Applied to me very much, or most of the time

For example, if you felt extremely nervous for a short period of time, or if you felt nervous for the whole day, you would answer with “3”. However, if you felt slightly nervous for a short period of time, or if you felt nervous for just some of the day, you would answer with “1”.

1. I was unable to relax
2. I was nervous
3. I was terrified, afraid, or scared
4. I felt tightness in my chest
5. I felt lightheaded
6. I felt hot or cold regardless of the surrounding temperature
7. I was trembling or unsteady
8. My heart was pounding in the absence physical exertion
9. I was sad
10. I disliked myself
11. I didn’t feel pleasure from the things I enjoy

12. I didn't feel interested in people or things
13. I felt like a failure
14. I felt my future was hopeless
15. I had trouble making decisions
16. I felt fatigue or heaviness
17. The world was not open and inviting to me, and it had less possibilities
18. It was hard to relate to others

1.2. The diary entry guidelines

The diary entry should generally encompass this day: from the moment you wake up (but you can also include the sleep preceding it) to:

a) the moment of starting the questionnaire if you are completing it in the same day, OR
b) the moment of going to sleep if you are completing it the next day.

Be mindful that your entry is approximately 150 words at minimum (approx. 15 lines in the app; for the website, you can copy-paste the text in a word editor to see the word count). There is no upper word limit. You can write freely and in any order; do not include the bullet points. You can also include text-based emojis. Try to write in a formal style of language, without using contractions (e.g., use "do not", not "don't"). Consider the following:

- Describe your mood(s) over the last day, how it (they) affected your experience of yourself, towards the world, and towards other people and things.
- If you noticed any change in this from the previous day, describe it. How did the change feel, how did you experience it?
- Add any additional experiences and factual information from the last day that you might want to highlight.

Both altogether should take about 15 minutes to complete.

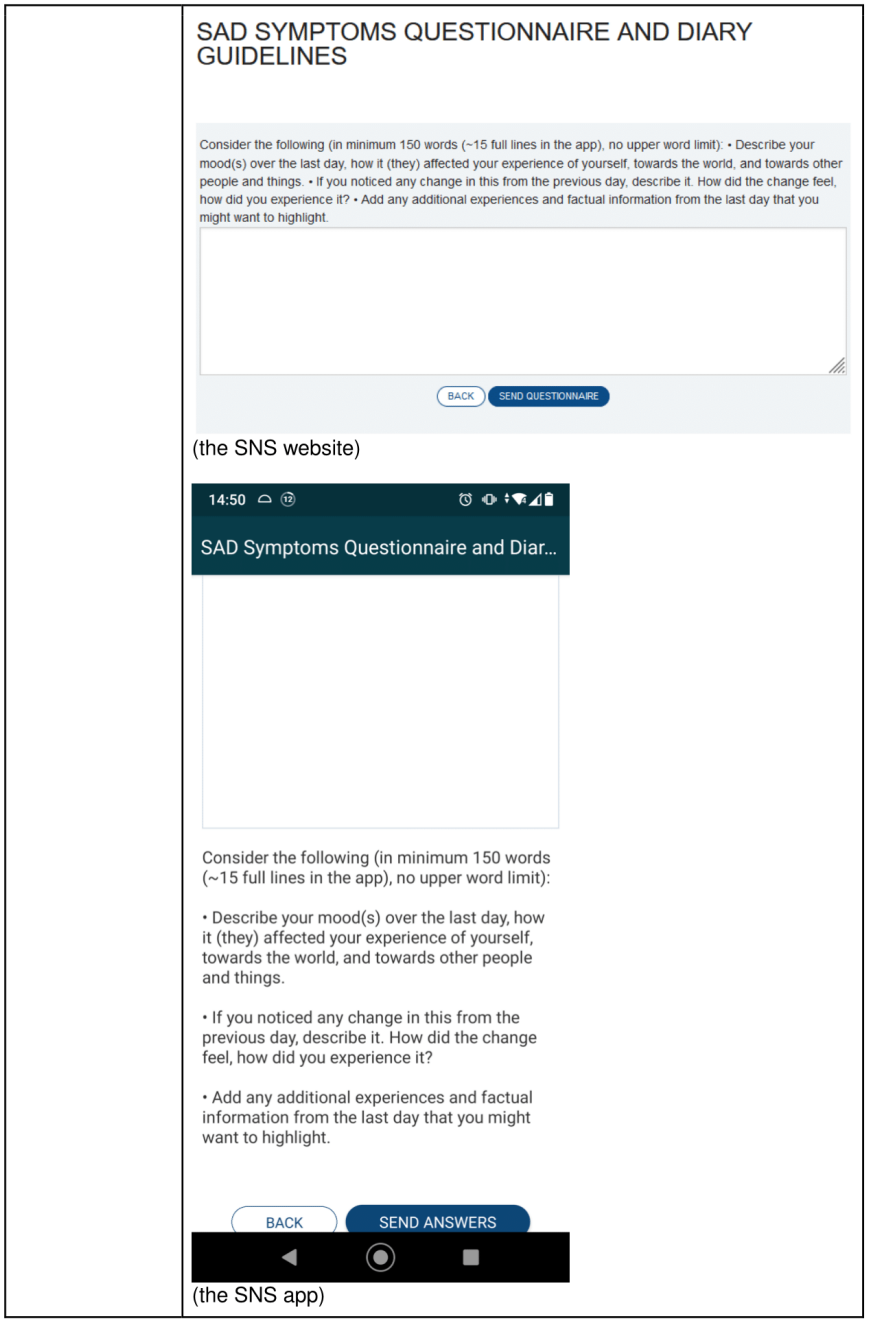
The study uses the Synergetic Navigation System (SNS) tool, which includes an app and a website for completing the daily questionnaire.

The study has been reviewed and approved by an ethics committee (*Ethical approval code: cafiancimhumema_2021-07-13*).

2. How to participate

Step	**Instruction**
Download the SNS app.	The SNS app can be found here (access the link through your phone or alternatively, find it in Google Play store/Apple store app on your phone by searching the app's name - Synergetic Navigation System): - for Android:

	https://play.google.com/store/apps/details?id=de.ccsys.sns.app&gl=US - For iOS: https://apps.apple.com/tr/app/sns/id1333494245
Change the SNS app settings on your phone to allow all the notifications.	Changing the app notification settings enables you to receive daily notifications for questionnaire taking. This will guarantee that you do not forget to provide the data as well as ensure some kind of equidistance in the data collection. You can do so by going into Settings → Apps & notifications → SNS and then allowing all the possible notifications. Some phones have a different path to enabling these options. If there are any problems, send an email to [anonymized for review purposes].
Sign the informed consent form.	The consent form can be downloaded here: https://drive.google.com/file/d/130oknM3-Nym7f1-HOqmQgLJO6hKWyA0C/view?usp=sharing. Sign it (physically or digitally) and email to: [anonymized for review purposes].
Log into your account after receiving your account information.	Sometime after completing the Google Forms form and before the study starts for you, you will receive your anonymized username (following this practice for generating anonymous IDs) and password that enable you to participate in the study. Upon receiving them, log in to your account on the SNS website: https://sfu-ljubljana.sns-live.de/login. After logging in for the first time, change your password. Then, log in to your account in the SNS app. You do so by: 1. opening the downloaded SNS app 2. clicking INPUT MANUALLY 3. inputting the following information: a. under URL, enter https://sfu-ljubljana.sns-live.de b. under Username and Password, enter your allocated username and the password you chose in the previous step. 4. clicking AGREE
Take the daily quantitative questionnaire and provide the diary entry.	Each day from 19:00 onwards, a questionnaire will open up and be available to you. You should receive a notification on your phone and through email, letting you know that you can take the questionnaire. When you want to complete the questionnaire (preferably as close to the end of the day as possible), you either enter the SNS website (https://sfu-ljubljana.sns-live.de) or open the app, and you should have a questionnaire available to you in the “Current questionnaires” section of the entry screen of the SNS app. If you missed a questionnaire from the previous day and still want to provide the information for it, the questionnaire will be available under the headline “Missed questionnaires”. If you prefer completing the questionnaire for a given day the following day, you can do that. The questionnaire consists of the symptoms questionnaire first and the diary second. After you finish with the first, be careful not to skip the diary part. The diary should appear like this:


SAD SYMPTOMS QUESTIONNAIRE AND DIARY GUIDELINES
Consider the following (in minimum 150 words (~15 full lines in the app), no upper word limit): • Describe your mood(s) over the last day, how it (they) affected your experience of yourself, towards the world, and towards other people and things. • If you noticed any change in this from the previous day, describe it. How did the change feel, how did you experience it? • Add any additional experiences and factual information from the last day that you might want to highlight.
BACK
SEND QUESTIONNAIRE
(the SNS website)
14:50
SAD Symptoms Questionnaire and Diar...
Consider the following (in minimum 150 words (~15 full lines in the app), no upper word limit):
• Describe your mood(s) over the last day, how it (they) affected your experience of yourself, towards the world, and towards other people and things.
• If you noticed any change in this from the previous day, describe it. How did the change feel, how did you experience it?
• Add any additional experiences and factual information from the last day that you might want to highlight.
BACK
SEND ANSWERS
(the SNS app)

	Only click “SEND QUESTIONNAIRE” (website)/”SEND ANSWERS” (app) after providing the diary entry. For the website, you can prepare the diary entry in any text editor beforehand and copy-paste it into the text box if that is easier to you. The questionnaire stays open until the end of the day in question. After that time window, the questionnaire will move into the “Missed questionnaires” section. Don’t worry, you can still take the questionnaire even when in that section. Lastly, please use English when writing your diary entry. If your English is not good, and you would prefer to use Slovene, send me an email at [anonymized for review purposes].

During the study, you will be also asked to complete a demographic and a personality questionnaire. The questionnaires will, similar to the daily one, appear in your app and in your website profile.

3. Other

3.1. Missing days in completing the questionnaire

If you miss a day due to any reason, that should be fine. You can also provide data on the missed day the following day. However, try to not miss two days in a row without providing data for the day before. If for any reason you don’t feel comfortable sharing some parts of your daily life in the diary entry, there is no problem in omitting them. E.g., if you feel comfortable sharing your mood for the last day but not the reason for it, feel free to do so. Only ever share what you feel like you want to or feel comfortable enough sharing.

3.2. Data privacy

The data will be thoroughly anonymized (method), will be held on a secure server and only used for research purposes. If at any point you decide to withdraw your data, we will delete it from the server.

3.3. Contact

If you have any questions regarding the study described in 1. or any problems with any of the steps in 2., please contact [anonymized for review purposes]. Unfortunately, taking part in this study has no material compensation, but you will have my eternal love and gratitude.

A.3 Study Diary Entry Guidelines

The diary entry should generally encompass this day: from the moment you wake up, but you can also include the sleep preceding it, to: a) the moment of starting the questionnaire if you are completing it in the same day, OR b) the moment of going to sleep if you are completing it the next day. Be mindful that your entry is approximately 150 words at minimum. There is no upper word limit. You can write freely and in any order; do not include the bullet points. You can also include text-based emojis. Try to write in a formal style of language, without using contractions (e.g., use “do not”, not “don’t”). Consider the following:

Describe your mood(s) over the last day, how it (they) affected your experience

of yourself, towards the world, and towards other people and things.

If you noticed any change in this from the previous day, describe it. How did the change feel, how did you experience it?

Add any additional experiences and factual information from the last day that you might want to highlight.

A.4 Study Demographic Questionnaire

1. Date of birth
2. Sex assigned at birth
3. Gender identity
4. Highest educational attainment
5. Overall how would you rate your mental health?
6. Have you ever been diagnosed with a mental disorder?
7. If you currently have a mental disorder diagnosis, what is it? (optional, you don't have to disclose it if you feel uncomfortable)
8. Have you had mental health-related therapy in the recent past?
9. Are you currently taking any medication for mental disorders?
10. How many hours do you sleep per day on average?
11. How is your quality of sleep on average?
12. How would you self-describe your emotional valence?
13. How would you self-describe your emotional arousal?

A.5 Study 18-item SAD Questionnaire

1. I was unable to relax
2. I was nervous
3. I was terrified, afraid, or scared
4. I felt tightness in my chest
5. I felt lightheaded
6. I felt hot or cold regardless of the surrounding temperature
7. I was trembling or unsteady
8. My heart was pounding in the absence of physical exertion
9. I was sad
10. I disliked myself

11. I didn't feel pleasure from the things I enjoy
12. I didn't feel interested in people or things
13. I felt like a failure
14. I felt my future was hopeless
15. I had trouble making decisions
16. I felt fatigue or heaviness
17. The world was not open and inviting to me, and it had less possibilities
18. It was hard to relate to others

A.6 Post-Study Questionnaire

1. Approximately how much time did you spend daily on the questionnaire?
2. Were the guidelines about the diary entry clear? Did anything bother or confuse you? Would you change anything about them?
3. What did you think about the 150 word minimum for the diary?
4. When you were completing the questionnaire, did you have in mind the last 24 hours (including the night of the previous day) or just the day in question?
5. When completing the questionnaire, what time (in the study, the questionnaire was open from 19:00 to 23:45) would be representative of your entire day for you to capture it in your answers?
6. Would you rather complete the questionnaire in the morning/before noon for the whole previous day?
7. How would you rate the SNS tool in terms of its usability and how comfortable it was to complete the questionnaires with?
8. Did you prefer using the app or the website?
9. Did you use the app or the website more when completing questionnaires?
10. While completing the questionnaires, did you feel your responses were influenced by you knowing what this research was about, or did you rather focus on the questions at hand without worrying about how the researchers want you to respond?
11. Do you have any other comments about the study and how you participated in it?
12. Would you be willing to participate in a similar research in the future?

References

[1] F. Ornell, W. V. Borelli, D. Benzano, *et al.*, "The next pandemic: Impact of covid-19 in mental healthcare assistance in a nationwide epidemiological study," *The Lancet Regional Health – Americas*, vol. 4, Dec. 2021, ISSN: 2667-193X. DOI: 10.1016/j.lana.2021.100061. [Online]. Available: https://doi.org/10.1016/j.lana.2021.100061.

[2] World Health Organization, *Mental Health Action Plan 2013-2020*. Geneva: World Health Organization, 2003.

[3] Mental Health Foundation, *Stress: Are we coping?* London: Mental Health Foundation, 2018.

[4] A. J. Baxter, K. M. Scott, T. Vos, and H. A. Whiteford, "Global prevalence of anxiety disorders: A systematic review and meta-regression," *Psychological Medicine*, vol. 43, no. 5, pp. 897–910, 2013. DOI: 10.1017/S003329171200147X.

[5] J. Twenge, "Time Period and Birth Cohort Differences in Depressive Symptoms in the U.S., 1982–2013," *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*, vol. 121, no. 2, pp. 437–454, Apr. 2015. DOI: 10.1007/s11205-014-0647-1. [Online]. Available: https://ideas.repec.org/a/spr/soinre/v121y2015i2p437-454.html.

[6] P. S. Wang, S. Aguilar-Gaxiola, J. Alonso, *et al.*, "Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys," *Lancet*, vol. 370, no. 9590, pp. 841–850, Sep. 2007.

[7] A. Schmidtke, U. Bille-Brahe, D. DeLeo, *et al.*, "Attempted suicide in Europe: rates, trends and sociodemographic characteristics of suicide attempters during the period 1989-1992. Results of the WHO/EURO Multicentre Study on Parasuicide," *Acta Psychiatr Scand*, vol. 93, no. 5, pp. 327–338, May 1996.

[8] World Health Organization, *Investing in Mental Health*. World Health Organization, 2003. [Online]. Available: https://apps.who.int/iris/handle/10665/42823.

[9] S. C. Curtin, M. Warner, and H. Hedegaard, "Increase in Suicide in the United States, 1999-2014," *NCHS Data Brief*, no. 241, pp. 1–8, Apr. 2016.

[10] P. Winkler, D. Krupchanka, T. Roberts, *et al.*, "A blind spot on the global mental health map: a scoping review of 25 years' development of mental health care for people with severe mental illnesses in central and eastern Europe," *Lancet Psychiatry*, vol. 4, no. 8, pp. 634–642, Aug. 2017.

[11] European Commission, *Inequalities in access to healthcare, A study of national policies*. European Commission, 2018. [Online]. Available: https://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=815.

[12] J. Auerbach and B. F. Miller, "Covid-19 exposes the cracks in our already fragile mental health system," *American Journal of Public Health*, vol. 110, no. 7, pp. 969–970, 2020, PMID: 32271609. DOI: 10.2105/AJPH.2020.305699. eprint: https://doi.org/10.2105/AJPH.2020.305699. [Online]. Available: https://doi.org/10.2105/AJPH.2020.305699.

[13] M. C. Angermeyer and H. Matschinger, "The effect of personal experience with mental illness on the attitude towards individuals suffering from mental disorders," *Social Psychiatry and Psychiatric Epidemiology*, vol. 31, no. 6, pp. 321–326, Nov. 1996.

[14] M. Moutoussis, N. Shahar, T. U. Hauser, and R. J. Dolan, "Computation in psychotherapy, or how computational psychiatry can aid Learning-Based psychological therapies," en, *Comput Psychiatr*, vol. 2, pp. 50–73, Feb. 2018.

[15] M. Gams and T. Kolenik, "Relations between electronics, artificial intelligence and information society through information society rules," *Electronics*, vol. 10, no. 4, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10040514. [Online]. Available: https://www.mdpi.com/2079-9292/10/4/514.

[16] B. Fogg, *Persuasive Technology: Using Computers to Change What We Think and Do*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[17] R. Orji and K. Moffatt, "Persuasive technology for health and wellness: State-of-the-art and emerging trends," *Health Informatics Journal*, vol. 24, no. 1, pp. 66–91, 2018. DOI: 10.1177/1460458216650979.

[18] J. Oakley, *Intelligent cognitive assistants (ica)*, 2018. [Online]. Available: https://www.nsf.gov/crssprgm/nano/reports/ICA2_Workshop_Report_2018.pdf.

[19] *Cognitive architecture*, http://cogarch.ict.usc.edu/, Accessed: 2020-05-30.

[20] S. Garrod and M. J. Pickering, "Why is conversation so easy?" *Trends in Cognitive Sciences*, vol. 8, no. 1, pp. 8–11, 2004, ISSN: 1364-6613.

[21] D. Bakker, N. Kazantzis, D. Rickwood, and N. Rickard, "Mental health smartphone apps: Review and evidence-based recommendations for future developments," *JMIR Mental Health*, vol. 3, no. 1, e7, Mar. 2016, ISSN: 2368-7959. DOI: 10.2196/mental.4984. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26932350.

[22] L. Laranjo, A. G. Dunn, H. L. Tong, *et al.*, "Conversational agents in healthcare: a systematic review," *J Am Med Inform Assoc*, vol. 25, no. 9, pp. 1248–1258, Sep. 2018.

[23] J. L. Z. Montenegro, C. A. da Costa, and R. da Rosa Righi, "Survey of conversational agents in health," *Expert Systems with Applications*, vol. 129, pp. 56–67, 2019, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.03.054. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417419302283.

[24] S. Provoost, H. M. Lau, J. Ruwaard, and H. Riper, "Embodied Conversational Agents in Clinical Psychology: A Scoping Review," *J. Med. Internet Res.*, vol. 19, no. 5, e151, May 2017.

[25] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape," *Can J Psychiatry*, vol. 64, no. 7, pp. 456–464, Jul. 2019.

[26] A. A. Abd-Alrazaq, A. Rababeh, M. Alajlani, B. M. Bewick, and M. Househ, "Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis," *J Med Internet Res*, vol. 22, no. 7, e16021, Jul. 2020, ISSN: 1438-8871. DOI: 10.2196/16021. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/32673216.

[27] H. Gaffney, W. Mansell, and S. Tai, "Conversational agents in the treatment of mental health problems: Mixed-method systematic review," *JMIR Ment Health*, vol. 6, no. 10, e14166, Oct. 2019, ISSN: 2368-7959. DOI: 10.2196/14166. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/31628789.

[28] E. Bendig, B. Erb, L. Schulze-Thuesing, and H. Baumeister, "The next generation: Chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review," *Verhaltenstherapie*, 2019, ISSN: 1016-6262. DOI: 10.1159/000501812. [Online]. Available: https://doi.org/10.1159/000501812.

[29] A. A. Abd-alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, and M. Househ, "An overview of the features of chatbots in mental health: A scoping review," *International Journal of Medical Informatics*, vol. 132, p. 103 978, 2019, ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2019.103978. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505619307166.

[30] T. Kolenik, "Methods in digital mental health: Smartphone-based assessment and intervention for stress, anxiety and depression," in *Integrating Artificial Intelligence and IoT for Advanced Health Informatics*, C. Comito, A. Forestiero, and E. Zumpano, Eds., In press, Springer, 2021.

[31] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020. arXiv: 2005.14165. [Online]. Available: https://arxiv.org/abs/2005.14165.

[32] D. M. Korngiebel and S. D. Mooney, "Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery," *npj Digital Medicine*, vol. 4, no. 1, p. 93, Jun. 2021, ISSN: 2398-6352. DOI: 10.1038/s41746-021-00464-x. [Online]. Available: https://doi.org/10.1038/s41746-021-00464-x.

[33] H. H. Thorp, *Chatgpt is fun, but not an author*, 2023.

[34] Z. Kansoun, L. Boyer, M. Hodgkinson, V. Villes, C. Lançon, and G. Fond, "Burnout in french physicians: A systematic review and meta-analysis," *Journal of affective disorders*, vol. 246, pp. 132–147, 2019.

[35] *Stress*, https://www.mentalhealth.org.uk/a-to-z/s/stress, Last accessed on 2021-05-29.

[36] R. B. Cialdini, *Influence : science and practice*. Boston: Pearson Education, 2009, ISBN: 0205609996 9780205609994 9780205663781 0205663788. [Online]. Available: http://www.amazon.co.uk/Influence-Practice-Robert-B-Cialdini/dp/0205663788.

[37] D. Kahneman, *Thinking, fast and slow*. New York: Farrar, Straus and Giroux, 2011, ISBN: 9780374275631 0374275637. [Online]. Available: https://www.amazon.de/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374275637/ref=wl_it_dp_o_pdT1_nS_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I3OCESLZCVDFL7.

[38] R. Thaler and C. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008, ISBN: 9780300146813. [Online]. Available: https://books.google.si/books?id=dSJQn8egXvUC.

[39] H. Berghel, "Malice domestic: The cambridge analytica dystopia," *Computer*, vol. 51, no. 5, pp. 84–89, 2018. DOI: 10.1109/MC.2018.2381135.

[40] C. Midden, T. Mccalley, J. Ham, and R. Zaalberg, "Using persuasive technology to encourage sustainable behavior," *Journal of Applied Mechanics-transactions of The Asme - J APPL MECH*, Jan. 2008.

[41] S. Gram-Hansen, T. Svarre, and C. Midden, *Proceedings of the 15th International Conference on Persuasive Technology (PERSUASIVE 2020), Aalborg, Denmark, April 20–23, 2020*. Jan. 2020, ISBN: 978-3-030-45711-2. DOI: 10.1007/978-3-030-45712-9.

[42] H. Oinas-Kukkonen and M. Harjumaa, "Persuasive systems design: Key issues, process model, and system features," *Communications of the Association for Information Systems*, vol. 24, Mar. 2009. DOI: 10.17705/1CAIS.02428.

[43] S. Gkika, M. Skiada, G. Lekakos, and P. E. Kourouthanassis, "Investigating the role of personality traits and influence strategies on the persuasive effect of personalized recommendations," in *EMPIRE@RecSys*, 2016.

[44] T. Kolenik and M. Gams, "PerMEASS – Personal Mental Health Virtual Assistant with Novel Ambient Intelligence Integration," in *http://ceur-ws.org/Vol-2820/*, CEUR-WS, Santiago de Compostela, Spain, 2020, pp. 8–12. [Online]. Available: http://ceur-ws.org/Vol-2820/AAI4H-2.pdf.

[45] H. Michener, J. DeLamater, and D. Myers, *Social Psychology* (Available Titles Cengagenow). Wadsworth/Thomson Learning, 2003, ISBN: 9780534583217. [Online]. Available: https://books.google.si/books?id=dgxHPgAACAAJ.

[46] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007, ISSN: 0092-6566. DOI: https://doi.org/10.1016/j.jrp.2006.02.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092656606000195.

[47] K. Lee and M. C. Ashton, "Hexaco model of personality structure, the," in *Encyclopedia of Personality and Individual Differences*, V. Zeigler-Hill and T. K. Shackelford, Eds. Cham: Springer International Publishing, 2020, pp. 1932–1936, ISBN: 978-3-319-24612-3. DOI: 10.1007/978-3-319-24612-3_1227. [Online]. Available: https://doi.org/10.1007/978-3-319-24612-3_1227.

[48] S. Lovibond and P. Lovibond, *Manual for the Depression Anxiety Stress Scales* (Psychology Foundation monograph). Psychology Foundation of Australia, 1996, ISBN: 9780733414237. [Online]. Available: https://books.google.si/books?id=mXoQHAAACAAJ.

[49] M. Ratcliffe, *Experiences of Depression: A Study in Phenomenology* (International Perspectives in). Oxford University Press, 2015, ISBN: 9780199608973. [Online]. Available: https://books.google.si/books?id=0UePBQAAQBAJ.

[50] F. Delahunty, I. D. Wood, and M. Arcan, "First insights on a passive major depressive disorder prediction system with incorporated conversational chatbot," in *Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, 2018, pp. 327–338.

[51] K. Denecke, S. Vaaheesan, and A. Arulnathan, "A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1, 2020. DOI: 10.1109/TETC.2020.2974478.

[52] M. Gjoreski, T. Kolenik, T. Knez, *et al.*, "Datasets for cognitive load inference using wearable sensors and psychological traits," *Applied Sciences*, vol. 10, no. 11, 2020, ISSN: 2076-3417. DOI: 10.3390/app10113843. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3843.

[53] Wikipedia contributors, *Eliza — Wikipedia, the free encyclopedia*, [Online; accessed 23-April-2021], 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=ELIZA&oldid=1012889844.

[54] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial," *JMIR Ment Health*, vol. 4, no. 2, e19, Jun. 2017, ISSN: 2368-7959. DOI: 10.2196/mental.7785. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28588005.

[55] *Microsoft's bing is an emotionally manipulative liar, and people love it*, The Verge, Feb. 2023. [Online]. Available: https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams.

[56] R. Daws, Applications, Ethics, Developers, Adoption, and Hardware, *Medical chatbot using openai's gpt-3 told a fake patient to kill themselves*, Oct. 2020. [Online]. Available: https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/.

[57] R. R. Morris, K. Kouddous, R. Kshirsagar, and S. M. Schueller, "Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions," *J Med Internet Res*, vol. 20, no. 6, e10148, Jun. 2018, ISSN: 1438-8871. DOI: 10.2196/10148. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29945856.

[58] H. N. Io and C. B. Lee, "Chatbots and conversational agents: A bibliometric analysis," in *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2017, pp. 215–219. DOI: 10.1109/IEEM.2017.8289883.

[59] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Computers in Human Behavior*, vol. 37, pp. 94–100, 2014, ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2014.04.043. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563214002647.

[60] P. McCrone, M. Knapp, J. Proudfoot, *et al.*, "Cost-effectiveness of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trial," *Br J Psychiatry*, vol. 185, pp. 55–62, Jul. 2004.

[61] J. Jetter, "The good, the bad, and the aesthetically challenged: The good, the bad, and the aesthetically challenged [opinion]," *IEEE Technology and Society Magazine*, vol. 38, no. 4, pp. 27–31, 2019. DOI: 10.1109/MTS.2019.2952297.

[62] B. Cliffe, A. Croker, M. Denne, and P. Stallard, "Clinicians' use of and attitudes towards technology to provide and support interventions in child and adolescent mental health services," *Child and Adolescent Mental Health*, vol. 25, no. 2, pp. 95–101, 2020. DOI: 10.1111/camh.12362. eprint: https://acamh.onlinelibrary.

wiley.com/doi/pdf/10.1111/camh.12362. [Online]. Available: https://acamh.onlinelibrary.wiley.com/doi/abs/10.1111/camh.12362.

[63] M. Price, E. K. Yuen, E. M. Goetter, *et al.*, "mHealth: a mechanism to deliver more accessible, more effective mental health care," *Clin Psychol Psychother*, vol. 21, no. 5, pp. 427–436, 2014.

[64] N. Freedman, J. D. Hoffenberg, N. Vorus, and A. Frosch, "The effectiveness of psychoanalytic psychotherapy: the role of treatment duration, frequency of sessions, and the therapeutic relationship," *J Am Psychoanal Assoc*, vol. 47, no. 3, pp. 741–772, 1999.

[65] R. Sandell, J. Blomberg, A. Lazar, J. Carlsson, J. Broberg, and J. Schubert, "Varieties of long-term outcome among patients in psychoanalysis and long-term psychotherapy. A review of findings in the Stockholm Outcome of Psychoanalysis and Psychotherapy Project (STOPP)," *Int J Psychoanal*, vol. 81 (Pt 5), pp. 921–942, Oct. 2000.

[66] P. Corrigan and A. Watson, "The impact of stigma on people with mental illness," *World psychiatry : official journal of the World Psychiatric Association (WPA)*, vol. 1, pp. 16–20, Mar. 2002.

[67] G. Thornicroft, S. Chatterji, S. Evans-Lacko, *et al.*, "Undertreatment of people with major depressive disorder in 21 countries," *Br J Psychiatry*, vol. 210, no. 2, pp. 119–124, Feb. 2017.

[68] I. Amaral and F. Daniel, "Ageism and it: Social representations, exclusion and citizenship in the digital age," in *Human Aspects of IT for the Aged Population. Healthy and Active Aging*, J. Zhou and G. Salvendy, Eds., Cham: Springer International Publishing, 2016, pp. 159–166, ISBN: 978-3-319-39949-2.

[69] M. Pigato, "Information and communication technology, poverty, and development in sub-saharan africa and south asia," 20, Washington, D.C.: The World Bank, 2001.

[70] S.-G. Lee, S. Trimi, and C. Kim, "The impact of cultural differences on technology adoption," *Journal of World Business*, vol. 48, no. 1, pp. 20–29, 2013, ISSN: 1090-9516. DOI: https://doi.org/10.1016/j.jwb.2012.06.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1090951612000405.

[71] *AfriCHI '18: Proceedings of the Second African Conference for Human Computer Interaction: Thriving Communities*, Windhoek, Namibia: Association for Computing Machinery, 2018, ISBN: 9781450365581.

[72] A. Yorita, S. Egerton, J. Oakman, C. Chan, and N. Kubota, "A robot assisted stress management framework: Using conversation to measure occupational stress," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 3761–3767.

[73] S. Avancha, A. Baxi, and D. Kotz, "Privacy in mobile technology for personal healthcare," *ACM Comput. Surv.*, vol. 45, no. 1, 2012, ISSN: 0360-0300. DOI: 10.1145/2379776.2379779. [Online]. Available: https://doi.org/10.1145/2379776.2379779.

[74] S.-S. Lee, Y.-k. Lim, and K.-p. Lee, "A long-term study of user experience towards interaction designs that support behavior change," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '11, Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 2065–2070, ISBN: 9781450302685. DOI: 10.1145/1979742.1979909. [Online]. Available: https://doi.org/10.1145/1979742.1979909.

[75] D. B. Klein, "Statist Quo Bias," *JEconomic Journal Watch*, vol. 1, pp. 260–271, 2004.

[76] N. Eyal and R. Hoover, *Hooked: How to Build Habit-forming Products*. Portfolio Penguin, 2014, ISBN: 9780241184837. [Online]. Available: https://books.google.si/books?id=YNZZoAEACAAJ.

[77] J. Pitt, "From trust and loyalty to lock-in and digital dependence [editorial]," *IEEE Technology and Society Magazine*, vol. 39, no. 1, pp. 5–8, 2020. DOI: 10.1109/MTS.2020.2967483.

[78] Australian Human Rights Commission, *Human rights and technology*, 2019. [Online]. Available: https://humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-discussion-paper-2019.

[79] R. Silva and F. Neiva, "Systematic literature review in computer science - a practical guide," Federal University of Juiz de Fora, Tech. Rep., 2016.

[80] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," *International Journal of Social Research Methodology*, vol. 8, no. 1, pp. 19–32, 2005. DOI: 10.1080/1364557032000119616. eprint: https://doi.org/10.1080/1364557032000119616. [Online]. Available: https://doi.org/10.1080/1364557032000119616.

[81] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*. Wiley, 2008, ISBN: 9781405150149. [Online]. Available: https://books.google.si/books?id=ZwZ1%5C_xU3E80C.

[82] J. Howland, T. Wright, R. Boughan, and B. Roberts, "How Scholarly Is Google Scholar? A Comparison to Library Databases," *College & Research Libraries*, vol. 70, pp. 227–234, May 2009. DOI: 10.5860/crl.70.3.227.

[83] W. Walters, "Google Scholar coverage of a multidisciplinary field," *Information Processing & Management*, vol. 43, pp. 1121–1132, Jul. 2007. DOI: 10.1016/j.ipm.2006.08.006.

[84] A.-W. Harzing and S. Alakangas, "Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison," *Scientometrics*, vol. 106, no. 2, pp. 787–804, Feb. 2016, ISSN: 1588-2861. DOI: 10.1007/s11192-015-1798-9. [Online]. Available: https://doi.org/10.1007/s11192-015-1798-9.

[85] E. D. López-Cózar, E. Orduna-Malea, and A. Martín-Martín, *Google Scholar as a data source for research assessment*, 2018. arXiv: 1806.04435 [cs.DL].

[86] H. Ritchie and M. Roser, "Mental health," *Our World in Data*, 2018. [Online]. Available: https://ourworldindata.org/mental-health.

[87] A. Abbott, "COVID's mental-health toll: how scientists are tracking a surge in depression," *Nature*, vol. 590, no. 7845, pp. 194–195, Feb. 2021.

[88] I. Medhi Thies, N. Menon, S. Magapu, M. Subramony, and J. O'Neill, "How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians," in *Human-Computer Interaction - INTERACT 2017*, R. Bernhaupt, G. Dalvi, A. Joshi, D. K. Balkrishan, J. O'Neill, and M. Winckler, Eds., Cham: Springer International Publishing, 2017, pp. 441–459, ISBN: 978-3-319-67744-6.

[89] G. Albright, C. Adam, D. Serri, S. Bleeker, and R. Goldman, "Harnessing the power of conversations with virtual humans to change health behaviors," *mHealth*, vol. 2, no. 11, 2016, ISSN: 2306-9740. [Online]. Available: https://mhealth.amegroups.com/article/view/12530.

[90] Linchpin, *25 chatbot stats and trends shaping businesses in 2021*, https://linchpinseo.com/chatbot-statistics-trends/, Last accessed on 2021-05-10, 2021.

[91] C. Yoon, C. Jeong, and E. Rolland, "Understanding individual adoption of mobile instant messaging: A multiple perspectives approach," *Information Technology and Management*, vol. 16, no. 2, pp. 139–151, Jun. 2015, ISSN: 1573-7667. DOI: 10.1007/s10799-014-0202-4. [Online]. Available: https://doi.org/10.1007/s10799-014-0202-4.

[92] R. Fulmer, A. Joerin, B. Gentile, L. Lakerink, and M. Rauws, "Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: Randomized controlled trial," *JMIR Ment Health*, vol. 5, no. 4, e64, Dec. 2018, ISSN: 2368-7959. DOI: 10.2196/mental.9782. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30545815.

[93] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," *JMIR Mhealth Uhealth*, vol. 6, no. 11, e12106, Nov. 2018, ISSN: 2291-5222. DOI: 10.2196/12106. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30470676.

[94] *Dialogflow documentation*, https://cloud.google.com/dialogflow/docs/. [Online]. Available: https://cloud.google.com/dialogflow/docs/.

[95] *Ibm watson*, https://www.ibm.com/watson. [Online]. Available: https://www.ibm.com/watson.

[96] *Microsoft bot framework*, https://dev.botframework.com/. [Online]. Available: https://dev.botframework.com/.

[97] *Gpt-3*, https://gpt3.website/. [Online]. Available: https://gpt3.website/.

[98] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021. DOI: 10.1136/bmj.n71. eprint: https://www.bmj.com/content/372/bmj.n71.full.pdf. [Online]. Available: https://www.bmj.com/content/372/bmj.n71.

[99] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Emma: An emotion-aware wellbeing chatbot," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 1–7. DOI: 10.1109/ACII.2019.8925455.

[100] S. Khadikar, P. Sharma, and P. Paygude, "Compassion driven conversational chatbot aimed for better mental health," *Zeichen Journal*, vol. 6, no. 9, pp. 121–127, 2020.

[101] S. Park, J. Choi, S. Lee, *et al.*, "Designing a chatbot for a brief motivational interview on stress management: Qualitative case study," *J Med Internet Res*, vol. 21, no. 4, e12231, Apr. 2019, ISSN: 1438-8871. DOI: 10.2196/12231. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30990463.

[102] S. Pola and M. Sheela Rani Chetty, "Behavioral therapy using conversational chatbot for depression treatment using advanced rnn and pretrained word embeddings," *Materials Today: Proceedings*, 2021, ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2021.02.521.

[103] C. Rishabh and J. Anuradha, "Counsellor chatbot," *International Research Journal of Computer Science*, vol. 3, no. 5, pp. 126–136, 2018.

[104] Wikipedia contributors, *Artificial linguistic internet computer entity — Wikipedia, the free encyclopedia*, [Online; accessed 24-April-2021], 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Artificial_Linguistic_Internet_Computer_Entity&oldid=993396811.

[105] A. Yorita, S. Egerton, C. Chan, and N. Kubota, "Chatbot for peer support realization based on mutual care," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 1601–1606. DOI: 10.1109/SSCI47803.2020.9308277.

[106] C. E. Gould, F. Ma, J. R. Loup, C. Juang, E. Y. Sakai, and R. Pepin, "Technology-based mental health assessment and intervention," *Handbook of Mental Health and Aging*, pp. 401–415, 2020. DOI: 10.1016/b978-0-12-800136-3.00024-7.

[107] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open source language understanding and dialogue management," *CoRR*, vol. abs/1712.05181, 2017. arXiv: 1712.05181. [Online]. Available: http://arxiv.org/abs/1712.05181.

[108] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Towards understanding emotional intelligence for behavior change chatbots," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 8–14. DOI: 10.1109/ACII.2019.8925433.

[109] A. M. Leslie, O. Friedman, and T. P. German, "Core mechanisms in 'theory of mind'," *Trends in Cognitive Sciences*, vol. 8, no. 12, pp. 528–533, 2004, ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2004.10.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661304002608.

[110] S. Suganuma, D. Sakamoto, and H. Shimoyama, "An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: Feasibility and acceptability pilot trial," *JMIR mental health*, vol. 5, no. 3, e10454–e10454, Jul. 2018, ISSN: 2368-7959. DOI: 10.2196/10454. [Online]. Available: https://doi.org/10.2196/10454.

[111] K. Oyibo, R. Orji, and J. Vassileva, "Investigation of the influence of personality traits on cialdini's persuasive strategies," in *PPT@PERSUASIVE*, 2017.

[112] J. J. Prochaska, E. A. Vogel, A. Chieng, *et al.*, "A therapeutic relational agent for reducing problematic substance use (woebot): Development and usability study," *J Med Internet Res*, vol. 23, no. 3, e24850, Mar. 2021, ISSN: 1438-8871. DOI: 10.2196/24850. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/33755028.

[113] A. Darcy, A. Beaudette, E. Chiauzzi, *et al.*, "Anatomy of a woebot® (wb001): Agent guided cbt for women with postpartum depression," *Expert Review of Medical Devices*, vol. 19, no. 4, pp. 287–301, 2022, PMID: 35748029. DOI: 10.1080/17434440.2022.2075726. eprint: https://doi.org/10.1080/17434440.2022.2075726. [Online]. Available: https://doi.org/10.1080/17434440.2022.2075726.

[114] H. M. Demirci, "User experience over time with conversational agents: Case study of woebot on supporting subjective well-being," M.S. thesis, Middle East Technical University, 2018.

[115] G. Schiepek, O. Gelo, K. Viol, *et al.*, "Complex individual pathways or standard tracks? a data-based discussion on the trajectories of change in psychotherapy," *Counselling and Psychotherapy Research*, vol. 20, no. 4, pp. 689–702, 2020. DOI: https://doi.org/10.1002/capr.12300. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/capr.12300. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/capr.12300.

[116] J. B. Hirsh, S. K. Kang, and G. V. Bodenhausen, "Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits," *Psychological Science*, vol. 23, no. 6, pp. 578–581, 2012, PMID: 22547658. DOI: 10.1177/0956797611436349. eprint: https://doi.org/10.1177/0956797611436349. [Online]. Available: https://doi.org/10.1177/0956797611436349.

[117] A. M. Leslie, O. Friedman, and T. P. German, "Core mechanisms in ‘theory of mind’" *Trends in Cognitive Sciences*, vol. 8, no. 12, pp. 528–533, Dec. 2004, ISSN: 1364-6613. DOI: 10.1016/j.tics.2004.10.001. [Online]. Available: https://doi.org/10.1016/j.tics.2004.10.001.

[118] A. Jain, H. Patel, L. Nagalapatti, *et al.*, "Overview and importance of data quality for machine learning tasks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20, Virtual Event, CA, USA: Association for Computing Machinery, 2020, pp. 3561–3562, ISBN: 9781450379984. DOI: 10.1145/3394486.3406477. [Online]. Available: https://doi.org/10.1145/3394486.3406477.

[119] N. Gupta, S. Mujumdar, H. Patel, *et al.*, "Data quality for machine learning tasks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD '21, Virtual Event, Singapore: Association for Computing Machinery, 2021, pp. 4040–4041, ISBN: 9781450383325. DOI: 10.1145/3447548.3470817. [Online]. Available: https://doi.org/10.1145/3447548.3470817.

[120] G. Schiepek, H. Eckert, B. Aas, S. Wallot, and A. Wallot, *Integrative Psychotherapy A Feedback-Driven Dynamic Systems Approach*. Hogrefe Verlag GmbH & Co. KG, 2015, p. 111, ISBN: 9780889374720. DOI: 10.1027/00472-000. [Online]. Available: https://elibrary.hogrefe.com/book/10.1027/00472-000.

[121] D. J. Pritchard, T. A. Hurly, M. C. Tello-Ramos, and S. D. Healy, "Why study cognition in the wild (and how to test it)?" *J Exp Anal Behav*, vol. 105, no. 1, pp. 41–55, Jan. 2016.

[122] M. Prince, "9 - Epidemiology," in *Core Psychiatry (Third Edition)*, P. Wright, J. Stern, and M. Phelan, Eds., Third Edition, Oxford: W.B. Saunders, 2012, pp. 115–129, ISBN: 978-0-7020-3397-1. DOI: https://doi.org/10.1016/B978-0-7020-3397-1.00009-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780702033971000094.

[123] N. van Berkel, "Data quality and quantity in mobile experience sampling," Ph.D. dissertation, 2019/09/19 2019. [Online]. Available: http://hdl.handle.net/11343/227682.

[124] T. Kubiak and J. M. Smyth, "Connecting domains—ecological momentary assessment in a mobile sensing framework," in *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*, H. Baumeister and C. Montag, Eds. Cham: Springer International Publishing, 2019, pp. 201–207, ISBN: 978-3-030-31620-4. DOI: 10.1007/978-3-030-31620-4_12.

[125] D. Colombo, J. Fernández-Álvarez, A. Patané, *et al.*, "Current state and future directions of technology-based ecological momentary assessment and intervention for major depressive disorder: A systematic review," *Journal of Clinical Medicine*, vol. 8, no. 4, 2019, ISSN: 2077-0383. DOI: 10.3390/jcm8040465. [Online]. Available: https://www.mdpi.com/2077-0383/8/4/465.

[126] M. Olden, R. Holle, I. M. Heid, and K. Stark, "IDGenerator: unique identifier generator for epidemiologic or clinical studies," *BMC Med Res Methodol*, vol. 16, p. 120, Sep. 2016.

[127] A. T. Beck, N. Epstein, G. Brown, and R. Steer, "Beck anxiety inventory," *Journal of Consulting and Clinical Psychology*, 1993.

[128] A. T. Beck, R. A. Steer, G. K. Brown, *et al.*, *Beck depression inventory*. Harcourt Brace Jovanovich New York: 1987.

[129] M. Ratcliffe, *Experiences of depression: A study in phenomenology*. OUP Oxford, 2014.

[130] M. T. Orne, "Demand characteristics," in *Introducing psychological research*, Springer, 1996, pp. 395–401.

[131] D. Friedman, *Machine learning from scratch*, `https://dafriedman97.github.io/mlbook/content/introduction.html`, 2020.

[132] B. Seref and E. Bostanci, "Performance comparison of naıv̈e bayes and complement naıv̈e bayes algorithms," in *2019 6th international conference on electrical and electronics engineering (ICEEE)*, IEEE, 2019, pp. 131–138.

[133] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[134] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[135] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[136] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, 2014, pp. 216–225.

[137] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.

[138] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[139] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The gad-7," *Archives of internal medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[140] J. DiNardo, "Natural experiments and quasi-natural experiments," in *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan UK, 2016, pp. 1–12, ISBN: 978-1-349-95121-5. DOI: `10.1057/978-1-349-95121-5_2006-1`. [Online]. Available: `https://doi.org/10.1057/978-1-349-95121-5_2006-1`.

[141] B. Arapovic-Johansson, C. Wåhlin, L. Kwak, C. Björklund, and I. Jensen, "Work-related stress assessed by a text message single-item stress question," *Occupational Medicine*, vol. 67, no. 8, pp. 601–608, 2017.

[142] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Applying the user experience questionnaire (ueq) in different evaluation scenarios," Jun. 2014, pp. 383–392, ISBN: 978-3-319-07667-6. DOI: `10.1007/978-3-319-07668-3_37`.

[143] K. Jokinen and M. McTear, *Spoken Dialogue Systems* (Synthesis lectures on human language technologies). Cham: Springer International Publishing, 2010.

[144] C. Grosan and A. Abraham, *Intelligent Systems* (Intelligent systems reference library), en. New York, NY: Springer, Jan. 2011.

[145] Wikipedia, *Cognitive behavioral therapy — Wikipedia, the free encyclopedia*, `https://en.wikipedia.org/wiki/Cognitive_behavioral_therapy`, [Online; accessed 01-September-2022], 2022.

[146] P. P. Schmidt and A. Fay, "Applying the domain-mapping-matrix to identify the appropriate level of detail of simulation models for virtual commissioning," *IFAC-PapersOnLine*, vol. 48, no. 10, pp. 69–74, 2015, 2nd IFAC Conference on Embedded Systems, Computer Intelligence and Telematics CESCIT 2015, ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2015.08.110. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896315009775.

[147] A. Alslaity and T. Tran, "The effect of personality traits on persuading recommender system users," in *IntRS'20-Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, 2020, pp. 48–56.

[148] K. Oyibo, R. Orji, and J. Vassileva, "Investigation of the influence of personality traits on cialdini's persuasive strategies.," *PPT@ PERSUASIVE*, vol. 2017, pp. 8–20, 2017.

[149] M. Khanahmadi, M. Malmir, H. Eskandari, and T. Orang, "Evaluation of Visual Information Processing Speed in Depressed People," *Iran J Public Health*, vol. 42, no. 11, pp. 1266–1273, Nov. 2013.

[150] O. Hovermale, "1 individual differences in the perception of emoji: Effects of depression and self-esteem," M.S. thesis, Ball State University Muncie, Indiana, USA, Ball State University Muncie, Indiana, USA, 2020.

[151] Wikipedia contributors, *Ratio — Wikipedia, the free encyclopedia*, [Online; accessed 6-January-2023], 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Ratio&oldid=1127160304.

[152] K. van Deemter, E. Krahmer, and M. Theune, "Squibs and discussions: Real versus template-based natural language generation: A false opposition?" *Computational Linguistics*, vol. 31, no. 1, pp. 15–24, 2005. DOI: 10.1162/0891201053630291. [Online]. Available: https://aclanthology.org/J05-1002.

[153] *Bad bad words dataset*, https://www.kaggle.com/datasets/nicapotato/bad-bad-words.

[154] *Toxic comments dataset*, https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data.

[155] E. Parliament, *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts*, 2022. [Online]. Available: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729512.

[156] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," en, *Genome Inform*, vol. 13, pp. 51–60, 2002.

[157] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 8, Feb. 2012, ISSN: 1472-6947. DOI: 10.1186/1472-6947-12-8. [Online]. Available: https://doi.org/10.1186/1472-6947-12-8.

[158] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, vol. 21, pp. 137–146, 2011.

[159] B. Lamichhane, *Evaluation of chatgpt for nlp-based mental health applications*, 2023. arXiv: 2303.15727 [cs.CL].

Bibliography

Publications Related to the Thesis

Journal Articles

M. Gjoreski, T. Kolenik, T. Knez, *et al.*, "Datasets for cognitive load inference using wearable sensors and psychological traits," *Applied Sciences*, vol. 10, no. 11, 2020, ISSN: 2076-3417. DOI: 10.3390/app10113843. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3843.

T. Kolenik and M. Gams, "Persuasive technology for mental health: One step closer to (mental health care) equality?" *IEEE Technology and Society Magazine*, vol. 40, no. 1, pp. 80–86, 2021. DOI: 10.1109/MTS.2021.3056288.

T. Kolenik and M. Gams, "Intelligent cognitive assistants for attitude and behavior change support in mental health: State-of-the-art technical review," *Electronics*, vol. 10, no. 11, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10111250. [Online]. Available: https://www.mdpi.com/2079-9292/10/11/1250.

T. Kolenik, G. Schiepek, and M. Gams, "Computational psychotherapy system for mental health prediction and behavior change with a conversational agent," *Frontiers in digital health*, 2023, In submission process.

Book Chapter

T. Kolenik, "Methods in digital mental health: Smartphone-based assessment and intervention for stress, anxiety and depression," in *Integrating Artificial Intelligence and IoT for Advanced Health Informatics*, C. Comito, A. Forestiero, and E. Zumpano, Eds., In press, Springer, 2021.

Conference Papers

T. Kolenik, M. Gjoreski, and M. Gams, "Designing an Intelligent Cognitive Assistant as Persuasive Technology for Stress, Anxiety and Depression Relief," in *15th International Conference on Persuasive Technology, Adjunct Proceedings (PERSUASIVE 2020)*, (Aalborg, Denmark), M. Skov, L. B. Bertel, S. B. Gram-Hansen, and R. Orji, Eds., ser. CEUR Workshop Proceedings, 2020. [Online]. Available: http://ceur-ws.org/Vol-2629/6%5C_poster%5C_kolenik.pdf.

T. Kolenik and M. Gams, "PerMEASS – Personal Mental Health Virtual Assistant with Novel Ambient Intelligence Integration," in *http://ceur-ws.org/Vol-2820/*, CEUR-WS, Santiago de Compostela, Spain, 2020, pp. 8–12. [Online]. Available: http://ceur-ws.org/Vol-2820/AAI4H-2.pdf.

S. Pajmon, L. Kovač, T. Pajk, Ž. Besal, and T. Kolenik, "Building blocks for a computational psychotherapy system: A partial implementation," in *Proceedings of the MEi: CogSci Conference*, vol. 16, 2022.

T. Kolenik and J. Caporusso, "The one-ness of change: An exploratory neurophenomenological single case study on change in mood," in *Cognitive science: proceedings of the 24th International Multiconference Information Society–IS*, 2021, pp. 330–336.

T. Kolenik and M. Gams, "Povečevanje enakosti (oskrbe duševnega zdravja) s prepričljivo tehnologijo," in *Proceedings of the 23rd International Multiconference Information Society–IS*, 2020, pp. 55–58.

Other Publications

Journal Articles

M. Gams and T. Kolenik, "Relations between electronics, artificial intelligence and information society through information society rules," *Electronics*, vol. 10, no. 4, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10040514. [Online]. Available: https://www.mdpi.com/2079-9292/10/4/514.

M. Gjoreski, B. Mahesh, T. Kolenik, *et al.*, "Cognitive load monitoring with wearables–lessons learned from a machine learning challenge," *IEEE Access*, vol. 9, pp. 103 325–103 336, 2021. DOI: 10.1109/ACCESS.2021.3093216.

V. Janko, G. Slapničar, E. Dovgan, *et al.*, "Machine learning for analyzing non-countermeasure factors affecting early spread of covid-19," *International Journal of Environmental Research and Public Health*, vol. 18, no. 13, p. 6750, 2021.

J. Jug, T. Kolenik, A. Ofner, and I. Farkaš, "Computational model of enactive visuospatial mental imagery using saccadic perceptual actions," *Cognitive Systems Research*, vol. 49, pp. 157–177, 2018.

A. Jankovič, T. Kolenik, and V. Pejović, "Can personalization persuade? study of notification adaptation in mobile behavior change intervention application," *Behavioral Sciences*, vol. 12, no. 5, p. 116, 2022.

T. Kolenik, "Seeking after the glitter of intelligence in the base metal of computing: The scope and limits of computational models in researching cognitive phenomena," *Interdisciplinary Description of Complex Systems: INDECS*, vol. 16, no. 4, pp. 545–557, 2018.

T. Kolenik, "Kaj imata skupnega heidegger in roomba: Utelešena umetna inteligenca kot peskovnik kontinentalne filozofije," *Analiza*, vol. 20, no. 3–4, pp. 147–164, 2016.

Conference Papers

T. Kolenik, "Embodied cognitive robotics, the question of meaning and the necessity of non-trivial autonomy," in *Cognitive science: proceedings of the 19th International Multiconference Information Society–IS*, 2016, pp. 24–27.

T. Kolenik and M. Gams, "Increasing mental health care access with persuasive technology for social good," in *IJCAI 2021 Workshop on AI for Social Good*, 2021.

T. Kolenik and M. Gams, "Progressing social good by reducing mental health care inequality with persuasive technology," in *AI for Social Good Workshop*, 2020.

L. Kovač and T. Kolenik, "Detection and forecasting of mental health phase transitions from text data," in *Proceedings of the MEi: CogSci Conference*, vol. 17, 2023.

P. Šiško and T. Kolenik, "Computational investigation of phase transitions in mental health," in *Proceedings of the MEi: CogSci Conference*, vol. 16, 2022.

A. Jankovič, T. Kolenik, and V. Pejović, "The role of personality-tailored notifications in mobile-based behavior change intervention," vol. 6, 2021.

G. Slapničar, V. Janko, T. Kolenik, M. Luštrek, and M. Gams, "Cognitive, psychological and social influence on spread of covid-19," in *Proceedings of the 23rd International Multiconference Information Society–IS*, 2020, pp. 56–59.

T. Kolenik and M. Gams, "The state of the integrated information theory, its boundary cases and the question of "phi-conscious" AI," in *Cognitive science: proceedings of the 23rd International Multiconference Information Society–IS*, 2019, pp. 25–29.

T. Kolenik and U. Kordeš, "Why true perceptions die out and how embodiment helps: Modelling evolution with genetic algorithms," in *The Science of Consciousness*, 2019, p. 211.

T. Kolenik, "Are truer perceptions really better perceptions? a genetic algorithm study," in *MEi: CogSci Conference 2018*, 2018, p. 32.

T. Kolenik, "Exploring features of cognitive science as natural epistemology," in *21st International Multiconference Information Society - IS 2018*, 2018, pp. 60–64.

T. Kolenik, "Symbol grounding through action and language in cognitive robotics," in *MEi: CogSci Conference 2016, Vienna*, 2016.

Biography

Tine Kolenik (born 28. 6. 1991) earned his Master of Science degree in Cognitive Science from a joint international master's programme offered by the University of Vienna, Medical University of Vienna, University of Ljubljana, Comenius University in Bratislava, and Eötvös Loránd University Budapest. He was honored with the Prešeren Award, Slovenia's highest academic decoration for outstanding academic work, for his master's thesis titled "Computer modelling of the influence of natural selection on perceptual veridicality". He completed his master's studies in the top 1% of his class.

During his doctoral studies at the Jožef Stefan International Postgraduate School, he achieved a 9.97 grade point average. He has authored over 10 articles in prestigious scientific journals and a book chapter published by Springer International Publishing. Additionally, he has presented more than 15 papers at major conferences, including AI for Social Good (2020, Harvard, USA), The Science of Consciousness (2020, Switzerland), European Conference on Artificial Intelligence (2021, Spain), and International Joint Conference on Artificial Intelligence (2021, Canada), where he received multiple best paper awards.

Kolenik has co-organized events such as multiple International multiconferences Information Society (IS) and served as a co-chair and co-editor for the IS's Cognitive Science conferences. He also co-organized UbiComp's UbiTtention workshop and is an assistant editor for Informatica, an international journal of computing and informatics. He has been a programme committee member for IS and the AI for Social Good workshop. He regularly reviews for journals like Data Mining and Knowledge Discovery, Machine Intelligence Research, Journal of Multidisciplinary Healthcare, Neuropsychiatric Disease and Treatment, Psychology Research and Behavior Management, and others.

He served as a teaching assistant for the Cognitive Sciences course at the Jožef Stefan International Postgraduate School and has supervised bachelor's theses as well as assisted students with their master's theses. Kolenik has also been invited as a guest lecturer at various institutes in Slovenia, such as the Faculty of Computer and Information Science, and internationally, including the Department of Psychiatry at Amsterdam University Medical Centers in the Netherlands.